Introducción a la genética de poblaciones y a la genética cuantitativa

Hugo Naya
Federica Marín
Mauricio Langleib
Graciana Castro

UNIVERSIDAD DE LA REPÚBLICA
COMISIÓN SECTORIAL DE ENSEÑANZA
Y EDUCACIÓN PERMANENTE

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Prorrectorado
de **Enseñanza**

La mosca y la sopa científica

El arte que se muestra en la página anterior es producto del trabajo de Mauricio Langleib de Souza, y fue realizado específicamente para la presente edición. Se trata de una obra digital desarrollada sobre una tableta electrónica, y combina el calco de un ejemplar real de *Drosophila melanogaster* con diseños de carácter más abstracto.

Este animal es el objeto central (la «mosca de la fruta»). Esto no es casualidad, sino un homenaje al aporte que el estudio de este organismo modelo brindó al desarrollo de la genética como disciplina.

El título de la obra puede hacer honor a la interacción entre la mosca y esta combinación de disciplinas genéticas; sin embargo, no puede negarse, para los entendidos en la materia, que el homenaje mayor presente en él es hacia un disco de Patricio Rey y sus Redonditos de Ricota.

Introducción a la genética de poblaciones y a la genética cuantitativa

Hugo Naya, Federica Marín y Mauricio Langleib
con la colaboración de Graciana Castro

Financiación: Comisión Sectorial de Enseñanza y Educación Permanente,
Universidad de la República

Introducción a la genética de poblaciones y a la genética cuantitativa /
Hugo Naya, Federica Marín, Mauricio Langleib y Graciana Castro. Montevideo: Universidad de la República. Comisión Sectorial de Enseñanza y Educación Permanente, 2026.

911 p. -- (Manuales didácticos / Comisión Sectorial de Enseñanza y Educación Permanente).

ISBN: 978-9974-0-2365-9

1. GENETICA DE POBLACIONES
2. GENETICA CUANTITATIVA
3. ENSEÑANZA SUPERIOR
4. UNIVERSIDADES

I. Naya, Hugo II. Marín, Federica III. Langleib, Mauricio IV. Castro, Graciana

CDD: 575.1

La publicación de este libro fue realizada con el apoyo de la Comisión Sectorial de Enseñanza y Educación Permanente (CSEEP) de la Universidad de la República (Udelar).

Este manual es el producto de la presentación de un proyecto en el llamado de la CSEEP de la Udelar «Elaboración de manuales didácticos para la enseñanza de grado», 2022, que fuera aprobado en el mismo año y financiado durante 2023.

Atribución-NoComercial-CompartirIgual 4.0, Hugo Naya, Federica Marín, Mauricio Langleib, Graciana Castro. [2025].
https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es

Comunicación y Publicaciones, CSEEP, Udelar.
Alberto Lasplaces 1620, 11600 Montevideo
Tel.: (+598) 2400 8393
www.cse.udelar.edu.uy comunicacion@cse.udelar.edu.uy

ISBN: 978-9974-0-2365-9
Coordinación editorial: Matías Núñez
Corrección de estilo y diseño editorial: Felipe Correa Buroni y Gabriela Basaldúa

Índice general

Prólogo 15

Prefacio 18
Nuestra filosofía del *no* (tanto) . . . 20
¿Qué ES y qué NO ES este libro? . . . 20
Bibliografía recomendada para este curso . . . 23
Responsabilidades y Agradecimientos . . . 24
Íconos utilizados en este libro . . . 27

Parte I: Genómica 29

1 Introducción a la genómica 29
1.1 Variabilidad genética . . . 30
1.1.1 Detectando la variabilidad: diferentes técnicas . . . 36
1.1.2 Metagenómica y metatranscriptómica . . . 38
1.2 Genómica composicional . . . 45
Contenido GC genómico, génico, correlaciones y GC_{skew} . . . 46
Uso de codones . . . 50
1.2.1 Exploración multivariada: análisis de correspondencia . . . 54
1.2.2 Composición de proteínas . . . 56
Diferencias en propiedades de los AA . . . 56
Diferencias de composición entre clases de proteínas . . . 58
1.3 Genómica comparativa . . . 61
Comparación de secuencias génicas . . . 62
Selección positiva . . . 63
Vías . . . 64
1.4 Genómica funcional . . . 65
Estudios epigenéticos y epigenómicos . . . 68
1.5 Conclusión . . . 69
1.6 Actividades . . . 70
1.6.1 Control de lectura . . . 70
1.6.2 ¿Verdadero o falso? . . . 70
1.6.3 Ejercicios . . . 71

Parte II: Genética de poblaciones 77

2 Variación y equilibrio de Hardy-Weinberg 77
2.1 El equilibrio de Hardy-Weinberg . . . 79
Supuestos que asumimos se cumplen para H-W . . . 82
2.2 Hardy-Weinberg en especies dioicas (dos sexos) . . . 82
2.3 H-W: la frecuencia de heterocigotas en función de la frecuencia alélica . . . 85

2.4 El equilibrio de Hardy-Weinberg en cromosomas ligados al sexo 87
2.5 Tres o más alelos 90
2.6 La estimación de frecuencias y el equilibrio (o no) 92
2.7 El sistema ABO 98
2.8 ¿Dónde se «esconden» los alelos recesivos? 100
2.9 Hardy-Weinberg en especies poliploides 103
2.10 Geometría y genética: los diagramas de de Finetti 105
2.11 La estimación de frecuencias en el locus ABO 108
2.12 Conclusión 112
2.13 Actividades 113
2.13.1 Control de lectura 113
2.13.2 ¿Verdadero o falso? 113
2.13.3 Ejercicios 114

3 Deriva genética **125**
3.1 El rol de los procesos estocásticos en la genética 126
3.2 El modelo de Wright-Fisher 132
3.3 La subdivisión poblacional
y la evolución de las frecuencias alélicas 138
Homocigosidad y heterocigosidad 141
3.4 Cadenas de Markov 149
3.5 Tamaño efectivo poblacional 154
3.6 Aproximación de difusión 160
Caminatas al azar y procesos de difusión 160
Kolmogorov Forward Equation 168
Kolmogorov Backward Equation 169
Solución de las ecuaciones 171
3.7 Probabilidad de fijación y tiempos de fijación 173
3.8 El modelo coalescente 177
3.9 Conclusión 183
3.10 Actividades 184
3.10.1 Control de lectura 184
3.10.2 ¿Verdadero o falso? 184
3.10.3 Ejercicios 185

4 Selección natural **195**
4.1 El concepto de *fitness* 196
4.2 Selección natural en el modelo de un locus con dos alelos 197
4.3 Diferentes formas de selección 204
4.3.1 Selección direccional 205
4.3.2 Selección estabilizadora 210
4.3.3 Selección disruptiva 215
4.4 El teorema fundamental de la selección natural 218
4.5 Equilibrio selección-mutación 223
4.5.1 El principio de Haldane-Muller 226
4.6 La fuerza de la selección natural 228
4.7 Equilibrio selección-deriva 232
4.8 Otros tipos de selección y complejidades 237

4.9 Conclusión . . . 241
4.10 Actividades . . . 242
4.10.1 Control de lectura . . . 242
4.10.2 ¿Verdadero o falso? . . . 242
4.10.3 Ejercicios . . . 243

5 Dinámica de dos loci 250
5.1 Desequilibrio de ligamiento y recombinación . . . 251
5.2 La evolución en el tiempo del desequilibrio de ligamiento . . . 262
5.3 Otras medidas de asociación . . . 267
5.4 Selección en modelos de dos loci . . . 269
Selección en un locus: impacto en loci en desequilibrio de ligamiento . . . 273
5.5 Arrastre genético (*genetic hitchhiking o genetic draft)* . . . 277
5.6 Causas del desequilibrio de ligamiento . . . 281
Desequilibrio debido al mestizaje . . . 282
Desequilibrio inducido por el sistema de apareamiento . . . 284
El cromosoma Y . . . 284
Desequilibrio debido a recombinación reducida . . . 285
5.7 Conclusión . . . 286
5.8 Actividades . . . 287
5.8.1 Control de lectura . . . 287
5.8.2 ¿Verdadero o falso? . . . 287
5.8.3 Ejercicios . . . 288

6 Apareamientos no-aleatorios 297
6.1 El concepto de «identidad por ascendencia» (IBD) . . . 298
6.2 Generalización de Hardy-Weinberg para apareamientos no-aleatorios . . . 302
6.3 *F* como correlación entre gametos unidos . . . 305
6.4 Endocría y depresión endogámica . . . 308
6.5 Un caso extremo: la autogamia . . . 310
6.6 El coeficiente de endocría y los estadísticos *F* . . . 312
6.7 El efecto Wahlund . . . 320
6.8 Subdivisión, migración y el modelo de islas . . . 327
6.9 Mecanismos de especiación . . . 335
6.10 Conclusión . . . 338
6.11 Actividades . . . 339
6.11.1 Control de lectura . . . 339
6.11.2 ¿Verdadero o falso? . . . 340
6.11.3 Ejercicios . . . 341

7 Dinámica y genética de poblaciones microbianas 351
7.1 Genómica y mecanismos de herencia en procariotas . . . 353
7.2 Dinámica de las poblaciones bacterianas . . . 357
El modelo exponencial . . . 358
El modelo logístico . . . 363
7.3 Modelos haploides de selección natural . . . 370
Selección haploide: modelo discreto . . . 370
Selección haploide: modelo continuo . . . 376

7.4 Modelos de Moran y de fisión versus Wright-Fisher . . . 379
7.5 El rol de la transferencia horizontal . . . 383
7.6 Seleccionismo versus neutralismo: los procariotas en el debate . . . 387
7.7 Genómica poblacional . . . 389
7.8 Genes de resistencia . . . 394
7.9 Introducción a la epidemiología: modelos compartimentales . . . 398
7.10 Conclusión . . . 402
7.11 Actividades . . . 403
7.11.1 Control de lectura . . . 403
7.11.2 ¿Verdadero o falso? . . . 403
7.11.3 Ejercicios . . . 405

Parte III: Genética cuantitativa **414**

8 El modelo genético básico **414**
8.1 Variación continua y discreta . . . 416
8.2 El modelo genético básico . . . 420
8.3 Modelo genético básico: un locus con dos alelos . . . 425
Media de la población . . . 431
8.4 Efecto medio . . . 435
8.5 Valor reproductivo (o valor de cría) . . . 439
Derivación de los efectos medios por mínimos cuadrados . . . 442
8.6 Desvío de dominancia . . . 444
8.7 Interacción genotipo-ambiente . . . 450
8.8 La varianza en el modelo genético básico . . . 451
Componentes genéticos de la varianza: un locus con dos alelos . . . 452
La varianza ambiental . . . 456
La varianza de la interacción . . . 457
8.9 Conclusión . . . 458
8.10 Actividades . . . 459
8.10.1 Control de lectura . . . 459
8.10.2 ¿Verdadero o falso? . . . 459
8.10.3 Ejercicios . . . 460

9 Parentesco y semejanza entre parientes **468**
9.1 Parentesco . . . 469
9.2 Parentesco aditivo . . . 471
Consanguinidad . . . 475
Diagramas de flechas . . . 477
El método tabular . . . 482
9.3 Parentesco de dominancia . . . 486
9.4 Semejanza entre parientes . . . 488
9.5 Estimación de las varianzas aditiva y de dominancia . . . 493
Propiedades deseables de los estimadores . . . 496
Estimación de la varianza aditiva a partir del análisis de varianza (ANOVA) 498
Estimaciones basadas en modelos mixtos lineales . . . 503
9.6 Parentesco genómico . . . 505
Contenido génico . . . 506

La matriz de similaridad genómica . . . 510
9.7 Estudios de ascendencia (ancestría) . . . 514
9.8 Conclusión . . . 522
9.9 Actividades . . . 523
9.9.1 Control de lectura . . . 523
9.9.2 ¿Verdadero o falso? . . . 523
9.9.3 Ejercicios . . . 524

10 Parámetros genéticos: heredabilidad y repetibilidad 530
10.1 Heredabilidad . . . 532
La heredabilidad como relación de varianzas . . . 532
La heredabilidad como regresión . . . 536
Regresión del valor de cría en el fenotipo individual . . . 537
Regresión del valor fenotípico de los hijos sobre el fenotipo de un progenitor . . . 542
Regresión del valor fenotípico de los hijos sobre el fenotipo de ambos progenitores . . . 547
Precisión en las estimaciones de heredabilidad: varianza del estimador . . . 548
10.2 Heredabilidad en sentido amplio y sentido estricto . . . 554
10.3 Heredabilidad lograda . . . 556
10.4 Heredabilidad en poblaciones agronómicas o de laboratorio versus poblaciones naturales . . . 562
Estimaciones de heredabilidad en animales domésticos . . . 563
Estimaciones de heredabilidad en plantas . . . 568
Estimaciones de heredabilidad en condiciones naturales . . . 568
10.5 Heredabilidad y filogenética . . . 569
10.6 Heredabilidad en la era genómica . . . 571
10.7 Métodos más avanzados de estimación . . . 572
10.8 Repetibilidad . . . 573
La repetibilidad como cociente de varianzas . . . 575
La repetibilidad como coeficiente de correlación intraclase . . . 575
La relación entre la heredabilidad y la repetibilidad . . . 577
La reducción en la varianza fenotípica con varias medidas . . . 578
10.9 La repetibilidad como herramienta en la predicción . . . 582
10.10 Evolucionabilidad . . . 585
10.11 Conclusión . . . 586
10.12 Actividades . . . 587
10.12.1 Control de lectura . . . 587
10.12.2 ¿Verdadero o falso? . . . 587
10.12.3 Ejercicios . . . 588

11 Selección artificial I 593
11.1 Factores de corrección . . . 594
Factores de corrección aditivos . . . 596
Factores de corrección multiplicativos . . . 599
Factores de corrección lineales . . . 602
11.2 La respuesta a la selección y su predicción . . . 604
11.3 Diferencial de selección e intensidad de selección . . . 611
11.4 Intensidad de selección y proporción seleccionada . . . 614

¿De dónde sale la relación $i = \frac{z}{p}$? 619
La relación entre diferentes formas de calcular la intensidad de selección . . 620
11.5 Intervalo generacional 621
11.6 Medidas de la respuesta 625
11.7 Progreso genético generacional y anual 626
11.8 Cambio en las frecuencias alélicas bajo selección artificial 628
11.9 El efecto de la selección en la variabilidad genética: el efecto Bulmer 635
11.10 El diferencial de selección direccional y la identidad de Robertson-Price . . 639
11.11 Conclusión 641
11.12 Actividades 642
11.12.1 Control de lectura 642
11.12.2 ¿Verdadero o falso? 642
11.12.3 Ejercicios 643

12 Correlaciones y respuesta correlacionada **649**
12.1 Causas genéticas y ambientales de las correlaciones 650
Causas de las correlaciones genéticas 650
Causas de las correlaciones ambientales 655
12.2 Introducción al *path analysis* 657
La relación entre las correlaciones fenotípica, genética y ambiental 659
12.3 Métodos para determinar la correlación genética 663
12.4 La correlación fenotípica y su relación con otras correlaciones 666
12.5 Respuesta correlacionada 671
Eficiencia en relación con la selección directa 675
El signo de la correlación y el beneficio-demérito de la misma 678
12.6 Matrices de varianza-covarianza 679
12.7 La forma generalizada de la ecuación del criador 685
12.8 Conclusión 691
12.9 Actividades 692
12.9.1 Control de lectura 692
12.9.2 ¿Verdadero o falso? 692
12.9.3 Ejercicios 693

13 Selección artificial II **698**
13.1 Criterios y objetivos de selección 699
13.2 Selección basada en un solo tipo de fuente de información 702
Diferentes tipos de selección basados en estructuras de parentesco definidas 703
Cálculo de las heredabilidades en los distintos tipos de selección 705
13.3 Combinando la información proporcionada por diferentes tipos de parientes 710
Cálculo de los elementos de la matriz de varianzas-covarianzas fenotípica . . 712
Cálculo de los elementos del vector de varianzas-covarianzas genéticas . . . 713
Información propia y de un progenitor 713
Información propia y de varios hermanos enteros 718
Pruebas de progenie 722
13.3.1 Límites de la eficiencia en pruebas de progenie 723
Precisión del valor de cría a partir del índice 725
13.4 Métodos de selección para varias características 726
Selección en tándem 727
Niveles independientes de rechazo 729

Índices de selección 730
13.5 Índices de selección para múltiples características 734
El índice de Smith-Hazel 737
13.6 Métodos y técnicas avanzadas de selección 740
BLUP: las ecuaciones del modelo mixto lineal 741
13.7 Conclusión 747
13.8 Actividades 748
13.8.1 Control de lectura 748
13.8.2 ¿Verdadero o falso? 748
13.8.3 Ejercicios 749

14 Endocría, exocría, consanguinidad y depresión endogámica 753
14.1 El aumento de la consanguinidad a partir del número de individuos 755
¿Qué proporción de machos es necesaria para mantener un tamaño efectivo poblacional mínimo? 760
14.2 El coeficiente de consanguinidad en razas lecheras 763
14.3 Depresión endogámica 767
Los riesgos del uso masivo de reproductores y la endogamia elevada 772
Redistribución en la varianza genética 776
14.4 Exocría y heterosis 779
El modelo del templo griego 784
Un modelo sencillo de heterosis bajo exocría 787
Complementariedad 791
Tipos de cruzamientos 792
14.5 Modelo genético de cruzamientos 793
14.6 Conclusión 802
14.7 Actividades 803
14.7.1 Control de lectura 803
14.7.2 ¿Verdadero o falso? 803
14.7.3 Ejercicios 804

15 Normas de reacción e interacción genotipo-ambiente 810
15.1 Plasticidad fenotípica y normas de reacción 811
15.2 Interacción G×E en dos ambientes 817
15.3 Correlación genética a través de dos ambientes 823
Procedimientos de estimación 827
15.4 Genética cuantitativa de la interacción G×E 829
15.5 Otros ejemplos de interacción genotipo-ambiente 836
15.6 Conclusión 839
15.7 Actividades 840
15.7.1 Control de lectura 840
15.7.2 ¿Verdadero o falso? 840
15.7.3 Ejercicios 841

Apéndice A: Conceptos matemáticos básicos **847**
Fracciones . . . 847
Fracciones equivalentes y fracciones irreducibles . . . 847
Operaciones con fracciones . . . 847
Común denominador . . . 848
Operaciones combinadas con fracciones . . . 850
Proporcionalidad . . . 850
Regla de tres . . . 850
Cifras significativas y redondeo . . . 851
Operaciones con cifras significativas . . . 851
Producto y división . . . 852
Notación científica . . . 852
Ecuaciones . . . 853
Ecuaciones polinómicas . . . 853
Funciones . . . 854
Ceros de una función . . . 855
Tipos de funciones: . . . 855
Sumatoria y productoria . . . 859

Apéndice B: Cálculo diferencial e integral **860**
Límites . . . 860
Límite finito . . . 860
Límite infinito . . . 860
Derivadas . . . 863
Derivada puntual . . . 863
Derivada de una función . . . 863
Gráfica de puntos críticos . . . 864
Integrales . . . 868
Integral indefinida o primitiva . . . 868
Integral definida . . . 868
Relación entre primitiva e integral definida . . . 869
Métodos de integración . . . 869
Ecuaciones diferenciales . . . 871

Apéndice C: Álgebra lineal y geometría analítica **873**
Matrices . . . 873
Relación con los sistemas de ecuaciones . . . 874
Álgebra de matrices . . . 875
Determinante de una matriz . . . 878
Norma, producto escalar y producto vectorial . . . 879
Vectores . . . 879
Distancia y norma . . . 880
Producto escalar . . . 880
Producto vectorial . . . 880
Transformaciones lineales . . . 881
Núcleo e imagen de una transformación lineal . . . 881
Geometría de matrices . . . 881
Valores y vectores propios . . . 882
De una transformación lineal: . . . 882

De una matriz: . . . 882
Diagonalización de una matriz . . . 883
Derivada de la forma cuadrática . . . 883

Apéndice D: Probabilidad y estadística **884**
Probabilidad y conteo . . . 884
Probabilidad condicional . . . 885
Variables aleatorias . . . 885
Variables aleatorias discretas . . . 886
Variables aleatorias continuas . . . 887
Variables independientes . . . 888
Valor esperado . . . 888
Varianza y desviación estándar . . . 889
Covarianza y coeficiente de correlación . . . 890
Ley de los grandes números . . . 890
Distribuciones conocidas . . . 891
Distribución normal . . . 892
Distribución binomial . . . 895
Distribución de Poisson . . . 896
Estimadores . . . 897
Estimadores frecuentistas . . . 897
Otras propiedades: . . . 898
Métodos de estimación . . . 898
Estimadores bayesianos . . . 899

Bibliografía **901**

Prólogo

Previo al descubrimiento de las leyes de Mendel, Francis Galton observó que existía una relación casi lineal entre la estatura de los padres y la de sus hijos. En promedio, los descendientes de padres (madres) altos eran un poco menos altos, mientras que las progenies de madres (padres) bajos eran un poco más altas que sus progenitores. Sin embargo, la distribución de estaturas en todos los descendientes (en condiciones homogéneas) permanecía incambiada con respecto a la de la generación parental, debido al fenómeno entonces desconocido de la segregación mendeliana, que restaura variabilidad. Ni Darwin ni Galton podían explicar la regeneración de la diversidad, ya que la herencia de *blending* (mezcla) que postulaban llevaba, en teoría, a una pérdida de variación genética en generaciones siguientes. La estatura humana constituye un excelente ejemplo de lo que hoy se conoce como «característica compleja». La genética cuantitativa, disciplina relevante en agricultura y en predicción genómica aplicada a la medicina personalizada,[1] se ocupa de las mismas.

Una vez redescubiertas las leyes de Mendel a principios del siglo XX, Sewall Wright, Ronald A. Fisher y John B. S. Haldane sentaron las bases teóricas de la genética de poblaciones. Esta rama de la genética se ocupa de la dinámica de las frecuencias de variantes (alelos) en poblaciones finitas, de los efectos de fuerzas selectivas, y de los fenómenos de especiación y evolución. Wright (genetista) derivó medidas de consanguinidad y de parecido genético entre parientes; también demostró que, en poblaciones finitas, en el largo plazo se produce una extinción de la variación genética. Fisher (estadístico) introdujo el análisis de varianza, y propuso dividir la variabilidad entre individuos en componentes «causales», genéticos y no genéticos. Posteriormente introdujo la noción de verosimilitud estadística, sentando las bases de la estadística moderna. Haldane (matemático) desarrolló ecuaciones de equilibrio entre la mutación (modificación genética, generalmente aleatoria) y la selección. Esa era una época en la que prácticamente no había datos para analizar, pero en la que emergieron teorías útiles e influyentes como la de la importancia de variantes neutrales (Motō Kimura en Wisconsin y Japón) en procesos evolutivos.

Los desarrollos teóricos de Wright, Fisher y Haldane fueron integrados y adaptados a la producción animal por Jay Lush, profesor de Iowa State University en EE. UU. Así surgió el primer texto científico (*Animal Breeding Plans*, 1945) para un área que se denominaba Zootecnia, y que era meramente descriptiva. Lush reemplazó una visión veterinaria, basada en el individuo, por otra que usa la genética y la ciencia estadística, focalizándose en poblaciones, en inferencia y en predicción. En Uruguay, el texto de Lush fue introducido a principios de la década del sesenta en la Facultad de Agronomía de la Universidad de la República, Uruguay, por el ingeniero agrónomo Jaime Rovira (alumno de Lush). Otra influencia «local» fue la del Dr. Jorge de Alba (México) a través de su libro en castellano *Reproducción y Genética Animal*. Un necesario complemento al libro de Lush era un texto de D. S. Falconer (*Introduction to Quantitative Genetics*, 1961), profesor de Edimburgo en esa época. La mayoría de los estudiantes de agronomía consideraba a Falconer como

[1] De los Campos G. & Gianola D. 2023. *Genomic prediction in the big data era.* American Scientist. DOI: 10.1511/2023.111.5.286

«demasiado matemático», pero era muy difícil cuestionar el cuidado y la claridad meridiana con que ese libro había sido escrito.

Lush postulaba que las características de importancia económica en ganadería y agricultura, como el crecimiento y la reproducción animal, la producción de fibra y leche, el rendimiento vegetal (grano, área foliar), la resistencia a las enfermedades y la adaptación (termo-regulación, equilibrio metabólico) dependen de la acción concertada de cientos de genes en compleja interacción con el medio ambiente. Lush notaba que lo que se puede medir u observar no es más que una indicación indirecta del «valor de cría», especialmente en el caso de variables que reflejan la aptitud reproductiva, fuertemente influidas por efectos ambientales. Las ideas de Lush, inspiradas por Wright y por problemas específicos de producción animal, fueron posteriormente incorporadas al mejoramiento genético vegetal.

Como los genes «causales» y su acción a nivel molecular no se conocían, la situación estimuló el desarrollo de modelos matemáticos y estadísticos para describir y explicar los fenómenos evolutivos y los efectos esperados de la selección natural y artificial. Prácticamente, en más de medio siglo se acumularon desarrollos teóricos que «aguardaban» al conocimiento que posteriormente derivaría a partir de la decodificación del genoma. El descubrimiento del ADN (ácido desoxirribonucleico) como material molecular responsable de la información genética representó una especie de bomba atómica detonada en las ciencias biológicas. Una enorme inversión pública y privada posibilitó que se llegara a «leer» los 3000 millones de letras que corresponden a cuatro bases nitrogenadas en el genoma humano (A, T, G, C), así como secuenciar genomas de cerdos, vacunos, peces, maíz, trigo, café, árboles, bacterias y diferentes tipos de virus. Comenzó entonces la «era genómica», con desarrollo posterior de estudios de expresión génica, de la epigenómica (estudio de zonas del genoma asociadas a metilación diferencial), la corroboración de la teoría evolutiva a nivel molecular, la paleo genómica (¿tiene ancestros Neanderthal en su familia?), la astrogenómica (astrobiología o búsqueda de formas de vida en el espacio exterior), la puesta en acción de la epidemiología genómica, siguiendo en tiempo real la evolución del SARS-CoV-2, la producción de vacunas de ARN mensajero, la «edición» génica (ya hay un tratamiento aprobado para la anemia perniciosa), ¡y más!

Hasta principios del siglo XXI, la genómica había tenido un impacto marginal en la agricultura. Los proyectos de secuenciación de animales y plantas, más el desarrollo de plataformas de «genotipado» y de técnicas de bioinformática (algoritmos de alineamiento y ensamblaje, por ejemplo), permitieron identificar millones de biomarcadores moleculares. Estos explican poca variación genética individualmente, pero colectivamente (usando métodos estadísticos y de machine learning) pueden tener una capacidad sustancial para predecir la performance. Surge entonces la «selección genómica», basada en evaluaciones genéticas más precisas de animales y plantas a partir de miles de marcadores, y con estimaciones más confiables del «valor genómico» medido en fases juveniles, acortando así intervalos intergeneracionales. Un correlato es la predicción genómica de enfermedades complejas en seres humanos, todavía en fase incipiente.

De estos temas se trata este libro, que parece único en varios aspectos. El lector es introducido rápidamente a la organización genómica de los organismos. Este tratamiento es radicalmente diferente del que se ha presentado en muchos textos clásicos y recientes de genética cuantitativa. En estos últimos el genoma es tratado como una serie linealmente organizada de posiciones estadísticas en los cromosomas. En segundo término, y manteniendo una perspectiva genómica, el texto efectúa una presentación moderna de la genética de poblaciones, incluyendo el concepto de coalescencia. El tratamiento de temas como el de

deriva aleatoria, la evolución de las frecuencias génicas en poblaciones finitas y la aplicación genómica en epidemiología, es ágil y riguroso. La tercera parte del libro considera nociones básicas de genética cuantitativa. La cobertura es clásica, pero completa y actualizada; por ejemplo, el tratamiento de la interacción genotipo-medioambiente incluye el uso de normas de reacción.

Las experiencias colectivas de los autores en áreas de biología evolutiva, bioinformática y producción animal y vegetal generan una sinergia que confiere una elogiable originalidad a la obra. No tengo conocimiento de la existencia de textos similares en inglés, francés o castellano.

Si bien el libro coloca a la pedagogía en un plano elevado, la obra no es para débiles de corazón. Los autores advierten e invitan a los estudiantes a considerar el tratamiento abstracto de los problemas, y proporcionan a tales efectos útiles esenciales. El libro incluye cuatro apéndices como, por ejemplo, una introducción al cálculo matemático y al álgebra matricial. Como se menciona en el Prefacio, el texto se complementa con laboratorios virtuales que incluyen ejercicios y los estudiantes son invitados a programar sus cálculos en el lenguaje R.

En resumen, este original libro de texto ofrece a los estudiantes de biología, agronomía y veterinaria un tratamiento moderno y riguroso de áreas típicamente consideradas como «difíciles», como son la genética de poblaciones y la genética cuantitativa. La obra no es redundante con respecto a textos de cobertura similar, puesto que estos carecen de la visión más general e integrativa que se ofrece en este volumen. Corresponde entonces felicitar a los autores de la obra, con la seguridad de que será útil y bienvenida en el mundo hispanoparlante.

Daniel Gianola
Ingeniero agrónomo, M. Sc., PhD, Dr. Sci. H. C. (multi)
Profesor emérito Sewall Wright
Universidad de Wisconsin, Madison
EE. UU.

Prefacio

Outside of a dog a book is a man's best friend. Inside of a dog it's too dark to read.[2]

Groucho Marx

El curso de Genética de Poblaciones y Genética Cuantitativa (también conocido como Genética II) es, de alguna manera, la continuación natural del curso de Genética I. Ambos cursos fueron desarrollados en conjunto, de manera tal que permitiesen al estudiante ir avanzando en la comprensión de los conceptos, desde las bases moleculares de la herencia hasta sus implicancias en la evolución de las especies. En esta separación (algo arbitraria), el curso de Genética I hace más foco en mecanismos y procesos que se dan a nivel del individuo y que sientan las bases para entender la evolución al nivel de las poblaciones, mientras que el curso de Genética II aprovecha de esta construcción para comenzar a desarrollar nuestra comprensión sobre los fenómenos que ocurren a esta última escala.

Cuando en conjunto con los docentes de otros cursos que involucran a la genética decidimos cómo organizar y distribuir los conocimientos en el marco del nuevo Plan de Estudios (2020), resultó que Genética II debería incorporar fuertes conocimientos en tres grandes temas: Genética de poblaciones, Genética cuantitativa y Genómica. Desafortunadamente, el número de créditos disponibles era bastante limitado, por lo que debíamos optar por armar un curso que sobrevolase los temas, sin dejar espacio para que los mismos sean tratados al nivel requerido o dejar fuera demasiados temas. Nuestra solución de compromiso fue de alguna manera ortogonal a estas opciones; hemos organizado el curso en dos grandes bloques, el primero de Genética de Poblaciones y el segundo de Genética Cuantitativa, mientras que decidimos ubicar a la Genómica de manera transversal, motivándola especialmente a través de los ejemplos y ejercicios que surgen en los distintos capítulos.

Una de las grandes diferencias que encontrarás entre el curso de Genética I y el de Genética II se encuentra en las herramientas que utilizaremos para comprender los fenómenos de nuestro interés. Mientras que en Genética I gran parte del esfuerzo fue en comprender e internalizar los conceptos mecanísticos de la herencia, en Genética II nos enfocaremos en entender y modelar los conceptos que surgen de la interacción de un número importante de individuos, así como de generar predicciones a corto, mediano y largo plazo de la evolución de las poblaciones. Esto nos lleva directamente a uno de los focos metodológicos del presente curso: la capacidad de modelar sistemas biológicos, de comprender las implicancias y limitaciones de nuestros modelos, así como de interpretar sus predicciones. En particular, nos interesa desarrollar en este curso tus habilidades para

[2] *The Essential Groucho: Writings By, for and About Groucho Marx Stefan Kanfer*. La traducción podría ser algo así como «Fuera de un perro, un libro es el mejor amigo del hombre. Dentro del perro está muy oscuro para leer». Al comienzo de cada capítulo hay una nota de este tipo, en general en el idioma o dialecto original. Si bien en algunos casos esto puede sonar a pedantería, obviamente no manejamos con fluidez casi ninguno de ellos (aunque con algún esfuerzo y ayuda los entendemos), pero resulta casi tan importante como mantener la diversidad genética mantener la diversidad cultural; los lenguajes han evolucionado en paralelo con la genética de las distintas poblaciones humanas, como lo muestra Luigi Luca Cavalli-Sforza en su magnífico libro *Genes, Peoples, and Languages*.

comprender las relaciones que se generan entre los aspectos determinísticos de los sistemas biológicos y los aspectos estocásticos (lo que hace al azar). Para esto, en el contexto de la genética, es necesario contar con un conjunto mínimo de habilidades matemáticas, sin las cuales sería imposible siquiera adentrarse a la tarea. Sin duda, todas estas habilidades y conocimientos has tenido la oportunidad de adquirirlos en todo tu transcurso por el sistema educativo, ya que no se trata de nada extraordinario: buenos conocimientos (y sobre todo mucha práctica) de aritmética y álgebra elemental, teoría combinatoria, conceptos básicos de cálculo integral y diferencial, algunas nociones de álgebra lineal, un conocimiento básico de teoría de probabilidad y distribuciones, así como algunos conceptos básicos de estadística. A lo anterior deberás agregarle, un buen dominio de los conocimientos biológicos adquiridos en Genética I, así como las bases de la vida y las particularidades de la biología de los distintos dominios de la vida.

Uno de los objetivos del presente curso es desarrollar tus capacidades de razonamiento, utilizando el modelado de fenómenos evolutivos como el medio con el cual vamos a trabajar. La comprensión cabal de los modelos requiere, por lo general, de la capacidad de resolver diversos ejercicios y problemas, con variados niveles de dificultad, por lo que harás uso abundante del papel y el lápiz. A decir verdad, posiblemente sea provechoso copiar e ir desarrollando por ti mismo las distintas ecuaciones que irán apareciendo en el texto. Además, otro de los objetivos comprende el desarrollo de tus capacidades científicas, experimentando, generando hipótesis, buscando cómo probarlas y reportando los resultados de tus experimentos. Para ello hemos desarrollado laboratorios virtuales, donde podrás experimentar a tu placer. Si además eres un estudiante al que le gustan los desafíos y oportunidades que brinda el poder desarrollar tu propio código, también contamos con la posibilidad de ayudarte a desarrollar los ejercicios en el software libre R.

Por todo lo anterior, Genética II es un curso que probablemente requerirá de un importante esfuerzo de tu parte. Si bien hemos puesto todo nuestro esfuerzo en mantener la matemática empleada en el curso en un nivel básico, entendemos que reducirla más implicaría una pérdida de calidad injustificable. Por lo tanto, para poder avanzar de forma razonable te aconsejamos enfáticamente que realices la prueba diagnóstica en la forma más honesta posible. El objetivo principal de la prueba es ayudarte a detectar los principales temas, conceptos y habilidades que has olvidado y que te dificultarán el avance en nuestro curso. Te encontrarás con que gran parte de las preguntas deberás resolverlas sin el auxilio de una calculadora ni computadora; esto es porque el seguimiento de una parte importante de nuestro curso se basa en tus habilidades algebraicas y en tu capacidad de razonar en forma abstracta. Además, hemos puesto a tu disposición un material de apoyo sobre los conceptos y habilidades matemáticas básicas que asumimos para el desarrollo del curso. Puedes consultar este material antes o después de la prueba diagnóstica, como te resulte más cómodo y si aun así luego de leerlo no consigues buenos resultados estamos a tu disposición para ayudarte a encontrar la forma de superar tus dificultades.

Esta obra está bajo una Licencia Creative Commons Atribución-CompartirIgual 4.0 Internacional

Nuestra filosofía del *no* (tanto)

- **No renunciamos a usar y enseñar matemáticas** en los lugares y momentos que entendamos adecuados, para modelar y entender la realidad, para conectarnos con otras teorías o simplemente por diversión.
- **No nos molesta volver a explicar y trabajar los conceptos previos**, solo pedimos tu compromiso con el aprendizaje (es posible que debas dedicar más horas de las previstas para poder alcanzar la nivelación de cada tema).
- **No asumimos conocimientos previos profundos de matemáticas (ni de nada)**, pero nos encantaría que antes de empezar te aseguraras de confirmar que posees (o estás dispuesto a repasar/aprender bajo nuestra guía):
  - Un manejo fluido de álgebra básica en los campos numéricos de reales, enteros y racionales (propiedades, operaciones, etcétera)
  - Conceptos básicos de cálculo diferencial e integral, nociones de geometría analítica, así como idea de qué son las ecuaciones diferenciales
  - Nociones básicas de álgebra lineal en R^2 y R^3, en especial operaciones con matrices y vectores
  - Rudimentos de teoría de probabilidad, conocimientos básicos de distribuciones (Normal, Binomial, Poisson, etcétera)
  - Conceptos básicos de estadística, estadísticos resumen, prueba de hipótesis, análisis de varianza
  - Bases de la genética mendeliana, conocimientos básicos de genética molecular y genómica

¿Qué es y qué no es este libro?

El curso de Genética II, con las bases de la genética de poblaciones y de la genética cuantitativa en un mismo material, representa un gran desafío. Normalmente se trata de dos cursos separados, a veces conectados, pero los pocos casos que hemos podido encontrar con materiales disponibles no se ajustaban al nivel deseado, a los tiempos disponibles, o ambas cosas. Más aún, existen diferentes libros disponibles (mayormente en inglés) que cubren todos los temas que trabajaremos en nuestro curso, pero estos suelen tratar solo uno de los dos grandes bloques temáticos, o casi en la totalidad dedicarse a uno de los dos con el otro apenas cubierto en uno o dos capítulos. Es por estas razones que nos hemos decidido a preparar nuestro propio material, siguiendo la estructura de temas en base semanal que tiene nuestro curso, de manera de hacerte más fácil el seguimiento del curso.

Como irás viendo al leerlo, hemos procurado no asumir demasiados conocimientos previos, aunque estamos seguros de que ya has pasado por ellos previamente. En este sentido, puede que sientas cierta pesadez en explicaciones detalladas «de más», pero en cierto sentido es el precio a pagar para acompañar a los que llegan con menos práctica o conocimientos. Para nosotros, la matemática (elemental) en la que se basa nuestro curso es como la gimnasia o cualquier deporte; es posible que al comienzo aún la operación más sencilla te demande cierto esfuerzo mental y te sientas cansado al avanzar algunas pocas páginas. Como en la gimnasia o el deporte, la práctica progresiva, aumentando ligeramente la carga y dificultad

de los ejercicios, te permitirá (con suerte) llegar en mucho mejor forma hacia el final del curso y con mucha más adrenalina y ganas de continuar aprendiendo.

Este libro **no es original**, al menos en sus contenidos. Tampoco nos planteamos que lo fuese, pues nuestro objetivo era y es bien otro: preparar un material que se pudiese seguir (con voluntad y trabajo) por todos nuestros estudiantes, pese a las diferencias importantes en las historias de aprendizaje de cada uno, sin descuidar el rigor del tratamiento matemático de los problemas y sin renunciar al desarrollo de las capacidades que entendemos fundamentales en el estudiante de carreras con base científica. Es así que, en muchos casos, luego de analizar y leer diferentes libros y materiales, nos hemos decidido por seguir determinada línea de tratamiento de cada tema, muchas veces bastante pegados a algún texto original (lo que reconocemos particularmente cuando sentimos que esto se aparta de lo esperado) en aras de una mejor comprensión de cada tema. Una virtud adicional de esta aproximación es la facilidad para que puedas extender tu conocimiento sin tener que cambiar de notación o de línea argumental. Sin duda, la tentación de «reinventar la rueda» está presente en todos nosotros y es así que este enfoque nos representa un pequeño sacrificio: la originalidad. Creemos firmemente que nuestra pequeña renuncia será de enorme beneficio para nuestros estudiantes que quieran ampliar sus conocimientos.

En particular, del libro de John H. Gillespie (2004) sacamos la organización de capítulos para la parte de genética de poblaciones. Se trata de un excelente libro por el tratamiento conciso (como dice su título), con gran profundidad, pero al mismo tiempo con un tratamiento riguroso y simple. De hecho, fue posiblemente uno de los grandes inspiradores, junto al libro de Rice (2004) (dentro de nuestros 5 favoritos entre todos los campos) del deseo de escribir este manual (si bien en la extensión vamos en sentido contrario). Más aún, de John H. Gillespie (2004) sacamos varias aproximaciones a diferentes secciones y capítulos, cosa que en algunos casos se aprecia notoriamente. Otro libro excelente para la parte de genética de poblaciones y de donde también hemos sacado abundante material o seguido sus líneas de razonamiento y exposición es el Hartl and Clark (2007). Sin duda un material que podría usarse directamente en un curso del nivel que nos proponemos, aunque se encuentra organizado como curso de genética de poblaciones, lo que lo hace extenso para la mitad de nuestro curso (a esto se suma de que se trata de un libro sin traducción al castellano, al menos hasta donde sabemos). Un tercer libro del que hemos sacado ideas es el de Hamilton (2009). Se trata de un libro que hace un esfuerzo importante por dar un tratamiento intuitivo a muchos temas y en algunos casos hemos sacado ideas o aun líneas de razonamiento del mismo. En lo que hace a la parte de genética cuantitativa, la principal referencia es sin dudas el libro de Douglas S. Falconer y Mackay (1996). Se trata de un libro excepcional, que ha permanecido a lo largo de los años como referencia y que hemos sabido usar como base de otros cursos (por ejemplo, en el curso de Zootecnia de Facultad de Agronomía). Nuevamente, se trata de un libro cuyo énfasis se encuentra en una de las partes del curso (genética cuantitativa), por lo que el tratamiento de la parte de genética de poblaciones es algo escaso para nuestro objetivo; de lo contrario, hubiese sido nuestra elección primaria. Finalmente, las «biblias» de la genética cuantitativa evolutiva, el de Lynch y Walsh (1998) y el Walsh y Lynch (2018), constituyen la referencia última en casi cualquiera de los temas tratados en el presente manual. En algún caso, especialmente en algún tópico avanzado los hemos utilizado como referencia inmediata. Sin embargo, tanto el volumen y extensión del tratamiento (que juntan casi 2500 páginas entre los dos), así como la profundidad de tratamiento los dejan como material de consulta y profundización.

Si algún dejo de originalidad aún le podemos encontrar a este libro, posiblemente tenga que ver con el nivel de conocimientos previos esperados con relación a los objetivos planteados. Sin duda se trata de un plan ambicioso y para ello contamos con tu involucramiento. Con este objetivo en mente es que nos hemos esforzado por incluir la mayor cantidad posible de ejemplos y ejercicios que tengan que ver con tu futuro desarrollo profesional, intentando además evitar demasiada estrechez de miras. Al mismo tiempo, al tratarse de un curso para la carrera de Ingeniero Agrónomo, al tiempo que nos obliga a un tratamiento inclusivo de casi todos los dominios de la vida y sus mecanismos de herencia, en cierta forma (por las limitaciones de tiempo y espacio) nos obliga a reducir en gran medida el tratamiento de los detalles de la evolución molecular. Lamentablemente, esta también es una particularidad de este libro.

Posiblemente a esta altura, en vista de todas las dificultades que te esperan, te estarás haciendo varias preguntas. ¿Para qué tanta genética? ¿Tiene el Ingeniero Agrónomo que saber de genética de poblaciones y de genética cuantitativa? ¿Hace falta tener tanta matemática? ¿No alcanza con una descripción conceptual de los fenómenos que me quieren transmitir? Nuestras respuestas a estas interrogantes, como te habrás imaginado, apuntan a que es necesario e importante construir de la forma que pretendemos en este curso. El ingeniero agrónomo, como todo ingeniero, debe ser capaz de trascender respuestas genéricas y responder a preguntas concretas con soluciones específicas y adaptadas a la situación particular de cada sistema de producción y cada productor. A eso solo se llega a través del análisis de alternativas y su contraste numérico, en busca de óptimos. Como en cualquier edificación, no se construyen complejos pisos sobre débiles cimientos. Tanto Genética I como Genética II fueron y son parte de una coordinación compleja entre docentes de diferentes materias, en especial lo que llamamos la «secuencia de Genéticas», que te irán llevando desde las bases hacia las aplicaciones, cada vez más específicas, cada vez más profundas, posiblemente cada vez más cerca de ti y de lo que elegiste como vocación. Ya desde el próximo semestre de la carrera comenzarás a experimentar estos cursos más cercanos a la producción y disfrutarás de la base que adquiriste en Genética I y Genética II. Por ejemplo, tendrás el curso de Fitotecnia para el cual te serán fundamentales los conceptos de mecanismos de la herencia y equilibrio de Hardy-Weinberg, entre muchos otros. Además, si tu orientación es hacia la producción animal, tendrás el próximo semestre Genética IIIa, que será tu primera experiencia en mejoramiento genético animal y donde verás las distintas alternativas para incorporar el mejoramiento genético en especies domésticas.

En resumen, *¡ánimo!, no lo olvides*, estamos a tu disposición para ayudarte a avanzar y nuestro deseo más profundo es que, más allá de todo el esfuerzo, puedas disfrutar de seguir nuestro curso tanto como nosotros de armarlo para ti. ¡Que tengas un muy buen comienzo!

El equipo docente de Genética II

Bibliografía recomendada para este curso

La siguiente es una serie de libros que recomendamos tener a mano para seguir el curso y profundizar en aspectos que no hayan quedado claros con nuestro breve tratamiento. Los primeros conforman una lista de materiales en los que hemos basado gran parte de este libro. En algunos de ellos encontrarás un nivel más profundo y en otros más claro, por lo que especialmente te recomendamos tenerlos a mano.

Fuertemente recomendados para parte importante de los temas

- *Population Genetics: a concise guide*, John H. Gillespie (The Johns Hopkins University Press 2nd. Edition).
- *Principles of Population Genetics*, Daniel L. Hartl & Andrew G. Clark (Sinauer 4th Edition).
- *Introduction to Quantitative Genetics*, Douglas S. Falconer & Trudy F.C. Mackay (Pearson Prentice Hall).
- *Population Genetics*, Matthew B. Hamilton (Wiley-Blackwell).
- *Genética Cuantitativa*, Armando Caballero Rúa (Editorial Síntesis)
- *Mejora Genética Animal*, Agustín Blasco (Editorial Síntesis).

Recomendada para diferentes temas

- *Mejoramiento Genético Animal*, Ricardo Cardellino & Jaime Rovira (Editorial Hemisferio Sur).
- *A Primer of Population Genetics and Genomics*, Daniel L. Hartl (Oxford 4th Edition).
- Population Genetics with R: An Introduction for Life Scientists, Áki Jarl Láruson & Floyd Allan Reed (Oxford).
- Genetics and Analysis of Quantitative Traits, Michael Lynch & Bruce Walsh (Sinauer).
- *Evolution and Selection of Quatitative Traits*, Bruce Walsh & Michael Lynch (Oxford).
- *Chance in Biology Using Probability to Explore Nature*, Mark Denny & Steven Gaines (Princeton University Press).
- *Evolutionary Theory. Mathematical and Conceptual Foundations*, Sean H. Rice (Sinauer).
- *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*, Sarah P. Otto & Troy Day (Princeton University Press).

Responsabilidades y agradecimientos

El presente material es parte de un trabajo colectivo que hemos desarrollado para que los estudiantes del curso Genética II cuenten con lo necesario para introducirse en un enfoque ciertamente demandante de la genética. Estos materiales incluyen, además del presente libro, una *Guía práctica* con abundantes ejercicios, una *Guía de referencia rápida* con los elementos imprescindibles para resolver los ejercicios, diferentes presentaciones que iremos poniendo a disposición con el avance del curso y la plataforma AGROS, desde donde desarrollaremos el curso. Por lo tanto, es difícil delimitar las contribuciones y responsabilidades individuales concretas en todos los materiales.

En lo que respecta al libro, la presente versión se basa en el material creado para la primera edición del curso por HN,[3] tanto en lo que hace a la estructura, como en lo que hace a los contenidos principales. En dicha versión, el magnífico trabajo de edición fue responsabilidad de FM,[4] que durante el transcurso de la escritura se encontró con innumerables errores de tipeo (cuando no ortográficos), algunas frases de difícil lectura y otras sin sentido aparente, todo superado de forma eficiente y con gran independencia. Además, Federica fue responsable de una parte importante de los «Para recordar» que encontrarás al final de cada sección relevante del libro y que además constituyen la base de la *Guía de referencia rápida*. Sin embargo, la presente edición tiene mejoras notables y una cierta «completitud» que le faltaba a la primera versión. Gracias al apoyo recibido a través de los proyectos concursables para la generación de manuales didácticos de la Comisión Sectorial de Enseñanza de la Universidad de la República (CSE), fue posible aspirar a completar los materiales, lo que nos lleva a la presente edición. En especial, dichos fondos permitieron financiar la incorporación al equipo de ML,[5] que realizó la edición completa del material, generó diversos ejercicios con soluciones, organizó desde el punto de vista pedagógico los materiales (objetivos del capítulo, ¿verdadero o falso?, controles de lectura), agregó y mejoró distintos contenidos y, por fin, aseguró la calidad visual de los materiales generados, una tarea harto difícil con las limitaciones impuestas por Bookdown (un precio por pagar por tener el material disponible en formato PDF y web). Finalmente, los contenidos que aparecen en los apéndices fueron creados casi en su totalidad por Graciana Castro,[6] bajo un esquema inicial propuesto por HN. Graciana también recibió financiación del proyecto CSE y, si bien no figura entre los autores principales, ya que los apéndices constituyen un material que puede leerse con completa independencia y que en general asumimos que se corresponde a conocimientos previos a la lectura del libro, el excelente tratamiento de estos en el breve espacio disponible equivale, en la práctica, a cuatro pequeños capítulos.

En términos generales, un aporte inestimable fue el del profesor Jorge I. Urioste, profesor titular y responsable del Grupo Disciplinario de Mejoramiento Genético Animal durante muchísimos años y hasta hace poco. En buena parte este libro es resultado de diversas conversaciones que tuvimos a lo largo de los años, en las que siempre surgía la necesidad de generar materiales adaptados a los cursos de nuestra responsabilidad. En particular, el curso de Zootecnia siempre fue un ejemplo en este sentido ya que contaba con el libro de Cardellino y Rovira (1987) como referencia directa, casi capítulo a capítulo, lo que permitía

[3] Genética Cuantitativa Evolutiva/GD Mejoramiento Genético Animal, Facultad de Agronomía, hnaya@fagro.edu.uy & Unidad de Bioinformática, Institut Pasteur de Montevideo, naya@pasteur.edu.uy

[4] GD Mejoramiento Genético Animal, Departamento de Producción Animal y Pasturas, Facultad de Agronomía.

[5] GD Mejoramiento Genético Animal, Departamento de Producción Animal y Pasturas, Facultad de Agronomía & Unidad de Bioinformática, Institut Pasteur de Montevideo.

[6] Apéndices, Instituto de Ingeniería Eléctrica, Facultad de Ingeniería.

que los estudiantes pudieran seguirlo, aun sin poder asistir a los teóricos. Precisamente, la necesidad de actualizar el libro de Cardellino y Rovira (1987) para acomodar los enormes avances en genómica que ocurrieron desde su publicación original fue uno de los impulsores para aventurarnos en la tarea. Más allá del enorme aporte de Jorge en este sentido más general, concretamente para este libro su aporte más importante consistió en la lectura dedicada de cada capítulo, las abundantes correcciones señaladas y, nuevamente, los aportes conceptuales sobre la docencia y la comunicación con los estudiantes.

María André y Andrea Larracharte fueron las responsables de completar con ejemplos y ejercicios algunos de los diversos espacios que estaban disponibles al momento de su arribo, los cuales fueron numerosos, especialmente en la tercera parte del libro (la que tiene que ver con genética cuantitativa), así como aportar nuevos ejercicios a la *Guía práctica*. Micaela Botta, Paula Batista y Washington Bell son los responsables del armado de la mayor parte de las presentaciones que acompañan al curso. Washington, es además corresponsable junto con Ana Laura Sánchez de la plataforma AGROS. Ana Laura se hizo cargo de crear la nueva base de datos con ejercicios y evaluaciones. Además, con fondos del proyecto CSE, se encargó de agregar ejercicios a la *Guía práctica*. Rodrigo Vivián se responsabilizó del armado de la *Guía de referencia rápida* utilizando el material del libro. Por último, tanto algunos integrantes del Grupo Disciplinario MGA (Rodrigo López-Correa, Camila Etchevarría) como la gente del Departamento de Biología Vegetal (Clara Pritsch, Paola Gaiero y Marianella Quezada), siempre contribuyeron con comentarios y aportes, buscando una fuerte integración de los conceptos genéticos a lo largo de la carrera de Ingeniero Agrónomo. Otros varios amigos y colegas colaboraron con la lectura de algunos capítulos (Daniel Naya, Gabriel Naya) o simplemente cubriéndonos en tareas de nuestra responsabilidad (Martín Graña, Lucía Spangenberg). En todo caso, cualquier error u omisión en cualquiera de los trabajos es completa responsabilidad del primer autor (HN).

Luego de finalizado casi todo el trabajo de edición, y con muy poco margen para grandes cambios, decidimos enviarlo junto a otros materiales del curso a distinguidos colegas, a fin de recibir opiniones, tanto sobre los contenidos como sobre el tratamiento del material. Entre las opiniones recibidas destaco la de los profesores Armando Caballero y Agustín Blasco, así como la del Dr. Miguel Pérez-Enciso, quienes se tomaron el tiempo para responder algunas de nuestras preguntas. Eso no significa que respaldaran ninguna de nuestras decisiones o enfoques, pero en todos se apreció el esfuerzo por aportar otras visiones a un curso tan atípico como el nuestro. En particular, algún agregado de último momento responde a observaciones de estos colegas que los echaron en falta (efecto Bulmer, por ejemplo), o el responder a nuestro interés en particular por saber si el nivel de matemática requerido o empleado era adecuado al público al que destinamos la obra. A todos ellos, y a otros que no nombramos, muchas gracias.

El párrafo en el que nos encontramos es de pura responsabilidad del primer autor (HN) y solo refleja su experiencia (FM deja constancia de no compartir lo expresado). Durante las cuatro ediciones que llevó el curso, desde su inicio hasta el presente, mientras ajustábamos los materiales, como responsables del curso Genética II (HN) hemos recibido algunas críticas, muchas veces duras, pero con honestidad intelectual y el deseo de mejorar la educación de nuestros estudiantes. Sin duda, hacemos llegar nuestro agradecimiento a tan sinceras opiniones que nos permiten mejorar constantemente. Al mismo tiempo, también recibimos críticas de las otras, de las que buscan lastimar más que mejorar, agujeros debajo de la línea de flotación hechos desde adentro, provenientes en su mayoría de personas con escasa capacidad de desarrollar propuestas nuevas, conservadoras de lo peor, que culpan siempre

a otros de sus fracasos, que prefieren explicaciones místicas en lugar del trabajo y del estudio, del amiguismo en lugar de la amistad, que siempre buscan los atajos. Algunas veces esto llegó a afectar nuestra salud y en consecuencia a nuestra familia. Sin embargo, a estas personas también les damos las gracias sinceras, por obligarnos a permanecer eternamente vigilantes, buscando identificar aún los puntos más inesperados que serán sometidos al ataque malintencionado y transformarlos en oportunidades de introducir mejoras... *Fluctuat nec mergitur*.[7]

Finalmente, un pequeño espacio para un par de agradecimientos personales. Durante el año que llevó la escritura de los materiales originales, contamos (HN) con el apoyo institucional, tanto de la Facultad de Agronomía-Udelar, como del Institut Pasteur de Montevideo, lo que nos permitió dedicar gran parte de nuestro tiempo a la tarea (a veces en desmedro de otras actividades como la investigación). Sin embargo, el agradecimiento mayor es para Lucía, Guille, Agus, Marti, Alfo, Xime y Santi, así como para la familia toda (mis viejos, Carlos y Ana, mi padrino Carlos Alberto, mis suegros Eddy y Mariela, Pablo y Silvia, Bea y Marce, Gabriel y Adri, Daniel, Noyi y *el Negro*, etcétera), ya que aún en plena pandemia (con todas las dificultades adicionales que eso representó) sacrificaron una parte importante de los tiempos familiares para que pudiésemos concretar la tarea en tiempo y forma. En lo académico, a la OEG toda (Musto, *Yuyo*, Ale, etcétera), al *Pelo* y a todo el piso 4, que fueron los que me enseñaron a hacer ciencia, así como a Otto Pritsch, con el que, muchas veces sin saberlo, compartimos la visión global sobre lo que debería ser la ciencia. ML quisiera agradecer a su familia, a sus amistades y a la barra de UBi y Genómica Evolutiva por el apoyo de siempre. Además, vaya un agradecimiento especial a Zelig Langleib por inculcar el amor a los libros y la ciencia.

Para ellos entonces es la dedicatoria de nuestro humilde trabajo.

Los autores

[7] El famoso lema de la ciudad de París, que podría traducirse del latín como «se mueve pero no se hunde». En nuestro caso hacemos referencia a la canción de Georges Brassens *Les copains d'abord*, que habla de la amistad sincera.

Íconos utilizados en este libro

A lo largo de este libro se pueden apreciar algunos íconos que procuran guiar a la persona que lee respecto al contenido de algunas secciones o ejercicios. Los íconos son:

Sección con contenido avanzado, del cual se puede prescindir en una primera lectura. Estas lecturas de profundización son recomendadas para personas que quieran descubrir aplicaciones cercanas al estado del arte de una temática o aplicaciones fuera del uso sobre el cual se informa comúnmente en un libro de texto de grado del área.

Ejercicios con formato clásico encontrado en libros de genética de poblaciones (de allí que el ícono sea la cabeza de una *Drosophila melanogaster*). Generalmente refieren a los *loci* segregando en una población de organismos imaginarios o situaciones similares.

Ejercicios con particular interés para las agrociencias. Suelen referir al mejoramiento genético de animales o plantas, o refieren a organismos de especial interés para esta disciplina.

Ejercicios inspirados en experimentos de ciencia básica (por ejemplo, biología molecular, genética, bioquímica, microbiología, ecología, etcétera).

Ejercicios de demostración. Se pide demostrar alguna propiedad matemática a partir de las fórmulas presentadas en el capítulo. En general, los resultados demostrados ilustran aspectos biológicos interesantes que se desprenden de las bases teóricas desarrolladas en el texto.

Ejercicios de particular interés para las áreas biomédicas.

Parte I: Genómica

Capítulo 1

Introducción a la genómica

I'm sure the universe is full of intelligent life. It's just been too intelligent to come here.[1]

Arthur C. Clarke

El presente capítulo es una somera introducción a la genómica y apenas tiene como cometido el asegurarnos de que algunos conceptos básicos sobre genómica que usaremos durante el curso se encuentren disponibles. El foco estará en diferentes metodologías para el estudio de los genomas, muchas de las cuales usaremos en distintos ejemplos y secciones del curso. Poco esfuerzo se hace por explicar las bases de la herencia y la estructura y organización de los genomas eucariotas y procariotas, las cuales se asumen conocidas.

Los organismos en nuestro planeta derivan todos de un organismo ancestral, desconocido y, por lo tanto, es de esperar que sus genomas compartan determinadas características. Sin embargo, al mismo tiempo podemos observar una enorme diversidad entre ellos, por lo que resulta fundamental entender los procesos que llevan a esas diferencias. Dichos procesos operan sobre el componente heredable de los organismos, su material genético. Este componente está formado por una secuencia ordenada de nucleótidos, por lo cual es común referirse a este coloquialmente como 'secuencia(s)' en el ámbito de la genómica y las disciplinas afines. Conocer las secuencias de los genomas de diferentes organismos y compararlas para determinar las diferencias entre estas es, por lo tanto, el primer paso por dar para comprender los procesos que llevaron a dichas diferencias. La secuenciación de genomas es una historia que comenzó en las últimas décadas del siglo pasado y desde ese momento no ha dejado de crecer, en ciertos momentos a velocidad vertiginosa. En particular, la aparición de las denominadas 'tecnologías de segunda y tercera generación' ha permitido pasar de la secuenciación de un genoma por especie a la secuenciación de cientos o miles de individuos de una misma especie. De esta forma, en la actualidad se está logrando caracterizar mucho mejor la variabilidad biológica existente.

La diversidad existente a nivel de organismos macroscópicos es evidente. El ser humano interacciona con la flora y la fauna desde sus inicios como especie. Entender, preservar e influir sobre esta diversidad biológica es lo que ha permitido en última instancia el desarrollo de las civilizaciones. Resulta intuitivo que enriquecer nuestra comprensión de esta diversidad con aportes desde la genómica pueda resultar valioso. Por otra parte, aunque invisibles los microorganismos juegan un papel relevante en casi todos los ecosistemas a diferentes escalas: desde bosques, praderas, ríos y mares hasta los ecosistemas internos (por ejemplo,

[1] Frase que se atribuye a este autor en el dominio público, pero que no cuenta con una referencia concreta. Podría traducirse como «Estoy seguro de que el universo está lleno de vida inteligente. Solo que ha sido demasiado inteligente como para venir aquí».

aparato digestivo) y externos (piel) de los organismos. A pesar de su importancia, la diversidad microbiana apenas ha sido explorada, fundamentalmente debido a las dificultades asociadas al cultivo de los microorganismos, muchos de ellos directamente no-cultivables. En este sentido, también la genómica ha jugado un rol destacado en los avances que se han producido en este campo en los últimos años. La metagenómica, la metagenómica funcional y la metatranscriptómica han contribuido a entender la composición de los ecosistemas microbianos, las funciones de los genes involucrados y la regulación de su expresión, respectivamente. Mencionaremos cada una de estas aproximaciones en secciones posteriores de este capítulo.

Aunque a primera vista la informática y la biología puedan parecer áreas de estudio muy alejadas (¿qué se aleja más de la vida que una fría máquina?), veremos que la obtención y el procesamiento masivo de datos biológicos han permitido enriquecer de forma importante la perspectiva que tenemos sobre los organismos vivos y la transmisión de información de generación en generación, tanto en el plano individual como poblacional. A lo largo del libro iremos profundizando en esta perspectiva, empleando recursos informáticos para modelar procesos de la genética poblacional y cuantitativa, de forma de hacer un uso enriquecedor de las tecnologías disponibles en la actualidad.

OBJETIVOS DEL CAPÍTULO

☐ Abordar el fenómeno de la variabilidad genética, discutiendo el uso de métricas que nos permiten cuantificarla introduciendo tecnologías con las que se identifican variantes genéticas en poblaciones.

☐ Obtener una perspectiva de las distintas áreas de estudio de la genómica, incluyendo: - el estudio de la composición nucleotídica de los genomas, - el estudio evolutivo de los organismos, - el análisis de los patrones de expresión génica.

☐ Repasar conceptos básicos de genética y biología molecular, ilustrando en cada caso cómo el uso de herramientas estadísticas y computacionales enriquece nuestra perspectiva de los fenómenos biológicos.

1.1 Variabilidad genética

Todos los organismos presentan diferencias observables (*fenotípicas*) entre ellos. Muchas de estas son producto del desarrollo en determinados ambientes (específicas para cada individuo), mientras que otras se corresponden con las diferencias en el genoma. Excepto los clones (copias idénticas de un mismo individuo), todos los organismos son de alguna manera diferentes entre sí. Aún entre gemelos es posible postular que la epigenética[2] puede jugar un

[2]Se denomina como 'epigenético' todo cambio heredable que no tiene como base un cambio a nivel de nucleótido en la secuencia de ADN. A modo de ejemplo, cambios en el estado de metilación de un nucleótido pueden influir en el estado de compactación de la cromatina y así en la expresión de algunos genes; esto puede resultar en cambios fenotípicos heredables, aun cuando en ningún momento se haya reemplazado un nucleótido por otro.

rol diferente en su desarrollo. Estas diferencias a nivel genético o genómico son las que nos resultan fundamentales cuando hablamos de mejorar una especie de nuestro interés o en cualquier instancia en la que busquemos seleccionar algunos individuos de una población.

Muchas otras veces estamos más interesados en entender las diferencias entre especies y el rol que juegan las proporciones de sus individuos en la dinámica de un ecosistema. Claramente, la robustez de un ecosistema suele estar asociada al balance de diversas especies, cada una jugando roles diferentes. Pero ese balance no solo depende de la presencia de algún miembro de cada especie, sino de las proporciones que esas especies ocupan en el ecosistema. Tener índices que nos permitan cuantificar y comparar la diversidad existente en diferentes escenarios resulta por lo tanto útil. Tradicionalmente, en ecología se definen tres tipos de medidas de diversidad: la *diversidad alfa*, la *diversidad beta* y la *diversidad gamma*, de acuerdo con la escala del fenómeno estudiado.

Se llama *diversidad alfa* a la que describe la diversidad de especies dentro de una comunidad[3] a una escala pequeña o local. Esta escala es normalmente del tamaño de un ecosistema y es a la que generalmente nos referimos al hablar de una *zona*.

La *diversidad beta* describe la diversidad de especies entre dos comunidades o ecosistemas. Involucra una escala mayor, ya que incluye ecosistemas (que correspondían a la diversidad alfa) y en general suele haber alguna distinción geográfica o barrera importante entre las comunidades de referencia.

La *diversidad gamma* se refiere a una escala de estudio mucho mayor (por ejemplo, un bioma), donde se compara la diversidad de especies entre muchos ecosistemas.

Aunque la diversidad biológica puede cuantificarse de muchas maneras diferentes, los dos factores principales que se tienen en cuenta para medir la diversidad son la *riqueza* y la *uniformidad*. La riqueza representa el número de tipos (especies) diferentes entre todos los que identificamos en el área de interés. Sin embargo, nuestra intuición nos dice que la uniformidad representa también una parte importante de la diversidad. Es decir, si en un ambiente tenemos un individuo de cada una de cuatro especies y 996 individuos de una quinta especie, claramente identificaremos una menor diversidad respecto a otro ambiente que tiene 200 individuos de cada una de las especies. En general, la uniformidad compara la similitud del tamaño de la población de cada una de las especies presentes.

Para cada uno de estos tipos de diversidad se han desarrollado índices que permiten obtener estimaciones de su importancia relativa a otros sistemas. Por ejemplo, para la diversidad alfa el índice más sencillo de entender es la entropía de Shannon (llamada así por el matemático Claude Shannon, que la propuso para un problema de transmisión de información), que mide la incertidumbre al asignar la identidad de un individuo elegido al azar de la población.

[3] En ecología se define a una *comunidad* como un conjunto de poblaciones de dos o más especies que conviven al mismo tiempo en una misma área geográfica. Un *ecosistema* está compuesto de una comunidad y del entorno abiótico que la comprende. Ambos conceptos son amplios y pueden referir a unidades ecológicas de tamaños y características diversas.

Si le llamamos p_i a la proporción (frecuencia relativa) de individuos de la especie i, de un total de k especies identificadas en el ecosistema, entonces el **índice de Shannon** se calcula como

$$H' = -\sum_{i=1}^{k} p_i \ln p_i \tag{1.1}$$

Es decir, cada frecuencia es multiplicada por su logaritmo y sumada, con el resultado final multiplicado por -1 ya que cualquiera de los logaritmos será menor o igual a cero (porque para todas las especies $0 \leqslant p_i \leqslant 1$). En el caso de una sola especie $\ln p_i = \ln 1 = 0$, por lo que la diversidad será cero. Cuanto más homogéneo sea el reparto de las especies, mayor será el índice de Shannon. En particular, el máximo del índice de Shannon se dará cuando $p_i = \frac{N/k}{N} = \frac{1}{k}\ \forall\ i$, en cuyo caso

$$H_{max} = -\sum_{i=1}^{k} \frac{1}{k} \ln \frac{1}{k} =$$

$$= k\ \frac{1}{k}[\ln 1 - \ln k] = -[0 - \ln k] \therefore$$

$$H_{max} = \ln k$$

Es decir, el máximo del índice será el logaritmo del número de especies, lo que ocurre cuando todas las especies tienen el mismo número de integrantes. Podemos ver entonces que el índice de Shannon no solo nos permite comparar la diversidad biológica de forma cuantitativa, sino que además nos permite predecir bajo qué condiciones esta se maximiza. Lo que es más, esta predicción se corresponde con nuestra intuición de que una mayor diversidad debe reflejar un reparto más equitativo de los individuos para las diferentes especies (concepto de uniformidad mencionado anteriormente).

Resulta claro que los valores que se pueden obtener para el índice de Shannon dependerán del número de especies presentes (k) en el ecosistema estudiado. En este sentido, resulta útil un índice relacionado directamente al índice de Shannon es el propuesto por Pielou (1969), que estandariza el índice de Shannon por la riqueza, lo que lo deja como un índice de uniformidad (*evenness*, en inglés) que se calcula como

$$E = \frac{H'}{H_{max}} = \frac{H'}{\ln k} \tag{1.2}$$

Así, es posible tener una noción más clara respecto a la diversidad alfa presente en el ecosistema estudiado: cuanto más se acerca este índice E a 1, más cerca se estará del máximo teórico según el índice de Shannon.

Otro índice popular es el **índice de Simpson**. El índice de Simpson (D) mide la probabilidad de que dos individuos seleccionados al azar de una muestra pertenezcan a la misma especie o categoría.

Es decir, si n_i representa el número de individuos de la especie o categoría i y N el número total de individuos, es decir $N = \sum_i n_i$, entonces

$$D = \sum_{i=1}^{k} \frac{n_i(n_i - 1)}{N(N-1)} \tag{1.3}$$

La intuición de este índice es que si los individuos se concentran en pocos tipos diferentes el valor de D será más alto, es decir lo contrario a lo que esperaríamos para un índice de diversidad. De hecho, con este índice 0 representa una diversidad infinita y 1, ninguna diversidad. Es decir, cuanto mayor sea el valor de D, menor será la diversidad. Para superar este problema, se suele restar a 1 el valor de D, es decir se reporta $1 - D$. Otra alternativa es reportar el recíproco de D, es decir D. Cuando el número de individuos es muy grande, entonces en general $\frac{n_i - 1}{N-1} \to \frac{n_i}{N}$, por lo que

$$D = \sum_{i=1}^{k} \frac{n_i(n_i - 1)}{N(N-1)} = \sum_{i=1}^{k} \frac{n_i}{N} \frac{(n_i - 1)}{(N-1)} \approx \sum_{i=1}^{k} \frac{n_i}{N} \frac{(n_i)}{(N)} \therefore$$

$$D \approx \sum_{i=1}^{k} \frac{n_i^2}{N^2}$$

Notemos que hasta ahora no se han tenido en cuenta posibles errores al momento de determinar el número de especies *k* presentes en una muestra (por ejemplo, por falta de muestreo de organismos de una especie presente en el ecosistema en números muy bajos).

El **índice de cobertura de Good** es usado a veces como estimador de diversidad alfa y se calcula mediante la siguiente ecuación

$$C = 1 - \frac{F_1}{N} \tag{1.4}$$

donde F_1 es el número de OTU (*Operational Taxonomic Units*; unidades taxonómicas operacionales, en inglés) en la muestra con un solo representante. OTU es una forma de expresar una unidad taxonómica particular a un nivel de la taxonomía determinado. Por ejemplo, en algunos casos las OTU pueden representar especies, pero otras veces representan niveles más altos o más bajos de la taxonomía (de ahí el término *operacional*). La idea del índice es capturar de alguna manera la probabilidad de perder OTU debido a un tema de muestreo. Es decir, aquellas OTU que no fueron secuenciadas pero que se encontraban presentes, las cuales se hallarían de haberse realizado muestreos con tamaños muestrales mucho mayores.

La diversidad juega un papel fundamental en la resiliencia de los sistemas, tanto la diversidad entre especies como dentro de las especies. Por ejemplo, la pérdida de diversidad asociada a la revolución verde entre las décadas del sesenta y del ochenta del siglo XX representa una pérdida irrecuperable de patrimonio genético desarrollado durante miles y millones de años. Lamentablemente, es poco probable que la biotecnología pueda recuperar en breve lo que la naturaleza desarrolló durante millones de años, por lo que resulta trascendental pensar en conservar la diversidad en nuestro planeta.

Ejemplo 1.1

Luego de un muestreo realizado a partir de metagenómica en dos ambientes diferentes, la distribución de especies OTU en ellos es la siguiente:

Especie	Ambiente 1	Ambiente 2
OTU 1	60	1
OTU 2	150	1
OTU 3	0	297
OTU 4	390	1
OTU 5	400	700
Total	1000	1000

Calcular los diferentes índices para los dos ambientes y discutir los resultados.

Como el total de individuos (*reads*) obtenidos en ambas muestras es 1000, alcanza con dividir entre ese número para tener la frecuencia relativa. Por lo tanto, utilizando la ecuación (1.1), el índice de Shannon para el primer ambiente es igual a

$$\begin{aligned} H_1' &= -\sum_{i=1}^{k} p_i \ln p_i \\ &= -[0{,}06 \times \ln 0{,}06 + 0{,}15 \times \ln 0{,}15 + 0{,}39 \times \ln 0{,}39 + 0{,}4 \times \ln 0{,}4] \\ &\therefore H_1' = 1{,}1871 \end{aligned}$$

mientras que para el segundo ambiente el índice de Shannon es

$$\begin{aligned} H_2' &= -\sum_{i=1}^{k} p_i \ln p_i \\ &= -[0{,}001 \times \ln 0{,}001 + 0{,}001 \times \ln 0{,}001 + 0{,}297 \times \ln 0{,}297 \\ &\quad + 0{,}001 \times \ln 0{,}001 + 0{,}7 \times \ln 0{,}7] \\ &\therefore H_2' = 0{,}63096 \end{aligned}$$

Notemos que en el Ambiente 1 no se encontró la especie OTU 3, por lo que no se incluye en los cálculos (de hecho, $\ln 0 = -\infty$). De acuerdo con esto, según el índice de Shannon el Ambiente 1 es más diverso que el Ambiente 2. Para medir la uniformidad utilizando el índice propuesto por Pielou (1969), primero tenemos que calcular el máximo posible para el índice de Shannon, que a partir de la ecuación (1.1) sabemos que es igual al logaritmo del número de OTU, es decir $H_{max_1} = \ln k = \ln 4 = 1{,}3863$ y $H_{max_2} = \ln k = \ln 5 = 1{,}6094$. Por lo tanto, aplicando la ecuación (1.2), tenemos que el índice de uniformidad para los dos ambientes será

$$E_1 = \frac{H_1'}{H_{\max_1}} = \frac{1{,}1871}{1{,}3863} = 0{,}8563$$

$$E_2 = \frac{H_2'}{H_{\max_2}} = \frac{0{,}63096}{1{,}6094} = 0{,}3920$$

lo que habla de una uniformidad mucho menor en el Ambiente 2.

Para el índice de Simpson vamos a utilizar la ecuación (1.3), por lo que para el Ambiente 1 su valor será

$$\begin{aligned} D_1 &= \sum_{i=1}^{k} \frac{n_i(n_i-1)}{N(N-1)} \\ &= \left[\frac{60(59)+150(149)+390(389)+400(399)}{1000\times 999}\right] \\ &= 0{,}3375 \end{aligned}$$

mientras que para el Ambiente 2 su valor será

$$\begin{aligned} D_2 &= \sum_{i=1}^{k} \frac{n_i(n_i-1)}{N(N-1)} \\ &= \left[\frac{1(0)+1(0)+297(296)+1(0)+700(699)}{1000\times 999}\right] \\ &= 0{,}5778 \end{aligned}$$

Si utilizamos el criterio de reportar $1-D$, entonces $1-D_1 = 1-0{,}6625$, mientras que $1-D_2 = 1-0{,}4222$. Si utilizamos el criterio del recíproco, entonces $1/D_1 = 2{,}9626$ y $1/D_2 = 1{,}7307$, por lo que se aprecia claramente que la diversidad en el **Ambiente 1** es mayor que la del **Ambiente 2** de acuerdo con este índice. Finalmente, de acuerdo con la ecuación (1.4), teniendo en cuenta que en el primer ambiente no hay ningún OTU con un único representante mientras que en el segundo hay 3, el **índice de cobertura de Good** para los dos ambientes sería

$$C_1 = 1 - \frac{0}{1000} = 1$$
$$C_2 = 1 - \frac{3}{1000} = 0{,}997$$

PARA RECORDAR

- Tradicionalmente, en ecología se definen tres tipos de medidas de diversidad de acuerdo con la escala del fenómeno estudiado:
  - Se llama ***diversidad alfa*** a la que describe la diversidad de especies dentro de una comunidad a una escala pequeña o local. Esta escala es normalmente del tamaño de un ecosistema y es a la que generalmente nos referimos al hablar de una zona.
  - La ***diversidad beta*** describe la diversidad de especies entre dos comunidades o ecosistemas. Involucra una escala mayor, ya que incluye ecosistemas (que correspondían a la diversidad alfa) y en general suele haber alguna distinción geográfica o barrera importante entre las comunidades de referencia.
  - La ***diversidad gamma*** se refiere a una escala de estudio mucho mayor (por ejemplo, un bioma), donde se compara la diversidad de especies entre muchos ecosistemas.

PARA RECORDAR

- Aunque la diversidad biológica puede cuantificarse de muchas maneras diferentes, los dos factores principales que se tienen en cuenta para medir la diversidad son la ***riqueza*** **y la** ***uniformidad***. La ***riqueza*** representa el número de tipos (especies) diferentes entre todos los que identificamos en el área de interés, mientras que la ***uniformidad*** compara la similitud del tamaño de la población de cada una de las especies presentes.
- Para la diversidad alfa, el índice más sencillo de entender es la **entropía de Shannon**, que mide la incertidumbre al asignar la identidad de un individuo elegido al azar de la población. Si le llamamos p_i a la proporción (frecuencia relativa) de individuos de la especie i, de un total de k especies identificadas en el ecosistema, entonces el **índice de Shannon** se calcula como $H' = -\sum_{i=1}^{k} p_i \ln p_i$. Cuanto más homogéneo sea el reparto de las especies, mayor será el índice de Shannon. El índice máximo será cuando todas las especies tienen el mismo número de integrantes.
- El **índice de Simpson** (D) mide la probabilidad de que dos individuos seleccionados al azar de una muestra pertenezcan a la misma especie o categoría. Es decir, si n_i representa el número de individuos de la especie o categoría i y N el número total de individuos (por ejemplo, $N = \sum_i n_i$), entonces $D = \sum_{i=1}^{k} \frac{n_i(n_i-1)}{N(N-1)}$. La intuición de este índice es que si los individuos se concentran en pocos tipos diferentes el valor de (D) será más alto, al contrario de lo que esperaríamos para un índice de diversidad. Para superar este problema, se suele restar a 1 el valor de (D) (se reporta $1 - D$).

1.1.1 Detectando la variabilidad: diferentes técnicas

En las últimas décadas han ganado una enorme relevancia un conjunto de disciplinas con base en avances tecnológicos que permiten la obtención masiva de información biológica global para una condición dada. Esto se logra a través de la automatización de diversos protocolos de bioquímica y biología molecular, tales como la secuenciación de nucleótidos o péptidos provenientes de una muestra biológica. Estas disciplinas son identificables por la utilización del sufijo *-ómica* en su nombre. La genómica es una de estas disciplinas, focalizándose en el estudio del genoma de diferentes organismos a través de las tecnologías de secuenciación masiva. Existen también la *transcriptómica* (estudio global de los transcriptos expresados por un organismo o conjunto de células), la *proteómica* (estudio de las proteínas expresadas en una condición dada) o la *metabolómica* (estudio del perfil metabólico), entre otros. Este enfoque de análisis también se puede aplicar para estudiar varios organismos a la vez (por ejemplo, al tomarse muestras ambientales), surgiendo así un subconjunto de disciplinas *meta-*. A modo de ejemplo, la *metagenómica* se encarga del ensamblado y estudio de genomas microbianos en muestras provenientes de diferentes lugares. En la sección «Metagenómica y metatranscriptómica» se detalla un esquema típico de trabajo en metagenómica. Como es de esperar, estas técnicas de obtención masiva de datos han impactado de forma notable nuestro entendimiento de la variabilidad genética entre organismos.

Para entender la variabilidad genética es necesario determinar algún marcador genético cuyas variantes podamos detectar en muchos individuos (por ejemplo, obteniendo las secuencias para el marcador en esta muestra). Un ejemplo de estudios iniciales de diversidad genética es la utilización de grupos sanguíneos, lo que veremos en el próximo capítulo. Hasta hace algunos años la búsqueda de marcadores fiables de diversidad genética representaba un enorme desafío, y se desarrollaron una plétora de métodos para obtener marcadores moleculares de diversidad. Tanto el trabajo como los costos eran enormes, lo que dificultaba el desarrollo de estudios a gran escala en este sentido. Afortunadamente, a fines del siglo XX las técnicas de secuenciación masiva habían avanzado lo suficiente como para incluir centenas de individuos en los estudios. Más aún, el desarrollo de técnicas como los **microarrays** posibilitaron la automatización de muchos análisis (en especial los de expresión), pero también resultaron fundamentales a la hora de abaratar los costos para estudios masivos de diversidad.

Los microarrays son un arreglo de **sondas moleculares** colocadas en una placa. Originalmente contenían apenas unas decenas de sondas; la robotización y la aplicación de tecnologías que se usan en la fabricación de semiconductores permitieron alcanzar densidades del orden del millón de sondas por *array* (placa). Una sonda molecular, en este caso, es una secuencia específica para un gen (o una región de un gen) y que es, idealmente, única en el genoma de referencia. Esta sonda se encuentra fija en la placa, y la hibridación del **ADN** o el **ADN copia** (en los estudios de transcriptómica) de un individuo es acoplada a algún tipo de señal (por ejemplo, emisión de luz en determinada frecuencia al ser iluminada por un **láser**). La presencia de señal permite se asocia, por lo tanto, a la presencia del ADN de interés en la muestra proveniente del individuo.

Esta tecnología permite, por ejemplo, identificar si la secuencia presente en el organismo es de un tipo u otro en el caso de los **SNP** (polimorfismos de un solo nucleótido). En el caso de las especies diploides, es posible identificar el genotipo para un organismo. Para conocer la diversidad del genoma humano es usual utilizar chips de densidad media (10-60k de SNP) o de alta densidad (hasta un millón de SNP). En especies de interés comercial existen diferentes tipos de alternativas, desde arrays de baja densidad (apenas unos pocos miles de SNP), media, hasta los de alta densidad, como el **BovineHD** que comprende más de 777.000 SNP, compatible con todas las razas bovinas lecheras o de carne.

A pesar del enorme éxito de los microarrays para estudiar la diversidad genética en diversas poblaciones de diferentes especies, los mismos padecen de un problema fundamental: para crear una sonda para medir variabilidad hay que conocer las secuencias de regiones que contengan variabilidad. Es decir, para estudiar la variabilidad mediante esta técnica hay que tener alguna información previa de qué regiones son variables. Esto se debe a que la suma de los tamaños de las sondas es extremadamente pequeña en relación con el largo total de los genomas. Por ejemplo, en genoma humano un chip de alta densidad cubre 1 millón de posiciones de las más de 3000 millones que tiene el genoma (~0,0003 %). Para elegir bien ese millón de posiciones hay que tener muy buena información previa: si se eligen posiciones que casi no varíen en la población, se obtendrá muy poca información sobre la variabilidad genética en las poblaciones estudiadas. Además, otro problema derivado de los microarrays y de este sesgo en los marcadores a incluir es que luego de diseñado el array solo va a captar variabilidad en las posiciones del genoma para el que fue diseñado. Si una nueva población aparece con variabilidad importante en otras regiones, el array no va a ser capaz de detectarlo. Por ejemplo, en genoma humano, hasta hace poco tiempo los arrays poseían un importante sesgo eurocéntrico, porque la mayor parte de la información previa

venía de Europa. Esto dejaba fuera sitios de alta variabilidad en poblaciones asiáticas o amerindias, por ejemplo. Estudios que utilizaran dichos arrays terminarían determinando que estas poblaciones tenían menos variabilidad genética que la real.

Afortunadamente, las **tecnologías de secuenciación masiva** que se desarrollaron a comienzos del siglo XXI cambiaron definitivamente el panorama, ya que permiten obtener las secuencias de decenas o cientos de individuos a precio y tiempos razonables. Como referencia, el esfuerzo global para obtener la secuencia del primer genoma humano llevó más de diez años a decenas de laboratorios repartidos en el mundo, y costó más de 3000 millones de dólares. Actualmente, un genoma humano tarda en secuenciarse menos de una semana (el récord mundial, establecido recientemente, es de 5 horas y 2 minutos) y cuesta en promedio unos mil dólares (sin incluir los costos del análisis posterior). Más aún, si bien las tecnologías de segunda generación (**Roche 454**, **Illumina** y **SOLiD**) tuvieron un éxito enorme, su flexibilidad era algo limitada y en particular era muy difícil reducir los tiempos de secuenciación, además de tener ciertos sesgos específicos en los resultados. En la segunda década de este siglo comenzaron a aparecer y consolidarse otras nuevas tecnologías, que se dieron en llamar 'de tercera generación', y que incluyen los secuenciadores **IonTorrent**, los de **Oxford Nanopore Technologies** y los **PacBio.** En conjunto, estas tecnologías —sumadas a **Illumina**, que es la única de las de segunda generación que sobrevive con mucho éxito— permiten tener una flexibilidad enorme en tamaños y costos de equipamientos y reactivos, en sesgos y distribución de errores, así como en la velocidad de secuenciación.

1.1.2 Metagenómica y metatranscriptómica

En muchos casos nos interesa determinar la estructura de alguna comunidad biológica y estudiar su evolución en el tiempo, o simplemente comparar esa estructura en diferentes condiciones. Por ejemplo, la microbiota ruminal es fundamental para los procesos que determinan la degradación de la materia orgánica que el animal recibe, estando asociada también a qué productos de esta degradación recibirá el animal y en qué proporciones. ¿Qué variables afectan más la composición de la microbiota? ¿El pH ruminal? ¿El porcentaje de materia seca? ¿La proporción de fibra? En el plano de la salud, ¿qué microorganismos se encuentran presentes en la ubre sana y qué microorganismos en la glándula de un animal con mastitis? ¿La infección es a causa de un microorganismo dominante o el patógeno no precisa ser el que domine el ecosistema ruminal? ¿El rol del patógeno es directo sobre la salud del animal o su rol se limita a desestabilizar la ecología de la glándula?

Para poder responder a cualquiera de estas preguntas es necesario poder caracterizar la comunidad microbiana presente. A diferencia de lo que ocurre con especies de plantas, hongos y animales que generalmente podemos percibir y contar a simple vista, con los microorganismos nos ocurre que necesitamos identificarlos de formas más indirectas. Si bien usando el microscopio (con algo de suerte) podríamos ser capaces de reconocer el género de algunas bacterias presentes en una muestra, difícilmente seamos capaces de reconocer a la mayoría. Seríamos mucho menos capaces de reconocer las especies y cepas de estas bacterias, y muchísimo menos capaces de cuantificar la proporción de cada grupo bacteriano. Haciendo un paralelismo, para entender la comunidad vegetal de un bosque no alcanza con determinar la presencia de al menos un ejemplar de una especie: es fundamental entender la proporción que dicha especie ocupa en el ecosistema.

Como ya mencionamos, el hecho de que todos los organismos en nuestro planeta «desciendan» del mismo organismo ancestral, a veces conocido como **LUCA** (*Last Unknown Common Ancestor*, en inglés), nos sugiere que sería de esperar encontrar ciertas regularidades en la información genética de estos. Es decir, diferentes formas de vida son el producto de modificaciones a formas de vida preexistentes, al menos hasta llegar a LUCA, por lo que de alguna manera todos los seres vivos comparten ciertas características. Las diferencias entre especies, como sabemos, se corresponden en gran medida a cambios en su genoma, tanto a nivel de mutaciones puntuales como de inserciones y deleciones, duplicaciones de regiones del genoma o aun re-arreglos cromosómicos. La clave para identificar especies es ubicar secuencias lo suficientemente conservadas como para que deban encontrarse en todos los organismos de interés, pero al mismo tiempo con suficientes cambios entre especies como para poder identificarlas en forma unívoca.

Dentro de las secuencias con estas características, la maquinaria vinculada a la traducción de la información es ciertamente privilegiada ya que constituye un elemento clave para cualquier forma de vida: la capacidad de traducir la información genética en los efectores, que suelen ser las proteínas. La traducción de la información implica una maquinaria tan compleja que resulta completamente implausible su surgimiento de forma independiente en diferentes linajes de la vida, más aún dado que cualquier especie ya hubiese precisado de una para existir. Además, debido a su enorme complejidad se trata de una maquinaria muy ajustada y donde es posible introducir muy pocas variaciones sin que la maquinaria deje de funcionar completamente, en particular porque el ***fitness***[4] de los organismos tiene, en muchos casos, una vinculación directa con la tasa de producción de las proteínas. Un caso de particular importancia es el **ARN ribosomal** (**ARNr**). Este tipo de marcador molecular está constituido por una serie de moléculas de ARN que dependen de su estructura espacial para poder cumplir su rol. Esto incluye apareamientos Watson-Crick entre regiones complementarias, pero también bases modificadas post-transcripcionalmente para asegurar la estabilidad de la molécula y la fidelidad de la traducción.

En procariotas, el **ARNr 16S** es una molécula usada de forma extensiva para identificar el género y especie de los organismos. De hecho, la propuesta para un sistema de clasificación de la vida de Woese, Kandler y Wheelis (1990), representada en la figura 1.1, se basa en las diferencias observadas en la molécula de **16S** y su equivalente en eucariotas, el **ARNr 18S**. De acuerdo con esta forma de organizar las diferentes especies, existen tres grandes dominios de la vida: Bacteria, Archaea y Eukarya.[5] Es importante notar que, a diferencia de la taxonomía basada en similitud morfológica, la clasificación propuesta por Woese, Kandler y Wheelis (1990) se basa en similitud en las secuencias genómicas, donde de acuerdo con

[4] Diferentes individuos tenderán a reproducirse con mayor o menor éxito debido a una variedad de factores tales como la capacidad de sobrevida en un entorno dado o la fecundidad, entre otros. En la teoría evolutiva, este concepto se ve representado por el término *fitness*, que refiere a la contribución reproductiva de un individuo a su próxima generación. Así, diremos que individuos con un mayor fitness tienden a dejar mayor descendencia en la próxima generación. Algunos autores hacen referencia al fitness de un individuo, mientras que otros refieren al *fitness* de un genotipo particular o incluso de un alelo.

[5] Interesantemente, la metagenómica ha permitido el descubrimiento reciente de una diversidad arqueana antes desconocida, lo cual ha llevado a una fuerte controversia en esta área de la filogenética contemporánea. En particular, el descubrimiento en 2015 de un filo arqueano denominado Asgard y el análisis en conjunto de diferentes marcadores moleculares ha aportado evidencia congruente con una «hipótesis de dos dominios» para el árbol de la vida. Según esta hipótesis el dominio Eukarya sería en realidad un subgrupo del dominio Arquea, fuertemente emparentado con el filo Asgard. Más allá de qué hipótesis posea mayor respaldo, esta controversia ilustra hasta qué punto nuevas evidencias provenientes de la genómica y la metagenómica aportan a debates de primer nivel en la teoría biológica; otros ejemplos de relevancia teórica y práctica son mencionados más adelante en esta misma sección.

determinados modelos se pueden asignar relaciones de proximidad y distancia entre secuencias pertenecientes a diferentes organismos. En particular, la reconstrucción filogenética se realiza generalmente a partir de secuencias de genes *ortólogos*,[6] en este caso el **ARN ribosomal16S**, por ejemplo. Si bien desde el punto de vista morfológico los miembros del dominio **Archaea** no son, aparentemente, muy distintos de los miembros de **Bacteria**, desde el punto de vista evolutivo poseen grandes diferencias que los hacen acreedores de su propio dominio.

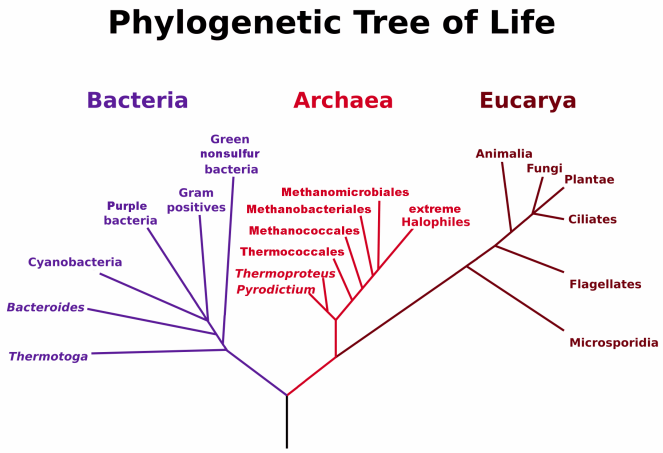


Figura 1.1. Árbol filogenético de los tres dominios biológicos, Bacteria, Archaea y Eukarya. Con una barra vertical se representa la existencia de **LUCA**, el último ancestro común, desconocido. (Por Maulucioni - Trabajo Propio, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24740337).

Las técnicas de secuenciación masiva implicaron un salto enorme para la caracterización de poblaciones microbianas, ya que permiten la secuenciación de miles y hasta millones de fragmentos de secuencias en muy poco tiempo y en forma paralela. La idea más básica consiste en secuenciar una determinada secuencia o región que se encuentre en todos los organismos de interés y que presente suficiente variabilidad, como es el caso del **ARNr 16S**. Esquemáticamente, los pasos serían los siguientes:

1. Determinar el grupo de organismos de interés, así como la molécula y región a secuenciar. Las regiones hipervariables **V3** y **V4** del **ARNr 16S** son buenas candidatas iniciales para procariotas. El gen tiene una longitud aproximada de 1500 pb y contiene nueve regiones variables intercaladas entre regiones conservadas. Qué región del ARNr 16S secuenciar es un área de debate, y puede variar dependiendo de cosas como los objetivos experimentales, el diseño y el tipo de muestra, aunque las regiones **V3** y **V4** son ampliamente utilizadas.

[6] En teoría evolutiva se dice que dos caracteres son *ortólogos* si derivan de un único ancestro común a dos especies. A modo de ejemplo, los genes codificantes para mioglobina de humano y chimpancé son ortólogos, ya que se infiere que el ancestro común a ambas especies poseía una única copia para este gen; una vez se dio el evento de especiación, pasó a existir una copia para cada especie.

2. Extraer los ácidos nucleicos de interés y amplificar por PCR las regiones de interés. Los protocolos permiten en general utilizar *barcodes*, que son secuencias específicas que se agregan a las secuencias (un barcode para cada muestra), lo que permite correr en paralelo hasta 96 muestras en una misma «corrida» del secuenciador.
3. Secuenciar las muestras preparadas de acuerdo con el protocolo. Existen diferentes tecnologías y en distintos tamaños, lo que permite secuenciar desde unas poquitas muestras hasta centenas de ellas a un costo razonable y en tiempos razonables. Las distintas tecnologías poseen problemas diferentes, pero en general no introducen sesgos importantes en la composición estimada de la muestra.
4. Control de calidad y limpieza de las secuencias obtenidas. Es una serie de procesos a nivel bioinformático que tienen por objetivo eliminar las secuencias de los adaptadores que se agregaron durante la preparación de las bibliotecas, separar las secuencias de acuerdo con el correspondiente barcode y calcular una serie de parámetros que indican la calidad de la corrida y de cada una de las muestras.
5. Agrupamiento de secuencias y mapeo contra bases de datos con secuencias identificadas. Se trata del corazón de la idea y consiste en que cada secuencia o fragmento producido durante la secuenciación debe tener uno idéntico depositado en una base de datos con secuencias de organismos. Hay diferentes estrategias respecto a cómo agrupar los fragmentos antes de «mapearlos» contra las bases de datos y diferentes bases de datos de organismos, pero esto ya es materia específica de la bioinformática.
6. Una vez mapeadas las secuencias obtenidas, es decir que a cada una le asignamos un **taxón** determinado (puede llegar a ser una especie, aunque a veces la identificación queda a un nivel más alto, por ejemplo, género), entonces alcanza con contar cuántas veces aparece cada organismo en una muestra (es decir, cuántos fragmentos secuenciados mapean al mismo taxón). Esto nos permite organizar la información en una tabla que puede tener las muestras en filas y los taxa en las columnas.
7. Análisis de la tabla de conteos. Por lo general la tabla generada de conteos tiene diferencias importantes en el número de secuencias obtenidas por muestra ya que no se obtiene la misma cantidad de ácidos nucleicos en cada muestra a pesar de los esfuerzos. Para la parte de exploración suelen usarse análisis exploratorios multivariados, como el **escalado multidimensional** (MDS), una técnica con cierta relación al PCA en su versión más clásica (que se conoce como **Principal Coordinates Analysis**, PCoA). La idea es reducir la dimensionalidad del problema de tal forma que se pueda representar en pocas dimensiones con poca pérdida de información. En el mejor de los casos, después es posible correlacionar directamente las posiciones de las muestras en las principales dimensiones con factores externos, tales como el pH, la temperatura, ubicación geográfica o cualquier otro factor que sea relevante. En la figura 1.2 se observa la representación gráfica de los dos primeros ejes en un análisis MDS realizado como control de calidad en un experimento de **transcriptómica RNA-seq** con ***Leptospira***

biflexa. Se observa que las muestras se agrupan de acuerdo con la condición biológica y que los dos factores relevantes desde el punto de vista biológico (fase de crecimiento y tiempo) son las que separan las muestras entre sí.

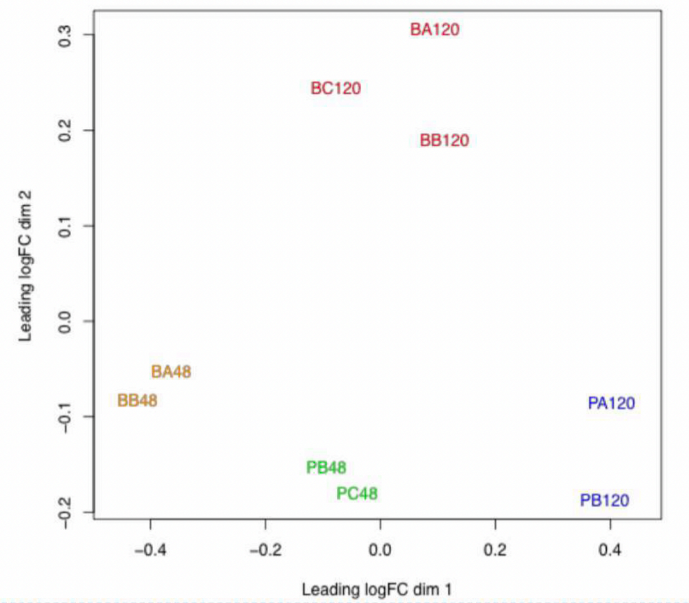


Figura 1.2. Escalado multidimensional en un experimento de transcriptómica de *Leptospira biflexa*. El experimento incluía dos factores, la fase de crecimiento (biofilm y planctónica) y dos tiempos (48 y 120 horas), en 3 réplicas (A, B y C) por condición. Alguna de las réplicas fue descartada por motivos de calidad (figura tomada de Iraola et al. 2016).

Con las tecnologías de segunda generación (Roche **454**, Illumina, SOLiD) era y aún es común secuenciar fragmentos, como vimos recién. Sin embargo, con tecnologías de tercera generación como la de **Oxford Nanopore Technologies** es común secuenciar toda la molécula de **16S** ya que el largo de la lectura no está determinado *a priori* (como sí en Illumina) y es normal obtener **reads** (lecturas del secuenciador) de miles o aun decenas de miles de pares de bases. Por otro lado, existen otras formas de metagenómica que no pasan por la secuenciación de una región o de una secuencia en particular. En **metagenómica funcional** la idea es secuenciar todo el **ADN** presente en la muestra, en lugar de determinadas secuencias en particular. Esto permite capturar no solo la variabilidad de organismos presentes en la muestra, sino que también identificar secuencias de ellos que puedan ser relevantes al problema biológico en cuestión.

Por ejemplo, si un determinado microorganismo se encuentra asociado a un determinado estado patológico de un sistema, conocer las secuencias de genes de este puede ser relevante para interpretar sus mecanismos de acción. Asumiendo que el organismo no se encuentre completamente secuenciado y disponible en las bases de datos públicos, la función de las secuencias relevantes se puede deducir a partir del mapeo contra las bases de datos de secuencias anotadas (por ejemplo, **GenBank**), usando **BLAST**. Por otro lado, si varios

organismos están expresando las mismas vías, por ejemplo, a causa del exceso de un determinado producto o metabolito en el medio, entonces aparecerá una plétora de secuencias relacionadas por la función, pero provenientes de diferentes organismos.

Un paso más adelante se encuentra la **metatranscriptómica** que tiene por objetivo la caracterización de todo el **ARN** mensajero (**ARNm**) presente en cada muestra. La diferencia más importante con la genómica funcional es que mientras que en esta última cada genoma aporta lo mismo y, por lo tanto, el factor de ponderación es la proporción de cada taxa (por ejemplo, especie) en la muestra, en la metatranscriptómica existe la misma ponderación, pero también la ponderación por expresión de las secuencias (en realidad un balance entre la tasa de transcripción y la de decaimiento). De esta forma, más allá de identificar qué secuencias están presentes en la muestra, podemos cuantificar (de forma relativa) la «importancia» de cada una. Por ejemplo, en situaciones de organismos patógenos, si el mismo está expresando genes de resistencia a antibióticos, podemos identificar las secuencias responsables de esto a través de la **metatranscriptómica**.

Las aplicaciones de todas estas técnicas, **metagenómica**, **metagenómica funcional** y **metatranscriptómica** han permitido avanzar en varios campos donde entender el funcionamiento del ecosistema microbiológicos es fundamental. Por ejemplo, en agricultura resulta muy claro que desde el punto de vista de las plantas el suelo no es solamente el sustrato y las sustancias químicas disponibles para las mismas; se trata de un ecosistema complejo donde los microorganismos juegan un rol fundamental, tanto los de vida libre como los asociados a las plantas. En este sentido, resulta fundamental entender la diversidad de microorganismos en los distintos suelos, su evolución a lo largo del tiempo y cómo eso se relaciona con la productividad del suelo.

Ejemplos en el área de producción animal también son ubicuos. El rumen es un órgano complejo cuya principal misión consiste en la fermentación de los alimentos ingeridos y la preparación para la posterior absorción de nutrientes a nivel intestinal. Se trata de un ambiente esencialmente anaerobio (ausencia de oxígeno) donde proliferan múltiples microorganismos, entre los que se destacan bacterias, protozoos, arqueas metanogénicas, hongos y virus. La mayor parte de los microorganismos presentes nunca han sido cultivados, en parte por la dificultad de cultivar anaerobios, pero las técnicas de metagenómica han permitido enormes avances al respecto. Claramente, la microbiota ruminal es dinámica y suele verse afectada por diversos factores, que incluyen la dieta, cambios en la fisiología del animal producto de la preñez, cambios ambientales o la estación del año. Los distintos tipos de bacterias que actúan en el rumen (de acuerdo con su función en la digestión) comprenden las celulolíticas, lactolíticas, sacarolíticas, lipolíticas, amilolíticas y metanogénicas ureolíticas, entre otras. Entender cómo las distintas variables afectan la composición de la microbiota ruminal es esencial para predecir los impactos en la salud animal y en la producción.

Tanto en salud animal como en salud vegetal resulta fundamental entender cuáles son los patógenos causantes de una enfermedad. En general, muchos de los microorganismos que se encuentran en el órgano afectado son desconocidos y en muchos casos no son cultivables. Como las condiciones de cultivo suelen ser fundamentales en la bacteriología clásica, existe un sesgo importante en asignar las causas de la patogenicidad a organismos cultivables. Esto nos deja con una parte de la historia conocida y otra parte (posiblemente igual de relevante) sin conocer, ya que en muchos casos el origen de una patología puede ser el producto de un desarreglo en la comunidad que lleve a patógenos oportunistas a tomar la oportunidad. Más aún, uno de los temas de mayor preocupación para la salud mundial, en el paradigma de **Una Sola Salud** (***One Health***, en inglés) es la resistencia a antibióticos que se

ha desarrollado a escala global y que amenaza dejarnos sin tratamiento ante enfermedades que parecían superadas. Frente a esto parece fundamental empezar a trabajar el tema desde una perspectiva racional, tanto en el ámbito de las políticas y la investigación como en el plano del trabajo de campo. Para un manejo racional es fundamental, a su vez, conocer los organismos causantes de cada patología específica, así como los potenciales mecanismos de resistencia desarrollados, ya que el uso infructuoso de antibióticos no hace otra cosa que aumentar las posibilidades de desarrollar otras resistencias.

En este último sentido, la metagenómica puede constituirse en una herramienta muy poderosa para la investigación, por ejemplo, en el estudio de la mastitis bovina, analizando los cambios en la flora microbiana de la glándula mamaria en animales sanos cuando pasan a un estado de mastitis. En principio, no solo es posible estudiar la composición microbiana durante la evolución de la enfermedad, sino que, mediante **metagenómica funcional**, y especialmente mediante **metatranscriptómica**, se puede analizar el **resistoma** (el conjunto de los genes relacionados a la resistencia) de la ubre (Hoque et al. 2020). Además, otras técnicas masivas que ya se están utilizando de forma rutinaria en el diagnóstico de enfermedades infecciosas en humano, como la espectrometría de masas MALDI-TOF,[7] prometen revolucionar la velocidad y precisión de diagnóstico en animales de producción, así como mejorar la predicción de resistencias antes de aplicar antibióticos (Nonnemann et al. 2019).

Ejemplo 1.2

Interpretar la figura 1.2 sabiendo que la letra inicial («P» de planctónica, es decir vida libre, o «B» de biofilm) representa la fase del crecimiento de colonias de la bacteria ***Leptopsira interrogans***, la segunda letra representa la réplica biológica del experimento («A», «B» o «C», para tres réplicas) y los números finales representan las horas a las que se extrajo la muestra a partir del inicio del experimento.

En general, en todos los experimentos biológicos donde queremos generalizar las conclusiones a la población debemos incluir una muestra representativa de individuos a comparar. Los experimentos de transcriptómica son lo suficientemente complejos y caros como para que el número de réplicas no sea elevado (3 por condición, en este caso).

Originalmente se trata de un experimento factorial balanceado del tipo 2×2, ya que tenemos dos factores (la fase de crecimiento —vida libre o biofilm— y los tiempos —48 y 120 horas—), cada uno con dos niveles; en cada celda del diseño factorial tenemos 3 réplicas. Se trata de un experimento de RNA-seq (transcriptómica), por lo que los datos que originan el gráfico salen de una tabla de conteo de *reads*.

Lo primero que se puede observar en la figura es que algunas muestras fueron descartadas del análisis, de acuerdo con los autores, debido a que no cumplían los requisitos de calidad previamente establecidos. Lo segundo que se observa es que las muestras se agrupan razonablemente de acuerdo con la condición biológica, es decir, se agrupan en cada celda del diseño factorial (que están representadas por distintos colores). Seguramente, las muestras no consideradas caerían lejos de su grupo correspondiente (lo que no es criterio para descartarlas), pero también lejos de todas las otras muestras.

Finalmente, la distribución de las muestras no sigue el patrón de que cada celda del diseño factorial (cada color en el gráfico) ocupe un cuadrante diferente, en particular cada

[7] Técnica analítica que permite determinar con alta precisión qué moléculas orgánicas o biomoléculas se encuentran presentes en una muestra.

eje representando uno de los factores. Entonces, podemos deducir que la relación entre los factores y las fuentes de variación no es totalmente ortogonal (por ejemplo, no son totalmente independientes). Sin embargo, es posible trazar una línea que divida a las de vida libre (planctónicas) de las sésiles (biofilm) y otra línea que separe las de 48 horas de las de 120 horas (de hecho, infinitas de estas rectas en ambos casos), por lo que parece existir «buena biología» en las causas de la separación.

PARA RECORDAR

- Para entender la variabilidad genética es necesario conocer las secuencias, o al menos algunas secuencias que nos puedan servir de marcador, en muchos individuos. Los **microarrays** son un arreglo de **sondas moleculares** colocadas en una placa.
- Una **sonda molecular**, en este caso, es una **secuencia específica a un gen o una región de un gen y que es (idealmente) única en el genoma de referencia.** Cuando el ADN o el ADN copia (en los estudios de transcriptómica) de un individuo se hibrida con dicha sonda (que está fijada en la placa) se produce algún tipo de señal, por ejemplo, la emisión de luz en determinada frecuencia al ser iluminada por un láser. Esta tecnología permite, por ejemplo, identificar si la secuencia presente en el organismo es de un tipo u otro en el caso de los **SNP** (polimorfismos de un solo nucleótido) e identificar el genotipo en el caso de las especies diploides.
- Es importante tener en cuenta que para crear una sonda para medir variabilidad hay que conocer de antemano las secuencias de regiones que contengan variabilidad, y que el array solo va a captar variabilidad en las posiciones del genoma para el que fue diseñado.
- La **metatranscriptómica** tiene por objetivo la **caracterización de todo el ARN mensajero (ARNm)** presente en cada muestra. La diferencia más importante con la genómica funcional es que mientras que en esta última cada genoma aporta lo mismo y, por lo tanto, el factor de ponderación es la proporción de cada taxa (por ejemplo, especie) en la muestra, en la metatranscriptómica existe la misma ponderación, pero también la ponderación por expresión de las secuencias. De esta forma, más allá de identificar qué secuencias están presentes en la muestra, podemos cuantificar (de forma relativa) la «importancia» de cada una.

1.2 Genómica composicional

Si bien todos los organismos derivan del mismo organismo ancestral (LUCA) y, por lo tanto, deben compartir ciertas características en sus genomas, estos evolucionan y son la base de las diferencias entre los distintos organismos. A nivel genómico la evolución se da por un conjunto de eventos como las mutaciones puntuales, las duplicaciones de genes y regiones, o las inserciones y deleciones de secuencias nucleotídicas. Claramente, la mayor parte de estos eventos producen cambios importantes a nivel composicional de los genomas. Por ejemplo, las mutaciones pueden llevar a una acumulación de cambios completamente al azar en los genomas, o puede tener determinados tipos de sesgos que favorezcan que las mutaciones sean hacia determinadas bases (por ejemplo, mutaciones desde las bases A

y T hacia las bases G y C; recordar que en el ADN doble cadena A se aparea con T y C con G). Los eventos de duplicación génica, por otro lado, generan cierta libertad para que ambas evolucionen más libremente, ya que en principio una de las dos secuencias sigue siendo funcional (hasta que también deje de serlo) y de esta forma se pueden explorar nuevas funciones para una de las dos secuencias.

En general, el conjunto de estos cambios hace que resulte interesante para entender las bases de la evolución en ambientes particulares o, aún en general, el conocimiento de las secuencias de los diferentes organismos y la comparación de las mismas. El estudio de las variaciones en la composición de los genomas es lo que se puede llamar **genómica composicional** y forma parte de la **genómica comparativa**. A continuación veremos el tema de la variación en composición, tanto desde el punto de vista de los nucleótidos como de los codones en las secuencias codificantes y finalmente de los aminoácidos que forman las proteínas.

Contenido GC genómico, génico, correlaciones y GC_{skew}

La proporción de bases guanina y citosina (G y C, respectivamente) que componen una secuencia se conoce como su contenido GC, o G+C, o aun GC % (cuando está expresado en porcentaje).[8] Por las reglas de apareamiento de bases (A-T y G-C), no tiene importancia en la determinación de dicho contenido cual hebra del ADN se considera para el cálculo de esta proporción.

Reglas de Chargaff

- **Primera regla de paridad:** La primera regla sostiene que una molécula de ADN de doble cadena tiene a escala global las siguientes igualdades en el porcentaje de bases: % A = % T y % G = % C. La validación rigurosa de la norma constituye la base del apareamiento de Watson-Crick de la doble hélice del ADN.
- **Segunda regla de paridad:** La segunda regla es que tanto % A = % T y % G = % C son válidos para cada una de las dos hebras de ADN. Esto describe solamente una característica global de la composición de bases en una sola hebra de ADN.

En principio, uno podría preguntarse cuál es la importancia de medir esta proporción en particular y no otra de las tres posibles combinaciones de dos bases distintas.[9] Es más, algunas de ellas resultan obviamente relevantes, como la proporción de purinas versus pirimidinas (AG vs. CT). Entonces, ¿qué es lo que hace tan relevante al contenido GC de las secuencias?

[8] El contenido GC se calcula usualmente como $GC = (N_C + N_G)/(N_A + N_C + N_G + N_T)$, donde $N_A ... N_T$ corresponde al número de veces que aparece cada base en una secuencia.

[9] Hay 6 de estas combinaciones, pero como la frecuencia de las dos elegidas más la de las otras dos debe sumar 1, basta con calcular la mitad; por ejemplo, GC=1-AT.

Transiciones y transversiones

- Los cambios entre **purinas** (A y G) o entre **pirimidinas** (C y T) se llaman **transiciones**.
- Los cambios de purinas a pirimidinas, o al revés, se llaman **transversiones**.
- Las transiciones son mucho más frecuentes que las transversiones.

Existen varias razones, como veremos más adelante, pero desde el punto de vista bioquímico tal vez la más obvia sea la diferencia en el número de puentes de hidrógeno entre los apareamientos A-T y los G-C, dos para el primero y tres para el segundo, así como diferencias en el *stacking* resultante.[10] Esto conduce directamente a importantes diferencias en las propiedades físicas de la doble hebra. Por ejemplo, existe un incremento en el *melting point* (temperatura en la cual la mitad de las dobles hélices de ADN se desnaturalizan) en función del contenido GC (Yakovchuk, Protozanova, and Frank-Kamenetskii, 2006). Pero más importante aún, en secuencias codificantes el código genético (tanto el universal como todas sus variantes) establece un mapeo entre la secuencia de ADN y la proteína que será traducida a partir de ella. Dada la estructura «redundante» del código genético, algunos aminoácidos presentan poca dependencia del contenido GC, mientras que otros se encuentran fuertemente asociados a este problema (ya que los codones correspondientes tienen un sesgo importante).

En la figura 1.3 se aprecia la tabla correspondiente al **código genético universal.**[11]

La tabla muestra los 64 codones posibles con sus correspondientes aminoácidos. El ARNm se lee en sentido 5' - 3'.

1.ª base	2ª base U	2ª base C	2ª base A	2ª base G
U	UUU (Phe/F) Fenilalanina	UCU (Ser/S) Serina	UAU (Tyr/Y) Tirosina	UGU (Cys/C) Cisteína
	UUC (Phe/F) Fenilalanina	UCC (Ser/S) Serina	UAC (Tyr/Y) Tirosina	UGC (Cys/C) Cisteína
	UUA (Leu/L) Leucina	UCA (Ser/S) Serina	UAA Parada (Ocre)	UGA Parada (Ópalo)
	UUG (Leu/L) Leucina	UCG (Ser/S) Serina	UAG Parada (Ámbar)	UGG (Trp/W) Triptófano
C	CUU (Leu/L) Leucina	CCU (Pro/P) Prolina	CAU (His/H) Histidina	CGU (Arg/R) Arginina
	CUC (Leu/L) Leucina	CCC (Pro/P) Prolina	CAC (His/H) Histidina	CGC (Arg/R) Arginina
	CUA (Leu/L) Leucina	CCA (Pro/P) Prolina	CAA (Gln/Q) Glutamina	CGA (Arg/R) Arginina
	CUG (Leu/L) Leucina	CCG (Pro/P) Prolina	CAG (Gln/Q) Glutamina	CGG (Arg/R) Arginina
A	AUU (Ile/I) Isoleucina	ACU (Thr/T) Treonina	AAU (Asn/N) Asparagina	AGU (Ser/S) Serina
	AUC (Ile/I) Isoleucina	ACC (Thr/T) Treonina	AAC (Asn/N) Asparagina	AGC (Ser/S) Serina
	AUA (Ile/I) Isoleucina	ACA (Thr/T) Treonina	AAA (Lys/K) Lisina	AGA (Arg/R) Arginina
	AUG (Met/M) Metionina, Comienzo	ACG (Thr/T) Treonina	AAG (Lys/K) Lisina	AGG (Arg/R) Arginina
G	GUU (Val/V) Valina	GCU (Ala/A) Alanina	GAU (Asp/D) Ácido aspártico	GGU (Gly/G) Glicina
	GUC (Val/V) Valina	GCC (Ala/A) Alanina	GAC (Asp/D) Ácido aspártico	GGC (Gly/G) Glicina
	GUA (Val/V) Valina	GCA (Ala/A) Alanina	GAA (Glu/E) Ácido glutámico	GGA (Gly/G) Glicina
	GUG (Val/V) Valina	GCG (Ala/A) Alanina	GAG (Glu/E) Ácido glutámico	GGG (Gly/G) Glicina

Figura 1.3. Tabla del código genético universal. En amarillo los aminoácidos apolares, en verdes los polares, azul básicos y rosa ácidos, y en blanco los codones de parada (de Wikipedia, https://es.wikipedia.org/wiki/C%C3%B3digo_gen%C3%A9tico).

[10] Los orbitales de los anillos aromáticos de las bases nitrogenadas de los nucleótidos que componen la molécula de ADN interaccionan a medida que se apilan en el polímero. Esta interacción no covalente es un factor fundamental que aumenta la estabilidad de la doble hélice de ADN.

[11] Si bien el código genético se encuentra ampliamente conservado en los tres dominios de la vida, este claramente no es universal. Mitocondrias, cloroplastos y algunos protozoarios unicelulares o bacterias del género *Mycoplasma*, entre otros, emplean códigos genéticos con variaciones en algunos codones.

Claramente los colores, que representan propiedades de los aminoácidos, se encuentran distribuidos de forma bastante organizada de acuerdo con estas propiedades. Cada columna marca una base diferente en segunda posición del codón (U, C, A, G). Por ejemplo, la tercera columna incluye los codones con segunda base **A** (adenina) y esta codifica para aminoácidos mayormente codificados por duetos (por ejemplo, por dos codones). La primera columna corresponde a codones con segunda base **U** (uracilo, correspondiente en ADN a **T**) y codifica para aminoácidos apolares.

Algunas propiedades del código genético universal

- 2 singletons (Met y Trp), 9 duetos (Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys), 1 terceto (Ile), 5 cuartetos (Val, Pro, Thr, Ala, Gly) y 3 sextetos (Leu, Ser, Arg).
- La tercera columna (A en la segunda base) codifica duetos: 2/3 de los cambios en tercera base no son sinónimos.
- La primera base del codón es muy conservada: un cambio de bloque-fila implica en general cambio de aminoácido, excepto en sextetos Arg y Leu.
- La segunda base del codón es la más conservada: un cambio de columna implica un cambio de aminoácido.
- La tercera base es la más neutra y, por lo tanto, menos conservada (rol de los cuartetos). Los cambios de aminoácidos (de cuál a cuál) no son todos equivalentes en su impacto: hay cambios conservativos y otros disruptivos.

Entre los aminoácidos codificados por codones con un fuerte sesgo GC(AT) tenemos: ***Phe*** (5 A/T en 6 posiciones), ***Ile*** (8 A/T en 9) y ***Tyr*** (5 A/T en 6). ¿Son utilizados estos aminoácidos de forma homogénea en los diferentes genes codificantes de los organismos? Veamos, por ejemplo, cómo es el uso de estos aminoácidos en el genoma de *Buchnera aphidicola str. Cc (Cinara cedri)*, una bacteria parásita intracelular obligada con contenido GC de 20,1 %. Mientras que en la proteína codificada por la secuencia con mayor contenido GC del genoma (36,7 %) el uso de estos 3 aminoácidos corresponde al 16 %, en la proteína con menor contenido GC (8,6 %) el uso de estos sube a 48 %. Además, nos podemos preguntar, ¿las diferencias en contenido GC a nivel genómico pueden dar lugar a diferencias en el contenido GC de las secuencias codificantes? ¿Esto puede llevar a diferencias entre organismos con genomas con contenido GC diferente? En principio, podemos pensar que estas diferencias no tendrían de por sí una consecuencia necesaria a nivel de diferencias entre organismos, ya que de hecho el contenido GC podría ser una simple consecuencia de los aminoácidos constituyentes de las proteínas codificadas en el organismo y no al revés. Es decir, la distribución del contenido GC de las proteínas de un organismo podría ser solo el reflejo de la constitución de estas.

Sin embargo, de ser esto cierto esperaríamos que organismos con una composición relativamente similar a nivel de sus proteínas constituyentes tuvieran similar contenido GC, y esto no es lo que ocurre. De hecho, es bien conocido que entre los organismos procariotas (Bacteria y Archaea) el contenido GC genómico varía desde el 25 % hasta el 75 % aproximadamente (Sueoka 1962), con algunos organismos aún más extremos. Estas variaciones marcadas se dan incluso entre organismos de un mismo filo. Más aún, la bacteria *Anaeromyxobacter dehalogenans 2CP-C*, primer cultivo puro de una mixobacteria capaz de crecer de forma anaeróbica, con un contenido GC genómico de 74,9 %, está constituida por proteínas

cuyo contenido GC va entre 56,5 % y 88,0 %. Comparando esto con los datos de *B. aphidicola* mencionados anteriormente, observamos que no existe superposición. Es decir, ambas bacterias no tendrían ninguna proteína de similar composición. Sabemos, sin embargo, que ambas comparten un origen evolutivo único y, por lo tanto, descienden de un mismo organismo. Además, ambas cuentan con una serie de proteínas con idéntica función, pero, de hecho, con la composición aminoacídica de estas es muy diferente. Claramente, estos hechos van en contra de la línea de razonamiento planteada anteriormente.

¿Qué es entonces lo que nos permite pensar que sería el contenido GC el que conduce el proceso de sustitución de aminoácidos? Hay varios motivos para pensar en este sentido. El primero es que los procariotas, cuyo genoma es en su mayoría **codificante**, poseen una importante homogeneidad composicional (restringida por la necesidad de usar todos los aminoácidos). Esto se manifiesta aun en las regiones intergénicas, que no tienen mayores restricciones composicionales. Un ejemplo del alto nivel de homogeneidad se puede ver al comparar los intervalos intercuartílicos de contenido GC codificante, que tanto en *B. aphidicola* como en *A. dehalogenans* es de 6 %; en otras palabras, la mitad de las proteínas dentro de cada bacteria difieren en menos de 6 % de contenido GC, en comparación con diferencias superiores al 50 % entre los contenidos GC genómicos de ambas especies.

Pero, además, existe otra evidencia fuerte. Así como calculamos el contenido GC de una secuencia, también podemos calcular el contenido GC en cada una de las posiciones de sus codones. Como hay tres posiciones por codón tenemos entonces GC_1, GC_2 y GC_3, de acuerdo con el contenido GC solo considerando una posición a la vez. Si observamos el código genético con detenimiento vamos a observar que la tercera posición es la «más sinónima», es decir, la posición en la que una mayor proporción de los cambios de base no afectan el aminoácido codificado. De hecho, excepto ***Met*** y ***Trp***, todos los aminoácidos poseen codones sinónimos, con opción entre G/C y A/T. De acuerdo con esto, si el contenido GC no estuviera conduciendo el proceso de sustitución de aminoácidos, se podría esperar que el GC_3 se mantuviera variando libremente entre los distintos genomas, con casi tanta varianza entre como dentro de genomas. Sin embargo, esto no es así, ya que existe una fuerte correlación positiva entre el contenido GC y el GC_3, tanto a nivel de secuencias como considerando los genomas enteros (en procariotas) y aun con las regiones intergénicas flanqueantes (Zerial et al. 1986).

Además de los sesgos en contenido GC que discutimos antes, existen otros sesgos menos obvios pero con muy importantes consecuencias prácticas. Por ejemplo, si consideramos una región particular de la mayoría de los genomas bacterianos es posible observar una diferencia importante (dentro de cada hebra) entre el número de bases G y C, así como entre A y T. De hecho, este fenómeno se encuentra asociado a la replicación de los genomas bacterianos (usualmente circulares) y típicamente la hebra *leading* («líder», en inglés) se encuentra enriquecida en G y T, mientras que la *lagging* («rezagada», en inglés) se encuentra enriquecida en C y A. Las desviaciones de las frecuencias G = C y A = T se conocen como GC_{skew} y AT_{skew}, respectivamente.[12] Una aplicación práctica evidente de esto es la determinación (aproximada) del origen de replicación, ya que en este (y en la posición opuesta dentro del genoma circular) habrá un cambio de signo del GC_{skew} (asociado a la densidad de genes en una hebra y otra). Una forma de visualizar esto es graficando el GC_{skew} en las ordenadas contra la posición del genoma (centro de la ventana) a partir del cual fue calculado, identificando el cambio de signo en las ordenadas. Otra alternativa, generalmente más efectiva,

[12] El GC_{skew} se calcula tomando una ventana de tamaño determinado y de acuerdo con la siguiente expresión: $GC_{skew} = (N_G - N_C)/(N_G + N_C)$.

es plotear el GC_{skew} acumulado contra la posición y observar los máximos y mínimos del gráfico, que usualmente están a una distancia de medio genoma (Grigoriev 1998).

Si bien existe una fuerte controversia sobre la explicación causal de todas estas correlaciones, así como de correlaciones entre el contenido GC y factores del ambiente donde viven los organismos, resulta claro que el contenido GC tiene un importante rol en la evolución de los organismos y en los procesos moleculares que los rigen (incluso se debe tener en cuenta a la hora de calcular algunos índices genómicos o en los procesos bioquímicos que requieran la desnaturalización del ADN). Existen diversas herramientas para el cálculo del contenido GC, alguna de ellas web (por ejemplo, Mobyle (http://mobyle.pasteur.fr/) que sirven para un cálculo rápido de una o pocas secuencias. Sin embargo, si el cálculo involucra muchas secuencias lo usual es recurrir a la programación de tipo *scripting*, usualmente en lenguajes como Python (http://python.org/), Perl (http://www.perl.org/) o Java (http://www.java.com; todos ellos poseen módulos `Bio` que facilitan todas las tareas), en R (http://cran.r-project.org/) a través de bibliotecas como `seqinr`, en Bash (cuando se trabaja en Linux-UNIX o Mac OS-X), o usando paquetes como EMBOSS (http://emboss.sourceforge.net/). Si se desea trabajar con genomas procariotas ya secuenciados, una página interesante es la de OligoWeb (http://insilico.ehu.es/oligoweb/), ya que además de las frecuencias de nucleótidos de los distintos genomas ya tiene calculados otros índices y estadísticos.

Uso de codones

Habíamos hablado antes de que los genomas procariotas eran extremadamente amplios en el rango de contenidos G+C y que, dado lo reducido de lo no-codificante, estas variaciones debían tener implicancias en el uso de codones. Es más, si comparamos entre genomas veremos que existe una clara correlación lineal entre el contenido GC del genoma con GC_1, GC_2 y GC_3, aunque con diferentes pendientes. En general GC_3 es una posición sinónima, es decir al cambiar de base en ella no hay cambio de aminoácido (debido a la degeneración del código genético; ver los cuartetos, por ejemplo). GC_1 y GC_2 son mucho más restringidas en su variación, dado que suelen implicar un cambio de aminoácido, lo que por lo general afecta negativamente la función de la proteína donde sucede dicho cambio.

Esta redundancia en el código genético, sumado a las posibilidades de reconocimiento alternativo codón/anticodón (no-Watson-Crick), permiten la evolución en los genomas de diferentes estrategias de uso de codones (UC) (Rocha 2004). Esto, que puede verse entre genomas, también tiene su correlato a nivel interno en muchos genomas. Es conocido que en algunos genomas procariotas los genes altamente expresados tienden a usar **codones mayores**, que suelen ser reconocidos en forma de apareamiento Watson-Crick por los tRNA más abundantes, aunque hay distintos modelos (ver la discusión en Rocha 2004). Lograr comprender los factores que inciden en el uso de codones (también puede pensarse para aminoácidos) suele requerir de alguna forma de «resumir» la información.

En un genoma procariota promedio tenemos unos 4000 genes y los codones de aminoácidos con degeneración son 59 (los 64 menos los 3 *stop*, menos ***Met*** y ***Trp***, que son codificados por un único codón cada uno), lo que se puede resumir en una tabla de esas dimensiones. Utilizando la tabla, podríamos analizar si distintos genes utilizan de forma diferente los codones, e hipotetizar cuál puede ser el motivo subyacente a este fenómeno. Para quitar la

influencia del uso de aminoácidos como factor que influye en esta perspectiva (cada proteína tiene un uso diferencial de AA), se puede estandarizar a la interna de cada aminoácido. Una de las estandarizaciones posibles es el RSCU (Sharp, Tuohy, and Mosurski 1986), en el que el número observado de codones para un aminoácido es dividido por el número esperado si el uso dentro de un aminoácido fuera equiprobable. Es decir,

$$\mathrm{RSCU}_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}} \tag{1.5}$$

donde x_{ij} es el número de ocurrencias del codón j para el aminoácido i y n_i es la multiplicidad del aminoácido i (es decir, el número de codones que lo codifican).

De cualquier forma, aun luego de estandarizar seguimos teniendo la tabla de 4000×59 (por ejemplo, cada gen en una fila, y el índice RSCU calculado para cada uno de los 59 codones considerados). La mayor parte de los humanos tenemos una limitada capacidad para visualizar datos en espacios de más de 3 dimensiones, por lo que necesitamos reducir la «dimensionalidad» del problema. Podríamos pensar en representar las 59 variables originales de a pares ($\binom{59}{2} = 59 \times 58/2 = 1711$), o aun de a tres ($\binom{59}{3} = 59 \times 58 \times 57/6 = 32.509$) para ver interacciones entre más variables, pero estos números hacen imposible poder extraer conclusiones relevantes de forma sistemática. Afortunadamente, existen diferentes técnicas estadísticas exploratorias multivariadas que nos permiten abordar este tipo de cuestiones.

Probablemente, de las técnicas usuales de análisis exploratorio multivariado, la más conocidas es el **Análisis de Componentes Principales** (PCA, del inglés *Principal Component Analysis*), aunque existen otros miembros prominentes de la familia. Uno de los más interesantes para nuestro tipo de problema es el **análisis de correspondencia** (COA, véase la sección «Exploración multivariada: análisis de correspondencia»). Este análisis nos permite reducir la variación en las 59 variables originales a unas pocas dimensiones (cada una de ellas una combinación de las originales) que capturan la mayor parte de la variabilidad. De esta forma, el problema pasa a ser identificar qué variables biológicas se asocian a cada una de estas dimensiones.

Ejemplos típicos de variables biológicas asociadas a la variación en el **uso de codones** dentro de genomas procariotas son el **contenido GC del gen**, su **nivel de expresión**, la **hidropatía** promedio de la proteína, la **hebra** en que se encuentra el gen, el $\mathrm{GC}_{\mathrm{skew}}$ y la **precisión en la traducción** (Ermolaeva 2001). Diferencias en el uso de codones entre genomas han sido reportadas asociadas a varios factores ecofisiológicos, entre ellos la **temperatura óptima de crecimiento** y adaptaciones a hipersalinidad. Un programa diseñado específicamente para el análisis composicional es el codonW (http://codonw.sourceforge.net/), que analiza las frecuencias de bases, codones y aminoácidos, calcula diferentes índices estadísticos, así como permite realizar un COA en codones y AA.

Existen diversos índices, calculados a partir de las frecuencias de codones, que permiten identificar sesgos importantes en UC, que pueden ser suficientemente indicativos de por sí, o combinados con la información del COA. Por ejemplo, el Nc (número efectivo de codones) es un índice del uso de codones intragénico. Puede tomar valores extremos de 20 —cuando un solo codón por aminoácido es utilizado en ese gen— hasta 61, cuando hay una equidistribución de estos (F. Wright 1990). Este índice es sensible al contenido GC del genoma, y posteriormente se sugirió una mejora que tiene en cuenta este factor (Novembre 2002). Otros dos índices, «codon bias index» (CBI) y «frequency of optimal

codons» (FOP), calculan el sesgo a partir de un juego de codones «óptimos», derivados respectivamente de un conjunto particular de genes o de la concentración de los tRNA. Una aproximación totalmente diferente, que no depende de un conjunto particular de genes, es el «codon adaptation index» (CAI) (Sharp and Li 1987), que ha demostrado ser un indicador muy razonable del nivel de expresión de los genes (basado en comparaciones con técnicas para medir experimentalmente la expresión génica, como los microarrays). El CAI es simplemente la media geométrica de los RSCU relativos (al máximo RSCU para ese AA) de todos los codones de una proteína. Con la misma notación que antes, si consideramos $w_{ij} = RSCU_{ij}/RSCU_{i,max}$, entonces se define $\text{CAI} = \exp(\frac{1}{L}\sum_{k=1}^{L} \ln w_{ij})$.

Pese a los buenos resultados del CAI, este presenta también importantes debilidades, como la no-linealidad y una sobreestimación de las desviaciones para secuencias cortas. En el año 2009, Roymondal y colaboradores (Roymondal, Das, and Sahoo 2009) propusieron una aproximación muy diferente, basada en las frecuencias de cada base en las distintas posiciones de los codones, que lleva al índice RCB. Estos autores proponen para cada codón *xyz* calcular el RCB de acuerdo con la fórmula

$$RCB_{xyz} = \frac{f(xyz)}{f_1(x)f_2(y)f_3(z)} \tag{1.6}$$

(donde $f_1(x)$, $f_2(y)$, $f_3(z)$ son las frecuencias de las bases x, y y z en las posiciones 1, 2 y 3 respectivamente), para luego resumir toda la información a través de la media geométrica, $RCB = (\prod_{l=1}^{L} RCB_{xyz}(l))^{1/L} - 1$. Este nuevo índice presenta ventajas respecto al CAI, pero aún presenta un problema importante para secuencias cortas, lo que llevó a sugerir una modificación del mismo (*Relative Codon Adaptation*, *RCA*), basada en el uso de pseudoconteos (Fox and Erill 2010).

Un índice de expresión tiene mucha utilidad en la exploración de los factores que influyen en la composición diferente de los genes, pero un punto interesante es la posible relación entre la expresión y el efecto de *gene dosage* (más copias de genes disponibles a medida que la replicación avanza), resultado de favorecer los genes altamente expresados en una posición cercana al origen de replicación. Para investigar este efecto basta con ver cuál es la relación entre el CAI y la posición. Una forma fácil consiste en un análogo del GC_{skew} acumulado, calculado como desvío del CAI promedio y compararlo con las posiciones genómicas. Si existe un efecto de gene dosage, el patrón de CAI debería mostrar un pico en la zona del origen de replicación.

Un soporte adicional a la hora de entender lo que ocurre en el uso de codones cuando comparamos dos grupos, por ejemplo, genes de alta expresión con los de baja expresión, es a través de las tablas de uso de codones. Estas contienen el conteo total por codón para cada grupo y si las agrupamos por AA podemos calcular la probabilidad de que las diferencias en uso de codones (para cada codón) se deban al azar. Para esto, dentro de cada AA se construye una tabla de 2 × 2 (dos grupos a comparar y codón a analizar versus la suma de los otros dentro de AA) y con la misma realiza un test de χ^2 con un grado de libertad. Es importante tener en cuenta las múltiples comparaciones a realizar (el número de hipótesis a ensayar, por ejemplo, el número de codones) a la hora de definir el umbral de significancia para corregir adecuadamente.

Finalmente, el uso de codones de un organismo puede tener implicancias directas de carácter tecnológico. Como vimos antes, entre las mayores fuentes de variación en UC en algunos organismos se encuentra la expresión, con codones muy usados en los genes de

alta expresión y un uso más plano en proteínas de baja expresión. Esto influye incluso en el plegamiento de las proteínas, cambiando su solubilidad (Cortazzo et al. 2002). Cuando deseamos expresar proteínas recombinantes resulta muy importante considerar no solo la secuencia de AA de la proteína a expresar, sino también que el UC se adapte al UC en genes de alta expresión en el organismo hospedero.

Ejemplo 1.3

Teniendo en cuenta todas las secuencias codificantes, en una bacteria de interés biotecnológico se observó el siguiente conteo de codones para dos aminoácidos:

Aminoácido	Codón	Conteo
Arg (R)	CGU	666
Arg (R)	CGC	1320
Arg (R)	CGA	1457
Arg (R)	CGG	1818
Arg (R)	AGA	5913
Arg (R)	AGG	5120
Val (V)	GUU	7123
Val (V)	GUC	2144
Val (V)	GUA	1125
Val (V)	GUG	8213

Calcular el **RSCU** para cada codón y determinar cuáles son los usados y si hay diferencias entre los aminoácidos en la forma en que usan los codones.

De acuerdo con la ecuación (1.5), el RSCU se calcula como

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}} \tag{1.7}$$

donde x_{ij} es el número de ocurrencias del codón j para el aminoácido i. Tenemos dos aminoácidos diferentes, por lo que vamos a calcular el RSCU de cada codón dentro de cada aminoácido. Lo primero es obtener la suma de codones en cada aminoácido; tenemos que para **Arginina** es igual a $666 + 1320 + 1457 + 1818 + 5913 + 5120 = 16.294$, mientras que para **Valina** es de $7123 + 2144 + 1125 + 8213 = 18.605$. Este es el número total de codones para cada uno de los aminoácidos o, lo que es lo mismo, el número de veces que aparece dicho aminoácido en las secuencias codificantes de nuestra bacteria. Para calcular el RSCU alcanza con dividir el número de ocurrencias de cada codón entre el total de veces que se utiliza el aminoácido y multiplicarlo por el número de codones que tiene el aminoácido ($n_i = 6$ en el caso de Arginina, $n_i = 4$ en el caso de Valina; véase el código genético).

Por lo tanto, utilizando la ecuación (1.5), tenemos que, por ejemplo, para el codón **CGU** de la **Arginina** su RSCU es igual a

$$RSCU_{R,CGU} = \frac{x_{CGU}}{\frac{1}{6}\sum_{j=1}^{6} x_{R,j}} = 6 \times \frac{666}{16.294} = 0{,}2452 \tag{1.8}$$

Para el codón **CGC** la cuenta es igual a

$$RSCU_{R,CGC} = \frac{x_{CGC}}{\frac{1}{6}\sum_{j=1}^{6} x_{R,j}} = 6 \times \frac{1320}{16.294} = 0{,}4861 \tag{1.9}$$

y así sucesivamente con todos los codones de este aminoácido. Esto nos deja con una tabla como la siguiente:

Codón	Conteo	RSCU
CGU	666	0,2452
CGC	1320	0,4861
CGA	1457	0,5365
CGG	1818	0,6694
AGA	5913	2,1774
AGG	5120	1.8854
Suma	16.294	6

De la misma manera, para la Valina, el RSCU del primer codón es igual a

$$RSCU_{V,GUU} = \frac{x_{GUU}}{\frac{1}{4}\sum_{j=1}^{4} x_{V,j}} = 4 \times \frac{7123}{18605} = 1{,}5314 \tag{1.10}$$

mientras que para el resto de los codones se calcula en forma análoga, hasta obtener la siguiente tabla:

Codón	Conteo	RSCU
GUU	7123	1,5314
GUC	2144	0,4610
GUA	1125	0,2419
GUG	8213	1,7658
Suma	18.605	4

En ambos aminoácidos se observa que el uso de codones se aparta bastante de la hipótesis de uso similar. Mientras que en el primer caso (Arginina) los codones favorecidos son los codificados por el dueto (del sexteto), es decir **AGA** y **AGG**, en el segundo caso (Valina) los codones favorecidos son los que terminan en **U** o en **G**.

1.2.1 Exploración multivariada: análisis de correspondencia

Este método fue desarrollado por Benzécri (1980), originalmente para trabajar con tablas de frecuencias. En resumen, la idea subyacente es, dada una tabla de $n \times p$ (casos × variables), encontrar un sistema de k ejes ortogonales (con $k < min(n,p)$) que condense el máximo de variación en los datos originales. Esto es posible en la medida en que las variables (casos) estén correlacionados en alguna medida y cuanto mayor sea esa asociación más varianza

será condensada en pocos ejes (dimensiones). La métrica adoptada es la de la distancia χ^2, o sea, para la celda correspondiente a la fila i y columna j, la distancia será

$$d_{ij} = (O_{ij} - E_{ij})^2 / E_{ij} \tag{1.11}$$

donde O_{ij} es el número observado y E_{ij} el número esperado para esa celda, usualmente el número total de observaciones multiplicado por las frecuencias relativas marginales en la tabla.

Aplicando el COA a una tabla de frecuencias observadas vamos a obtener una serie de valores propios y vectores propios asociados (eigen-values y eigen-vectors), tantos como el mínimo de Filas-1 y Columnas-1. Cada uno de los valores propios está relacionado con la varianza explicada por la nueva dimensión correspondiente. La suma de todos los valores propios es equivalente al 100 % de la varianza por lo que cada uno explicará una fracción correspondiente al valor propio sobre la suma de todos (la traza de la matriz de valores propios). A su vez, cada uno explica menor varianza que el anterior (están ordenados por el procedimiento de extracción). El punto importante es determinar con cuántos ejes (nuevas dimensiones) nos vamos a quedar. Cuantas más dimensiones retengamos mayor cantidad de la varianza original retendremos, pero al costo de dificultar la interpretación (el objetivo del COA es poder capturar la esencia de los datos, cuantas más dimensiones, más difícil de visualizar la situación). Existen varios métodos para determinar el número de dimensiones a retener, pero en última instancia esto es una materia subjetiva. Un enfoque es gráfico (*scriplot*) y consiste en encontrar el cambio en la pendiente de los valores propios graficados en función de su número ordinal.

Otra aproximación es retener los valores propios hasta que expliquen menos de lo que se esperaría por azar si cada una de las variables originales explicara la misma proporción de la varianza. Por ejemplo, en una tabla de uso de aminoácidos, dado que hay 20 AA la varianza promedio explicada por cada uno es 5 % y, por lo tanto, luego de un COA en AA retendríamos todos aquellos valores propios que expliquen más de 5 % de la varianza. Cada una de las observaciones originales tendrá ahora unas nuevas coordenadas en el nuevo sistema de ejes ortogonales (las dimensiones). O sea, si por ejemplo retuvimos de nuestro COA en uso de AA 3 dimensiones (combinaciones de los 20 originales) cada gen podrá representarse ahora en un espacio tridimensional. Asimismo, las variables originales están representadas de distinta manera en las nuevas dimensiones, es decir tienen distinta participación en cada una de ellas. Esto apunta a una de las características más destacables del COA, su simetría filas/columnas, lo que permite el auxilio de representar en forma conjunta (en el mismo gráfico) la distribución de observaciones y variables (genes y AA, por ejemplo) aunque las distancias entre observaciones y variables deben tomarse con cautela.

La clave del éxito con el COA es lograr «etiquetar» las nuevas dimensiones con factores entendibles, biológicamente significativos, que vayan más allá de las propias variables medidas. Ejemplo de esto en genómica podría ser la expresión diferencial, los compartimientos celulares, proteínas IMP (*integral membrane proteins*), etcétera, cuando se correlacionan con alguno de los ejes del análisis. La forma tradicional de trabajo, al menos en el campo de la genómica comparativa, consiste en determinar cuántos valores propios retener para luego graficarlos de a pares intentando ver la distribución conjunta de observaciones y variables en las nuevas dimensiones. Se puede además intentar correlacionar variables externas con las nuevas dimensiones, buscando «etiquetas» para las dimensiones. Un link interesante de análisis de correspondencia: http://www.micheloud.com/FXM/COR/index.htm

1.2.2 Composición de proteínas

Hasta ahora hemos centrado nuestros análisis composicionales en bases nucleotídicas y codones, pero la mayor parte de las secuencias procarióticas terminan siendo traducidas en proteínas. Por lo tanto, vamos a dedicar un breve espacio a tratar esta temática. En muchos casos los archivos de partida para un análisis bioinformático (en cualquiera de los formatos estándar, como FASTA o GenBank) solo incluyen la secuencia nucleotídica, dado que a partir de la misma es posible deducir directamente la secuencia de AA (si conocemos el código genético correspondiente).

Existen muchos programas para realizar la traducción de estas secuencias, pero el *transeq* (de la suite EMBOSS) es muy flexible y eficiente, aceptando diversos formatos de entrada. Dentro del EMBOSS también tenemos toda una serie de programas para facilitar el análisis de secuencias proteicas. Entre ellos (hay muchos más): *antigenic* para encontrar sitios antigénicos, *charge* que hace un diagrama de carga, *digest* reporta sobre enzimas proteolíticas, *emowse* para buscar en proteínas por peso molecular de fragmentos de digestión, *fuzzpro* para buscar patrones, *helixturnhelix* para identificar motivos de unión a ácidos nucleicos, *iep* para calcular el punto isoeléctrico, *makeprotseq* para crear secuencias aleatorias, *pepcoil* para predecir regiones coiled-coil, *pepstats* para calcular diversos estadísticos de proteínas, *preg* para buscar expresiones regulares en proteínas, *pepwindow* para realizar un ploteo de hidropatía, *tmap* para predecir segmentos trans-membrana, *psiphi* para calcular ángulos torsionales y *patmatmotifs* que escanea una proteína con motivos del PROSITE (http://prosite.expasy.org/).

Diferencias en propiedades de los aminoácidos

Existen 22 aminoácidos proteinogénicos o naturales (los que se incorporan directamente en las proteínas). De entre ellos, 20 son codificados directamente, mientras que la selenocisteína y la pirrolisina se codifican a partir de modificaciones en la lectura de codones *stop* (con símbolos IUPAC/IUBMB *Sec* o *U*, y *Pyl* u *O*, respectivamente). Como se puede apreciar en la figura 1.4, los 20 aminoácidos codificados directamente (que suelen ser los más abundantes) poseen importantes diferencias en las propiedades fisicoquímicas, lo que lleva a agruparlos de diferentes maneras, de acuerdo con el grupo de propiedades de relevancia.

Típicamente, podemos distinguir:

- *i)* Polares/hidrofílicos - *N*, *Q*, *S*, *T*, *K*, *R*, *H*, *D*, *E*, (*C*, *Y*),[13]
- *ii)* No-polares / hidrofóbicos - (*G*), *A*, *V*, *L*, *I*, *P*, *Y*, *F*, *W*, *M*, *C*,
- *iii)* enlace de H - *C*, *W*, *N*, *Q*, *S*, *T*, *Y*, *K*, *R*, *H*, *D*, *E*,
- *iv)* contienen azufre - *C*, *M*,
- *v)* con carga negativa a *pH* neutro / acídicos - *D*, *E*, (*C*),
- *v)* con carga positiva a *pH* neutro / básicos - *K*, *R*, (*H*),
- *vi)* Ionizables - *D*, *E*, *H*, *C*, *Y*, *K*, *R*,
- *vii)* Aromáticos - *F*, *W*, *Y*, (*H*, pero sin absorción significativa de UV),
- *viii)* Alifáticos - *G*, *A*, *V*, *L*, *I*, *P*,
- *ix)* Forman enlaces covalentes cruzados (puentes di-sulfuro) - *C*,
- *x)* Cíclicos - *P*.

(Véase, por ejemplo, http://www.mcb.ucdavis.edu/courses/bis102/AAProp.html).

[13] Figuran entre paréntesis cuando entran en la categoría parcialmente.

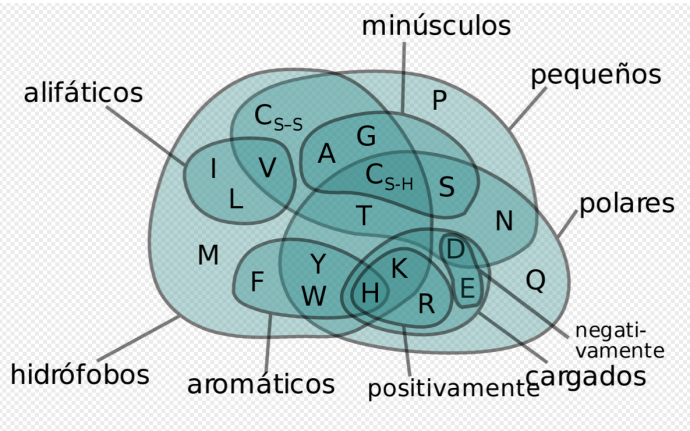


Figura 1.4. Diagrama de Venn con las propiedades de los diferentes aminoácidos. Figura tomada de Wikipedia (CC BY-SA 3.0), versión en castellano modificada a partir del archivo «Amino Acids Venn Diagram (es).svg»

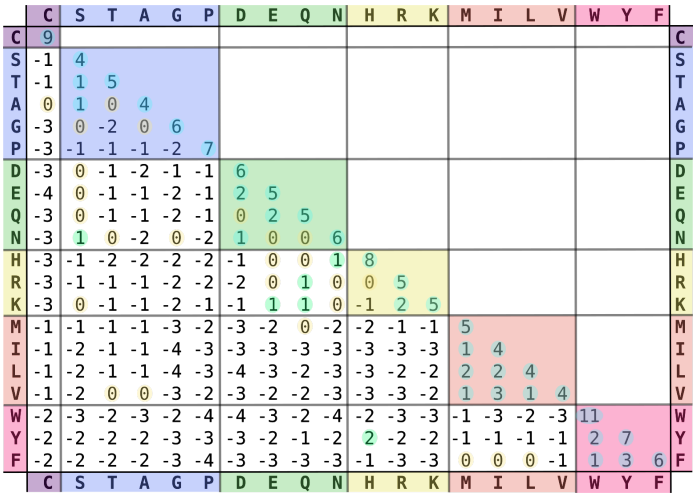
Las diferencias en las propiedades fisicoquímicas llevan a importantes diferencias en la estructura de las proteínas y hacen que los cambios observados (sustituciones) entre proteínas homólogas (descendientes de un ancestro común) tengan distribuciones particulares, las que se suelen representar en matrices de sustituciones[14] (ver figura 1.5). Estas diferencias, por lo tanto, tendrán un gran impacto en muchos de los procesos evolutivos, así como en su reconstrucción. De hecho, algunos AA poseen bastante similaridad estructural y fisicoquímica con otros, por ejemplo, ***Phe*** con ***Tyr*** e ***Ile*** con ***Val***, lo que hace que estos sean más propensos a intercambiarse cuando las presiones mutacionales son suficientemente fuertes.

Por ejemplo, mientras que en el caso de ***Phe*** con ***Tyr*** no hay diferencias en el contenido GC de los codones implicados, en el segundo caso (***Ile*** con ***Val***) se trata del cambio A↔G en primera posición del codón y, por lo tanto, estará sujeto a cambio cuando el sesgo mutacional sea hacia GC (AT). Otros casos de cambios relativamente frecuentes incluyen ***Asp*** con ***Glu***, ***Glu*** con ***Gln***, ***Ile*** con ***Leu***, ***Arg*** con ***Lys***, ***Leu*** con ***Met***, ***His*** con ***Tyr*** y ***Trp*** con ***Tyr***. De todos estos solo el último implica un cambio de dos bases, el resto de solo una base y casi todos independientes del contenido GC (de todos estos solo ***His*** con ***Tyr*** implica cambio de GC). Esto podría estar relacionado con un fuerte proceso selectivo para la optimización de la estructura del código genético (Novozhilov, Wolf, and Koonin 2007), aunque existen teorías alternativas.

Bajando un poco más el nivel de conservación (pero aún con *score* positivo en la matriz BLOSUM62) tenemos varios cambios que requieren cambio de contenido GC en los codones implicados.

Un recurso extremadamente útil para comparar las propiedades de los aminoácidos, así como para trabajar a partir de ellas, es la «AAindex database» (http://www.genome.jp/aaindex/) (Kawashima, and Kanehisa 2000). En esta se encuentra registrada información

[14] Las matrices de sustituciones aminoacídicas son matrices que reflejan la frecuencia evolutiva de sustitución de pares de aminoácidos. Existen diferentes variantes (por ejemplo, matrices PAM, matriz BLOSUM62), dependiendo del set de proteínas y las asunciones con las que estas fueron construidas. Estas matrices juegan un rol clave en la implementación de algoritmos de alineamiento de proteínas, como puede ser el empleado por el programa BLAST mencionado anteriormente.



	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

Figura 1.5. Matriz de sustitución aminoacídica BLOSUM62. Las letras en filas y columnas de la matriz refieren a los veinte aminoácidos naturales codificados directamente, y su color refleja sus tipos estructurales y fisicoquímicos. Las entradas numéricas de la matriz son proporcionales al logaritmo de la frecuencia con la que se observa la sustitución de un aminoácido por otro en proteínas homólogas. Valores positivos denotan mayores frecuencias de sustitución; lo contrario ocurre para valores negativos. Figura tomada de Wikipedia (CC BY-SA 3.0), versión original del archivo «Blosum62-dayhoff-ordering.svg»

para unas 500 propiedades diferentes, así como información adicional acerca de los autores, la publicación original y otras auxiliares. La posibilidad de acceder a estas propiedades a través de la biblioteca `seqinr` de R facilita enormemente la exploración de hipótesis más complejas que se relacionen con la sustitución de AA (ver, por ejemplo, Spangenberg et al. 2011).

Diferencias de composición entre clases de proteínas

Existen diferentes clases de proteínas de acuerdo con el criterio que se utilice para clasificarlas y, en general, de acuerdo con el criterio usado poseen una composición diferente. Por ejemplo, de acuerdo con la forma podemos clasificarlas en *fibrosas*, *globulares* y *mixtas*. Las fibrosas presentan largas cadenas polipeptídicas, estructura secundaria atípica, son insolubles en agua y en disoluciones acuosas (por ejemplo, queratina, colágeno y fibrina). Las globulares pliegan sus cadenas en una forma esférica apretada, con los grupos hidrófobos hacia el interior y grupos hidrófilos hacia afuera. Esto hace que sean solubles en disolventes polares como el agua. Entre ellas se encuentran los anticuerpos, las enzimas, algunas hormonas y proteínas de transporte. Finalmente, las mixtas poseen una parte fibrilar y otra parte globular.

Las proteínas solubles en agua suelen poseer los residuos hidrofóbicos (***Leu***, ***Ile***, ***Val***, ***Phe*** y ***Trp***) escondidos en el medio de la proteína (en su estructura), mientras que los hidrofílicos se encuentran expuestos en la superficie. Al mismo tiempo, las proteínas de membrana (IMP) suelen contar con anillos externos hidrofóbicos, los cuales sirven de anclaje en la bicapa lipídica. Por otro lado, proteínas que deben interactuar con moléculas de carga positiva

suelen tener residuos de carga negativa en su superficie (***Glu*** y ***Asp***) y viceversa (***Arg*** y ***Lys***). Algunos AA poseen propiedades muy particulares, con un efecto sustancial en la estructura de la proteína. ***Pro***, por ejemplo, que forma codos en la estructura y ***Cys***, que forma puentes di-sulfuro estabilizando estructuras.

Un grupo de particular interés son las proteínas de membrana, que juegan un papel muy importante en el intercambio de moléculas entre la célula y el exterior, así como en la transducción de señales. Por esta razón es muy importante poder predecir qué secuencias se corresponden con IMP. Existen básicamente dos tipos de IMP, las de α-hélices (TMH) y las de barriles-β (TMB). Las primeras se encuentran en la membranas internas bacterianas y en las membranas plasmáticas bacterianas, aunque a veces también en membrana externa. Es la mayor categoría de proteínas transmembrana y una parte importante de todas las proteínas humanas. Las proteínas de conformación tipo barriles-β se encuentran solo en la membrana externa de bacterias Gram-negativas, pared celular de Gram-positivas y membrana externa de organelos.

Aunque existen distintos algoritmos y programas, la predicción de la pertenencia a alguna de estas dos clases se basa esencialmente en el mismo paradigma, básicamente el alineamiento a otras proteínas o estructuras conocidas (Bigelow and Rost 2009). Un concepto clave en este sentido es el de «transferencia de homología a través de alineamientos». Bajo este paradigma, dos proteínas homólogas poseerán un alineamiento razonable. Por lo tanto, es posible transferir información derivada de evidencia experimental de una proteína a otra de interés para la que no se posee información. Claramente la capacidad de extender las conclusiones depende de la calidad del alineamiento. Ejemplo de esto serían regiones transmembrana en proteínas TMH: existen métodos específicos para la predicción de ambas clases de IMP. Los métodos para detección de TMH se basan en la presencia de una o más hélices transmembrana. Dentro de estos métodos se encuentra TMHMM (http://www.cbs.dtu.dk/services/TMHMM/).

Entre los programas que permiten discriminar entre proteínas TMB o no-TMB se encuentra TMB-Hunt, que utiliza un algoritmo k-Nearest Neighbour (k-NN) sobre la composición aminoacídica de las proteínas. El mismo permite además incluir pesos diferenciales para los aminoácidos, así como información evolutiva, lo que consigue una precisión mayor al 92 % con una sensibilidad mayor al 91 % (Garrow, Agnew, and Westhead 2005). Finalmente, una alternativa interesante es utilizar InterScanPro (http://www.ebi.ac.uk/Tools/pfa/iprscan/), una herramienta que combina la información de hasta 14 programas de predicción diferentes. Los programas que se pueden seleccionar a través de esta herramienta son: *BlastProDom* (busca en la base de dominios ProDom), FPrintScan (busca en la base PRINTS de grupos de motivos), HMMPIR (busca en modelos HMM de la Protein Sequence Database, PSD), HMMPfam (busca en modelos HMM de la base PFAM), HMMSmart (busca en modelos HMM de la base SMART), HMMTigr (busca en modelos HMM de la base TIGRFAMs), ProfileScan (busca contra perfiles de PROSITE), HAMAP (busca contra perfiles de HAMAP), PatternScan (nueva versión de búsqueda sobre PROSITE), SuperFamily, SignalPHMM, TMHMM, HMMPanther y Gene3D. El costo computacional de estas búsquedas es muy elevado, por lo tanto, se debe ser muy cauteloso a la hora de enviar varios trabajos simultáneamente.

Ejemplo 1.4

Escherichia coli es una bacteria común en el intestino de los mamíferos. En la misma, el regulón del fosfato comprende un conjunto de más de 20 genes no-ligados cuyos promotores pueden ser regulados simultáneamente por controles superpuestos y separados. En particular, los genes **phoS**, **phoA** y **phoE** son inducidos por la privación de fosfato y están sujetos a un control positivo por el producto del gen **phoB** y a un control negativo por el producto del gen **phoR**. Por otra parte, el sistema de transporte específico de fosfato (Pst) está codificado por cuatro genes estrechamente relacionados entre sí y situados juntos: **phoS**, **phoT**, **pstA** y **pstB**.

La siguiente es la secuencia de aminoácidos correspondientes al gen **pstA**, un transportador integral de membrana (Accesion K01992 del GenBank):

MAMVEMQTTAALAESRRKMQARRRLKNRIALTLSMATMAFGLFWLI
WILMSTITRGIDGMSLALFTEMTPPPNTEGGGLANALAGSGLLILW
ATVFGTPLGIMAGIYLAEYGRKSWLAEVIRFINDILLSAPSIVVGL
FVYTIVVAQMEHFSGWAGVIALALLQVPIVIRTTENMLKLVPYSLR
EAAYALGTPKWKMISAITLKASVSGIMTGILLAIARIAGETAPLLF
TALSNQFWSTDMMQPIANLPVTIFKFAMSPFAEWQQLAWAGVLIIT
LCVLLLNILARVVFAKNKHG

Si realizamos el análisis composicional de la misma llegamos la siguiente tabla:

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
38	1	3	11	13	21	2	27	10	41	17	9	12	8	13	16	22	18	9	5

El largo total es de 296 aminoácidos. La suma de los 4 aminoácidos hidrofóbicos más importantes (**A**, **I**, **L** y **V**) es de $38 + 27 + 41 + 18 = 124$, es decir, $\frac{124}{296} = 41{,}9\,\%$ del total de aminoácidos. Si bien la comparación tiene poco sentido si no conocemos la composición global del genoma (es decir, del resto de las proteínas del genoma), sin duda es una buena pista de que se trata de una proteína de membrana (donde los aminoácidos hidrofóbicos juegan un papel relevante debido a la naturaleza lipídica de la bicapa que forma la membrana).

PARA RECORDAR

- A nivel genómico la evolución se da por un conjunto de eventos como las mutaciones puntuales, las duplicaciones de genes y regiones o las inserciones y deleciones de secuencias nucleotídicas. El estudio de las variaciones en la composición de los genomas es lo que se puede llamar **genómica composicional** y forma parte de la **genómica comparativa**.
- La idea del **análisis de correspondencia** es, dada una tabla de $n \times p$ (casos × variables), encontrar un sistema de k ejes ortogonales (con $k < min(n,p)$) que condense el máximo de variación en los datos originales. **Esto es posible en la medida en que las variables (casos) estén correlacionados en alguna medida** y cuanto mayor sea esa asociación más varianza será condensada en pocos ejes (dimensiones).

- La métrica adoptada en el **COA** es la de la distancia χ^2, o sea, para la celda correspondiente a la fila i y columna j, la distancia será $d_{ij} = (O_{ij} - E_{ij})^2/E_{ij}$ donde O_{ij} es el número observado y E_{ij} el número esperado para esa celda, usualmente el número total de observaciones multiplicado por las frecuencias relativas marginales en la tabla. El punto importante es determinar con cuántos ejes (nuevas dimensiones) nos vamos a quedar. Cuantas más dimensiones retengamos mayor cantidad de la varianza original retendremos, pero al costo de dificultar la interpretación.
- Desde el punto de vista biológico, lo relevante del **COA**, además de identificar si podemos condensar la varianza en pocos ejes, es poder **etiquetar** cada eje con uno o unos pocos factores biológicos relevantes, tanto sean composicionales, de estructura del código, ambientales, etcétera.
- Los distintos aminoácidos poseen diferentes propiedades que permiten agruparlos en diversos grupos, de acuerdo con las mismas. Por ejemplo, tenemos aminoácidos polares, apolares, hidrofílicos, hidrofóbicos, pequeños, medianos y grandes, alifáticos y aromáticos, etcétera. La importancia de esto es que permiten generar proteínas con diferentes propiedades y estructura tridimensional.
- La redundancia en el código genético permite que un mismo aminoácido sea codificado por diferentes codones. Entre otras cosas, eso permite una mayor tolerancia a las mutaciones, ya que mutaciones que cambian el codón pero no el aminoácido no suelen tener mayor impacto en la estructura de la proteína (aunque puede ocurrir esto último debido a diferencias en la velocidad de traducción y su impacto en el plegamiento de la proteína).
- El contenido G+C de los genomas es una variable extremadamente importante, ya que se encuentra directamente asociada, entre otras cosas, a la composición de las proteínas. Diferentes hipótesis se han propuesto para explicar la enorme variabilidad entre los procariotas (20 % – 80 %), pero en esencia se dividen en **neutralistas** y **seleccionistas**.

1.3 Genómica comparativa

Como vimos antes, solo es posible entender la biología de los organismos si entendemos los procesos evolutivos subyacentes (Dobzhansky 1964). De hecho, pese a que nuestra visión antropocéntrica nos incline a vernos sustancialmente diferente de otras especies, desde el punto de vista genómico compartimos una enorme similaridad con otras especies. A nivel genómico la similitud nucleotídica es mayor al 98 % en algunos casos, como, por ejemplo, con el chimpancé (*Pan troglodytes*) (Sequencing and Consortium 2005). Las diferencias que observamos entre las especies pueden, en general, correlacionarse con diferencias a nivel genómico, tanto por la presencia/ausencia de genes así como diferencias en las secuencias que producen proteínas con diferencias estructurales/funcionales o modificaciones en las redes regulatorias. La genómica comparativa trata precisamente de identificar diferencias en los genomas, entender los procesos evolutivos que llevan a esas diferencias y en algunos casos de correlacionar estas diferencias con el fenotipo o aun el ambiente donde se desarrollan los organismos.

Comparación de secuencias génicas

La base de la genómica comparativa se encuentra en la determinación de homología entre secuencias, es decir la inferencia sobre el origen ancestral compartido de las mismas. En especial, la determinación de la relación de ortología (ancestría común por especiación, ver nota al pie 15) juega un papel central en la genómica comparativa ya que suele implicar (aunque no siempre) la realización de la misma función en diferentes organismos. La determinación de ortología es un problema usualmente complejo, particularmente cuando hay varias especies implicadas. La base de este proceso es la identificación de secuencias, en todas las especies de interés, que presenten un nivel de similaridad tal que permita inferir la relación de ortología. Debido a que es usualmente necesario explorar muchos pares (solo la comparación de dos bacterias requiere del orden de 10^7 comparaciones), se requiere de herramientas que identifiquen similaridad rápidamente, lo que lleva al tipo de alineamientos heurísticos como el del programa BLAST.

Usualmente, las secuencias ortólogas son las más similares en las otras especies. Esto ha llevado al criterio práctico del *Reciprocal Best Hit (RBH)*: un par de secuencias se consideran ortólogas si cada una de ellas es la mejor «selección» (la más similar) en su correspondiente genoma, de la otra. Un criterio de este tipo es necesario para evitar relaciones confusas debido a la existencia de parálogos.[15] Cuando se comparan dos especies este criterio está suficientemente bien definido; sin embargo, cuando se trata de más de dos especies la existencia de parálogos en alguna de ellas puede llevar a identificar distintas secuencias como pertenecientes al grupo de ortólogas en dicha especie de acuerdo con qué pares de especies se tomen en cuenta.

Este problema nos lleva a otros dos criterios de utilidad práctica a la hora de identificar ortólogos. Por un lado, cuando la similaridad entre secuencias es debida a eventos de especiación (definición de ortólogos, presentada anteriormente) el orden de los distintos genes dentro del genoma será relativamente similar durante mucho tiempo. Esto se debe a que los re-arreglos genómicos suelen tener un importante costo evolutivo; en procariotas en particular existen estructuras llamadas operones, que facilitan la expresión coordinada de un conjunto de genes cuyos productos interactúan directamente y que tienden a conservarse en la evolución. Este ordenamiento puede ser usado para determinar las relaciones de ortología, ya que además del alto nivel de identidad de las secuencias podemos requerir que los genes en el vecindario de ambas especies sean los mismos y estén ordenados de similar manera. Todo esto es altamente dependiente de la distancia evolutiva entre los organismos considerados, ya que a mayor tiempo de divergencia menos conservación de estos patrones podemos esperar.

La otra aproximación, conocida como *filogenómica*, implica una cuidadosa reconstrucción filogenética,[16] a partir de la cual se intenta conciliar los resultados de los alineamientos de

[15] Se dice que dos secuencias son parálogas si son producto de una duplicación ancestral. Un caso clásico para ilustrar este concepto es el de la familia de las hemoglobinas. Existen genes codificantes para α y β-hemoglobina en humano y chimpancé, y se ha inferido que estos genes también existían en el ancestro común a ambas especies. Los genes de α-hemoglobina y los de β-hemoglobina de las especies son ortólogas entre sí. Sin embargo, diríamos que la α-hemoglobina de humano es paráloga a la β-hemoglobina de chimpancé, ya que su «parentesco» no se debe al proceso de especiación, si no a la duplicación del gen codificante para la hemoglobina ancestral.

[16] Una filogenia representa la historia evolutiva inferida para un conjunto de diferentes entidades, incluyendo patrones de ancestría y las tasas de cambio que llevaron a la diversidad observada. Si bien el ejemplo más paradigmático puede ser el de una filogenia de especies (donde se representa la evolución del grupo) también es posible pensar en una filogenia de genes o cualquier otra entidad autorreplicante que se modifique a lo largo del tiempo.

las secuencias homólogas. Con un poco de suerte es posible determinar correctamente las relaciones de ortología y paralogía de nuestras secuencias. Los resultados obtenidos son altamente dependientes de la reconstrucción filogenética, por lo que es necesario tener mucho cuidado en la selección de secuencias para realizarla, en la correcta selección del modelo de sustitución (tanto aminoácidica como nucleotídica) y el algoritmo de reconstrucción filogenética. En la medida de lo posible, es bueno explorar la sensibilidad del análisis frente a posibles cambios en el proceso (por ejemplo, la selección de otro método de reconstrucción).

Existe una gran cantidad de software específico para atacar el problema de la detección y determinación de ortólogos, así como bases de datos para distintos grupos filogenéticos. Entre estas últimas podemos destacar COG (Cluster of Orthologous Groups, http://www.ncbi.nlm.nih.gov/COG/), una base de datos construida a partir de RBH con consistencia en al menos tres linajes para procariotas y su extensión eucariótica KOG, así como EGO (http://compbio.dfci.harvard.edu/tgi/ego/) una base de ortólogos para eucariotas. Entre los programas podemos mencionar Inparanoid (http://inparanoid.sbc.su.se/) que consiste en RBH seguido de agrupamiento de in-parálogos (parálogos dentro de una especie), así como OrthoMCL (http://www.orthomcl.org/), que a diferencia del primero usa **E-value** y porcentaje de identidad para asignar las proteínas a diferentes agrupamientos. Finalmente, un programa muy interesantes es Orthostrapper (http://sonnhammer.sbc.su.se/download/software/), y su visualizador OrthoGUI, que permite obtener valores de soporte estadístico para la ortología de cada par de secuencias en un alineamiento.

Selección positiva

Otro ejemplo interesante de las posibilidades de la genómica comparativa es la determinación de sitios en el genoma (o genes) que se encuentran evolucionando bajo selección positiva o, dicho en otras palabras, donde la selección se orienta hacia una adaptación a un ambiente particular (mejora en el fitness). Existen diversas formas de estudiar el efecto de la selección a nivel molecular, pero uno de los usualmente más aceptados es el de la relación $\frac{K_a}{K_s}$ (también conocida como ω, o dN/dS). Esta es la relación entre el número de sustituciones no-sinónimas por sitio no-sinónimo con respecto al número de sustituciones sinónimas por sitio sinónimo. Como las sustituciones sinónimas es de esperar que sean encontradas a una tasa usualmente mucho más alta que las no-sinónimas (ya que no generan cambios aminoacídicos y, por lo tanto, no repercuten en la función proteica), aquellos sitios con una relación $\frac{K_a}{K_s} > 1$ se consideran bajo selección positiva. El problema principal radica en determinar cuáles apartamientos del 1 resultan estadísticamente significativos.

Existen diferentes formas de estimar K_a y K_s, los más complejos usando estimación por máxima-verosimilitud,[17] pero un método aproximado suele dar resultados equivalentes cuando los datos son suficientes. En esencia, en el método aproximado, a partir de un alineamiento de secuencias, se realiza el conteo de sustituciones sinónimas y no-sinónimas, con corrección por sustituciones múltiples, para luego calcular la relación $\frac{K_a}{K_s}$. Es importante tener en cuenta que el alineamiento debe realizarse a nivel de aminoácidos, para luego realizar la traducción reversa a nucleótidos, asegurándose de esta forma que la comparación se realiza codón a codón. Dentro del software clásico para estimar esta relación podemos

[17] Dado un modelo probabilístico, la estimación de parámetros por máxima-verosimilitud es un método que busca estimar los valores que maximizan la posibilidad de observar un conjunto de datos dado según el modelo.

mencionar PAML (http://abacus.gene.ucl.ac.uk/software/paml.html) que realiza la estimación por máxima-verosimilitud, mientras que entre las fáciles de usar podemos mencionar la función *KaKs* de la biblioteca `seqinr` de R.

Vías

Las proteínas suelen realizar sus funciones particulares como partes de procesos más complejos, que envuelven otras proteínas, en diferentes etapas de esos procesos. El conjunto de proteínas (u otras moléculas) que participan en un determinado proceso se conocen como vía. A modo de ejemplo podemos recordar la vía de las pentosas fosfato, durante la cual se utiliza la glucosa para generar ribosa y poder reductor a través de NADPH. Las distintas vías existentes en cada organismo se constituyen en el «mapa» de los distintos procesos biológicos que cada organismo es capaz de realizar. Si bien el número de organismos donde todos los procesos bioquímicos son conocidos al detalle es relativamente escaso, la genómica comparativa es una alternativa importante para determinar la capacidad de realización de distintos procesos en otros organismos.

La idea central en esta estrategia consiste en ubicar en el genoma de interés ortólogos de los genes que constituyen una vía en un organismo modelo, como puede ser *E. coli*. Asumiendo que los genes ortólogos posean aún la misma función, la completitud de la vía (es decir, la presencia en el organismo de interés de todos los genes que la constituyen en el modelo) permite inferir la capacidad de realizar el proceso bioquímicos en el organismo de interés. Existen diferentes bases de datos con representaciones de las diferentes vías en varios organismos, pero probablemente la más conocida es KEGG Pathway (http://www.genome.jp/kegg/pathway.html). Esta base permite la búsqueda por diferentes criterios, posee información de varios organismos, muestra la información en forma gráfica y existe una API (*Application Programming Interface*) que facilita el acceso a los datos a través de software. Por otro lado, la existencia de información de las diferentes vías puede ser usada durante el proceso de anotación genómica, ya que si conocemos que un organismo realiza determinado proceso bioquímico entonces deben estar presentes todos los genes que constituyen esa vía. Si bien en algunos organismos las vías pueden estar formadas por diferentes genes, en general es posible determinar la falta de genes claves en las distintas vías, lo que usualmente se debe a errores en el proceso de anotación o a cambios desconocidos en las vías (esto último, de gran interés científico).

PARA RECORDAR

- La **genómica comparativa** trata de identificar diferencias en los genomas, entender los procesos evolutivos que llevan a esas diferencias y en algunos casos correlacionar estas diferencias con el fenotipo o aún el ambiente donde se desarrollan los organismos.
- La base de la genómica comparativa se encuentra en la determinación de homología en secuencias, es decir la inferencia sobre el origen ancestral compartido de estas. En especial, la determinación de la relación de ortología (ancestría común por especiación, mencionado anteriormente) juega un papel central en la genómica comparativa, ya que suele implicar (aunque no siempre) la realización de la misma función en diferentes organismos. Usualmente las secuencias ortólogas son las más similares en las otras especies y

esto ha llevado al criterio práctico del **Reciprocal Best Hit (RBH)**, es decir, un par de secuencias se consideran ortólogas si cada una de ellas es la mejor «selección» (la más similar) en su correspondiente genoma, de la otra. Cuando la similaridad entre secuencias es debida a eventos de especiación el orden de los distintos genes dentro del genoma será relativamente similar durante mucho tiempo.

- La aproximación **filogenómica** implica una cuidadosa reconstrucción filogenética, a partir de la cual se intenta conciliar los resultados de los alineamientos de las secuencias homólogas. Los resultados obtenidos son altamente dependientes de la reconstrucción filogenética, por lo que es necesario tener mucho cuidado en la selección de secuencias para realizar la misma, en la correcta selección del modelo de sustitución (tanto aminoácidica como nucleotídica) y el algoritmo de reconstrucción filogenética.
- Existen diversas formas de estudiar el **efecto de la selección a nivel molecular**, pero uno de los usualmente más aceptados es el de la relación $\frac{K_a}{K_s}$, que es la relación entre el número de sustituciones no-sinónimas por sitio no-sinónimo con respecto al número de sustituciones sinónimas por sitio sinónimo. Como las sustituciones sinónimas es de esperar que sean encontradas a una tasa usualmente mucho más alta que las no-sinónimas, **aquellos sitios con una relación $\frac{K_a}{K_s}$ por encima de 1 se consideran bajo selección positiva.** El problema principal radica en determinar cuáles apartamientos del 1 resultan estadísticamente significativos.

1.4 Genómica funcional

Las proteínas son los principales efectores a nivel celular (por ejemplo, las enzimas), así como también juegan un rol fundamental en las estructuras celulares y las diferentes maquinarias, por lo que si intentamos entender el funcionamiento de células y tejidos a este nivel es necesario conocer cuándo, cuánto y dónde se producen. Por ejemplo, si queremos ver las diferencias entre un tejido determinado en estado normal y en el mismo tejido pero con un tumor, podríamos medir el nivel de cada proteína en ambos tejidos y comparar luego los niveles correspondientes para entender las diferencias. Desafortunadamente, la medición de proteínas a escala masiva (lo que se conoce como **proteómica**) es un tema aún relativamente en pañales, si bien existen cientos de estudios desarrollados a esta escala en diferentes especies. Se trata de técnicas caras y complejas, que no han experimentado aún una explosión similar a la que se experimentó a nivel de la medición de ácidos nucleicos.

De acuerdo con una versión esquemática del paradigma de la biología molecular (paradigma que imperó durante muchos años), la información se almacena en el ADN, la que luego es transcrita hacia el ARN y que luego se traduce en proteínas. Si juntamos esto con el hecho de que es relativamente sencillo medir cantidades de **ARN** en forma masiva (en realidad **ADNc**), la idea de utilizar la medición de los transcritos como un ***proxy*** a la medición de proteínas surge de inmediato. Es decir, si suponemos que la tasa de producción de proteínas es directamente proporcional a la tasa de transcripción, entonces medir la cantidad de transcritos es una buena y fiable indicación de la producción de proteínas (aunque existe una compresión de rango entre el logaritmo de la relación de proteínas y el logaritmo de

la relación de mensajeros, como lo muestran Spangenberg et al. (2013). Claramente, a pesar de su enorme importancia en el desarrollo de la disciplina, esta visión simplificada del paradigma de la biología molecular ha sido ampliamente sobrepasada por la información generada en los últimos años, incluyendo los descubrimientos de los microARN y de un gran número de otros ARN que juegan un papel fundamental en la regulación de la expresión génica. Más aún, la vida media de los diferentes transcritos no es necesariamente igual, lo que conlleva un error a la hora de utilizar su cantidad (no su tasa de producción) como indicador de la producción de proteínas.

A pesar de todos estos problemas, el análisis de la expresión de los **ARN mensajeros** (**ARNm**) sigue siendo la principal herramienta para entender los cambios a nivel de expresión en los distintos tipos de órganos y tejidos, así como en células bacterianas o de otras especies. En un principio, los **microarrays** fueron fundamentales para esta tarea ya que permitían interrogar miles de genes al mismo tiempo. Desde los originales **spot arrays** que eran impresos por un robot cartesiano mediante un bloque de agujas que depositaban las sondas en una placa de vidrio, diferentes tecnologías se fueron desarrollando. Una de las principales limitaciones de los primeros arrays era la necesidad de utilizar algún tipo de «referencia», es decir algo contra lo que contrastar las mediciones debido a las imperfecciones de las impresiones con agujas. Posiblemente de las tecnologías más conocida es la desarrollada por la empresa **Affymetrix** (fundada por el bioquímico uruguayo Alejandro Zaffaroni) que utilizando la misma tecnología que se utiliza para la fabricación de chips de semiconductores (esencialmente la fotolitografía) consiguieron desarrollar arrays de altísima densidad. La figura 1.6 muestra dos arrays de esta marca, a la izquierda un chip para genoma humano, mientras que a la derecha aparece un chip para genoma de ratón; en la parte inferior aparece un fósforo a efectos de imaginar el tamaño de los mismos. La principal ventaja de estos arrays respecto a los típicos **spotted** era la muy alta precisión con los que eran fabricados, lo que permitía evitar tener que usar una referencia. Sin embargo, el principal problema con estos arrays era el costo de desarrollo asociado a cada nuevo diseño, por lo que quedaron reducidos a las pocas especies que eran modelo para la investigación (humano, ratón, *Arabidopsis thaliana*, etcétera).

Un poco más tarde, con la tecnología de la empresa **Agilent**, la **impresión láser** de los arrays permitió combinar una altísima precisión con la posibilidad de fabricarlos a gusto del consumidor y sin costos adicionales, lo que llevó a utilizarlos en decenas de nuevas especies. En efecto, en lugar de utilizar tinta de diferentes colores, el array se construía imprimiendo un nucleótido por vez en cada posición de una sonda, utilizando como tintas 4 soluciones con los nucleótidos A, C, G y T. La precisión de esta tecnología permitía realizar análisis en experimentos de un solo canal o en experimentos con dos canales, lo que en ciertos casos posibilitaba reducir los costos (por ejemplo, cuando existe una referencia natural para cada caso, utilizando el mismo array se elimina la necesidad de dos arrays).

Finalmente, en años recientes, las tecnologías de secuenciación masiva llegaron a este campo y mediante la secuenciación del **ARNc** permiten obtener una representación mucho más amplia y fidedigna de las proporciones de los distintos mensajeros. Además, como se obtienen las secuencias de los mismos, es posible identificar variantes (que son a su vez producto de variantes en el **ADN**) asociadas a cambios en la expresión entre individuos, lo que se conoce como **eQTL** (expression Quantitative Trait Locus). La principal diferencia entre el análisis de expresión con arrays respecto a la secuenciación (que se conoce como **RNA-seq**) es que en el primer caso la señal obtenida es analógica (un número en una escala continua),

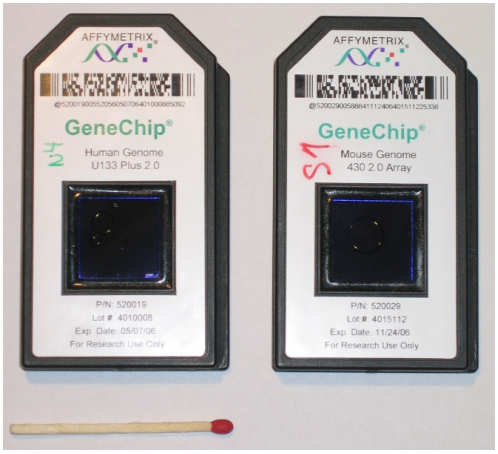


Figura 1.6. Dos chips de Affymetrix, uno para genoma humano, a la izquierda, y el otro para genoma de ratón. El fósforo debajo es solo a fines de comparación del tamaño. La fecha de expiración de ambos es en el año 2006, por lo que su uso seguramente fue anterior (Schutz, trabajo propio –based on copyright claims–, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=694717).

mientras que en el segundo caso es digital (en realidad el conteo de los reads que mapean a un transcrito particular).

La idea de las comparaciones es similar en las diferentes tecnologías, aunque cada tecnología requiere de detalles específicos a la hora de analizar los datos. En general, si vamos a comparar el perfil de expresión en dos tratamientos diferentes, por ejemplo, tejido sano y tumoral, requerimos de varios individuos en cada tratamiento a fin de tener en cuenta la variación entre individuos y que nuestras conclusiones se puedan extrapolar de alguna manera a la población general. Luego de haber obtenido los perfiles de expresión en cada individuo necesitamos realizar algún tipo de normalización, ya que las cantidades absolutas suelen depender de la preparación de las bibliotecas. Luego de normalizadas las muestras debemos comparar gen a gen, lo que suele hacerse mediante un modelo lineal sobre los datos transformados de alguna manera. En los casos en que se compara un tratamiento contra una referencia, se acostumbra a utilizar la distribución del logaritmo de la relación de cambio entre las dos condiciones (que se llama logFC). Es decir, si la expresión del gen i en el tratamiento es $\text{mRNA}_{i,1}$ y en la referencia o control es $\text{mRNA}_{i,2}$, entonces

$$logFC_i = log\left[\frac{mRNA_{i,1}}{mRNA_{i,2}}\right] \tag{1.12}$$

Las ventajas de esta transformación son varias. Por un lado, tanto las distribuciones de conteo (en el caso de RNA-seq) como las de señal luminosa tienen una cola larga a la derecha, es decir, una serie de valores muy altos y alejados de donde se concentra la mayor parte de la distribución; el logaritmo «corrige» de alguna forma eso, transformándolas en distribuciones parecidas a la normal. En segundo caso, mientras que la relación $\frac{mRNA_{i,1}}{mRNA_{i,2}}$ es asimétrica (cuando el numerador es más pequeño que el denominador la relación queda entre 0 y 1, mientras que al revés va de 1 a $+\infty$), la relación transformada $log\left[\frac{mRNA_{i,1}}{mRNA_{i,2}}\right]$ es simétrica y centrada en cero, además de relativamente normal. Con esta transformación, por ejemplo,

si hay solo dos grupos a comparar, la comparación podría ser a través de un test de *t* (la distribución). Como realizamos miles de comparaciones a la vez (una por gen), la significación estadística de cada gen debe ajustarse para múltiples ensayos, lo que usualmente se realiza a través de algún método como Benjamini-Hochberg.

Estudios epigenéticos y epigenómicos

Hasta ahora hemos manejado los aspectos de la genética que tienen que ver con la transmisión de información como algo bastante rígido y donde hay poco margen para cambiar lo recibido de los progenitores. Sin embargo, existen una serie de mecanismos que tienen que ver con la regulación que pueden transmitirse de padres a hijos y a su vez ser modificados posteriormente de recibidos. El estudio de estos mecanismos particulares es el objetivo de la epigenética y de la epigenómica. Básicamente, la epigenética se centra en los procesos que regulan la activación y desactivación de genes, tanto en lo que hace a cómo y cuándo se produce esto. Por otra parte, la epigenómica se refiere a los cambios globales que ocurren a nivel de genoma que permiten activar o desactivar un conjunto de genes, tanto a nivel de células como de organismos.

En general, los procesos epigenéticos son los responsables de controlar el crecimiento y desarrollo de un organismo. En situaciones particulares, la alteración del control puede llevar a procesos patogénicos, como, por ejemplo, el cáncer. Lo trascendente de este proceso es que factores ambientales pueden ser los causantes del desajuste en la regulación y, por lo tanto, en el desarrollo de patologías. Dentro de los factores ambientales conocidos, la exposición a determinadas sustancias químicas aparece como muy relevante, aunque también la dieta parece estar relacionada con alteraciones en la regulación. Además, existe cierta evidencia de que los cambios en la regulación podrían ser transmisibles.

Existen dos formas en las que el epigenoma es capaz de marcar el ADN para la regulación génica: a) la metilación y b) la modificación de las histonas. En la primera, grupos metilo se intercalan en la molécula de ADN modificando la función cuando esto ocurre en un gen promotor. En general, la metilación del ADN actúa reprimiendo la transcripción génica. En mamíferos entre el 60 % y el 90 % de todas las CpG están metiladas. Por otra parte, en la segunda forma de acción una serie de modificaciones en las histonas (acetilación, metilación, desmetilación, fosforilación, ubiquitilación, etcétera) modifican el arrollamiento del ADN y, por lo tanto, la disponibilidad para la transcripción.

Por supuesto, existen arrays disponibles para el estudio de las modificaciones epigenómicas, así como métodos para realizar los estudios a través de secuenciación masiva.

PARA RECORDAR

- Las proteínas son los principales efectores a nivel celular, así como también juegan un rol fundamental en las estructuras celulares y las diferentes maquinarias, sin embargo, desafortunadamente, la medición de proteínas a escala masiva (lo que se conoce como **proteómica**) es un tema aún relativamente nuevo: se trata de técnicas aún caras y complejas y que no han experimentado todavía una explosión similar a la que se experimentó a nivel de la medición de ácidos nucleicos.

- Como sabemos, la información se almacena en el **ADN**, la que luego es transcrita hacia el **ARN** y que luego se traduce en proteínas. Si a su vez consideramos el hecho de que es relativamente sencillo medir cantidades de ARN en forma masiva (en realidad ADNc), la idea de utilizar la medición de los transcritos como un proxy a la medición de proteínas surge de inmediato. Es decir, **si suponemos que la tasa de producción de proteínas es directamente proporcional a la tasa de transcripción, entonces medir la cantidad de transcritos es una buena y fiable indicación de la producción de proteínas.**

1.5 Conclusión

El estudio de la vida se ha nutrido a lo largo de los siglos del avance tecnológico; ejemplos claros de esto son la aparición del microscopio (que permitió vislumbrar que la vida habitaba lugares insospechados en formas diminutas) o los avances de la química, que permitieron comprender los mecanismos subyacentes al funcionamiento de la maquinaria celular y la herencia desde una perspectiva molecular. El uso de herramientas matemáticas también ha sido fundamental para el desarrollo teórico de la biología: pensemos en la importancia que tuvo la simple aproximación de Mendel al estudio de la herencia de caracteres.

No es de extrañar que en décadas marcadas por el avance de las ciencias de la informática y del procesamiento tecnológico de compuestos químicos también exista una nueva vigorización de nuestro entendimiento de la biología. En este capítulo presentamos varias de las áreas de estudio que han surgido o se han visto vitalizadas por la aparición de métodos de secuenciación masiva (por ejemplo, las distintas áreas de la genómica), discutiendo su impacto en la tipificación y estudio de la diversidad genética, la identificación y análisis evolutivo de organismos o el análisis de la expresión génica y uso de codones.

En el resto del libro trataremos dos áreas sumamente importantes de la genética: el estudio de la herencia a nivel poblacional y el análisis cuantitativo. Las tecnologías y la forma de análisis que vimos en este capítulo reaparecerán constantemente a medida que aprendamos cómo se observa la genética a través de estas dos potentes lentes conceptuales. Antes de seguir con ese recorrido, es importante remarcar nuevamente que estas dos áreas surgieron mucho antes que las tecnologías que mencionamos en este capítulo: una vez más en la historia de la biología, la aparición de tecnologías da nuevos datos y perspectivas de análisis a marcos teóricos que ya existían, ayudando a testear hipótesis, ampliar nuestro conocimiento e impulsando nuevas preguntas sobre el fenómeno de la herencia.

1.6 Actividades

1.6.1 Control de lectura

1. Existen muchas maneras de medir la diversidad biológica. Menciona al menos una, explicitando la forma de cálculo y en qué contexto puede ser útil la medida elegida.
2. ¿Qué es un microarray? Explica la utilidad de esta herramienta, poniendo algún ejemplo de su aplicación.
3. En las últimas décadas se han desarrollado un conjunto de disciplinas que permiten la obtención y análisis masivo de datos biológicos. Identifica el objeto de estudio de las siguientes:
    1. Genómica
    2. Transcriptómica
    3. Proteómica
    4. Metagenómica
    5. Metatranscriptómica
    6. Epigenómica
4. Actualmente se acepta que todos los organismos provenimos de un ancestro común (denominado LUCA). ¿Cómo se explica entonces que existan diferencias tan grandes entre los organismos vivos? Menciona al menos una aproximación para el estudio de esta diversidad biológica.
5. Define el concepto de codón. ¿Todos los organismos usan el mismo conjunto de codones? ¿Los codones son todos usados en igual proporción?

1.6.2 ¿Verdadero o falso?

1. Un codón solo codifica para un aminoácido, pero un aminoácido puede estar codificado por más de un codón.
2. La estructura del código genético refleja las propiedades fisicoquímicas de los aminoácidos codificados en el llamado «código genético universal».
3. La secuenciación y comparación de secuencias de ARNr 16S permite la identificación y estudio filogenético de todos los organismos vivos.
4. El contenido GC de los organismos no es uniforme, existiendo variaciones a lo largo del árbol de la vida.
5. Todos los cambios genéticos heredables se encuentran codificados en el genoma de los organismos, por lo que secuenciando y comparando secuencias genéticas podemos comprender en su totalidad el proceso de herencia de información entre individuos (y en las poblaciones).

Solución

1. **Verdadero.** Este concepto suele denominarse «redundancia» o «degeneración» del código genético. Entre otras cosas, eso permite una mayor tolerancia a las mutaciones, ya que mutaciones que cambian el codón pero no el aminoácido no suelen tener mayor impacto en la estructura de la proteína (aunque puede ocurrir esto último debido a diferencias en la velocidad de traducción y su impacto en el plegamiento de la proteína).

2. **Verdadero.** Los cambios sinónimos suelen llevar con más frecuencia a aminoácidos de características fisicoquímicas similares, hecho que podría adjudicarse a un régimen selectivo que favoreció el evitar que el proceso mutacional lleve a cambios radicales en las proteínas codificadas.
3. **Falso.** Si bien es cierto que la secuenciación de ARN ribosomal es útil para identificar tanto a organismos procariotas como eucariotas, es importante recordar que las subunidades secuenciadas varían en los dominios de la vida: en procariotas si se recurre al estudio de la subunidad 16S, mientras que en eucariotas se recurre al 18S (recordar que no existe subunidad 16S en eucariotas).
4. **Verdadero.** El contenido de GC (o lo que es lo mismo, AT) varía ampliamente entre organismos. Ya en la década de 1960 se determinó que existen bacterias cuyo genoma se encuentra en un rango que va desde 25 hasta 75 % de contenido GC. Tal es la diferencia, que en muchos casos es posible identificar transferencias de material genético entre diferentes especies bacterias observando que tramos del genoma desvían fuertemente de la tendencia general observada en el resto del genoma (se ahondará en este tema en el capítulo «Genética de poblaciones microbianas»).
5. **Falso.** Si bien este fue un paradigma dominante durante décadas, la existencia de transmisión de información epigenética (por ejemplo, a través de patrones de metilación y acetilación de histonas) muestra que es posible la herencia de información biológica que no se encuentra codificada a nivel nucleotídico. Disciplinas como la epigenómica intentan estudiar la relevancia de estos procesos en la herencia de los individuos e incluso la evolución.

1.6.3 Ejercicios

Ejercicio 1.1

Dada la siguiente secuencia «AUGGGAUUUCCGCCCAUUUACGUG»

a. Determina de qué ácido nucleico se trata (ADN o ARN). **b.** Determina si tiene sentido traducirla a una secuencia de aminoácidos tomándola como hebra **codificante** y en marco de referencia. **c.** Escribe en sentido 5'–>3' la hebra complementaria pero en ADN (hebra molde) **d.** Calcula el contenido G+C.

Solución

a. Se trata de ARN ya que entre sus bases aparece **U**racilo en lugar de **T**imina.

b. En principio parece tener sentido, por lo que asumiendo que está en marco de referencia correcto y que se trata de la hebra codificante, separamos en codones «AUG GGA UUU CCG CCC AUU UAC GUG», que utilizando el código genético se traduce como MGFPPIYV.

c. «CAC GTA AAT GGG CGG AAA TCC CAT».

d. 6G + 6C= 12 G o C en un total de 8 codones × 3 bases c/u = 24 bases, por lo tanto 12/24 = 0,5 o 50 %.

Ejercicio 1.2

Un investigador obtuvo la siguiente secuencia (según él) de un nuevo gen que sería responsable de la generación de radiactividad por parte de una especie bacteriana que lo posee (presentamos la secuencia ordenada en tripletes y en marco de lectura correcto, para mayor facilidad de visualización):

GTC TCT CGT GAT GCG GTT GTC ATA AGA GCT GGC TCT GAG CTG
AAT TGA AGC ATA TCC TTC GCA GTC TTA AAA GGG TTG GGA TTA
TCA ACC TCC CAC GCA ATA CCG GCG TTC AAC AGT TTA CCA TCA
TCT ACT GAC GCG TGA TGT TGC GGA GCG ACG GGA AGT GAT
AAA AGT TAC TTT ACA

La especie bacteriana de la que supuestamente fue aislado el gen posee un contenido GC del 80 %.

Por otra parte, un estudiante de su grupo afirma que no solamente no se trata de la secuencia de un nuevo gen, sino que no se trata de la secuencia de ningún gen. Más aún, plantea que esta secuencia fue generada de forma aleatoria por una máquina y que posiblemente el investigador ha sido engañado por alguien. El conteo de bases dio lo siguiente:

$$\begin{cases} A : 46 \\ C : 40 \\ G : 45 \\ T : 49 \end{cases}$$

mientras que el conteo de aminoácidos fue el siguiente:

$$\begin{cases} STOP : 2 \\ A : 7 \\ C : 2 \\ D : 3 \\ E : 1 \\ F : 3 \\ G : 5 \\ H : 1 \\ I : 3 \\ K : 2 \\ L : 5 \\ N : 2 \\ P : 2 \\ R : 2 \\ S : 11 \\ T : 4 \\ V : 4 \\ Y : 1 \end{cases}$$

a. ¿Quién de los dos parece tener razón? Argumenta las razones que te llevan a esa conclusión.

b. ¿Cuántas cisteínas hay codificadas en la secuencia y dónde se encuentran?

c. ¿Cuántas serinas y cuántas argininas serían de esperar en una secuencia de este largo si se tratase de una secuencia creada al azar y cuántas fueron observadas? ¿Y la suma de las dos (es decir, para los codones de los sextetos)?

d. Determina mediante un test estadístico si el número de serinas y de argininas observadas cae dentro de lo esperado, o si debemos rechazar dicha hipótesis nula.

Solución

a. Algunos hechos hacen sospechar que no se trata de una secuencia presente en un organismo vivo tal y como los conocemos:

- El codón de iniciación no es ATG ni tampoco ninguna de las formas alternativas más usuales (este argumento no es excluyente, pero es raro de observar).
- En la posición 46, es decir el inicio del codón 16, corresponde con el codón TGA que es un codón terminación, lo que nos dejaría con una proteína de 15 aminoácidos de largo. Lo mismo ocurre en la posición 139, es decir el inicio del codón 47, por lo que tampoco parece mucho mejor la situación del fragmento interno entre ambos codones STOP.
- La secuencia no termina en ninguno de los codones STOP conocidos.
- La distribución de la frecuencia de bases no se aparta mucho de lo esperado para una secuencia obtenida al azar ($\frac{180}{4} = 45$). En cambio, si el contenido del genoma bacteriano fuese de 80 %, entonces esperaríamos $1800{,}802 = 72$ bases para G y lo mismo para C, mientras que esperaríamos $1800{,}202 = 18$ bases para A y lo mismo para T.

b. Dos cisteínas, y se encuentran codificadas por los codones 48 y 49.

c. La serina está codificada por 6 codones y Arginina por otros 6 codones, de los 64 que hay en el código genético o, lo que es lo mismo, de $4^3 = 64$ posibles combinaciones de cada bases (A, C, G, T) en 3 posiciones. Como hay un total de 60 codones en la secuencia, esperamos: $\text{Ser} = 60\frac{6}{64} = 5{,}625$, $\text{Arg} = 60\frac{6}{64} = 5{,}625$. Se observaron 11 y 2 respectivamente, lo que en ambos casos parece algo alejado del esperado. Sin embargo, la suma de ambos sextetos da 13, mientras que el esperado es $\text{Ser} + \text{Arg} = 60\frac{12}{64} = 11{,}25$, lo que parece mucho más cercano.

d. Un test de chi-cuadrado de *goodness-of-fit* (bondad de ajuste) podría ser adecuado en este caso.

Ejercicio 1.3

Observando la tabla con el código genético universal, agrupa los aminoácidos por el número de codones que los codifican y resume el número de aminoácidos con cada multiplicidad. ¿Qué característica comparten la mayor parte de los duetos (los aminoácidos codificados por dos codones)?

Solución

- 1 codón: **Met, Trp**
- 2 codones: **Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys**
- 3 codones: **Ile**
- 4 codones: **Val, Pro, Thr, Ala, Gly**
- 6 codones: **Leu, Arg, Ser**

La característica que comparten 7 de 9 duetos es que se encuentran en la tercera columna de la tabla del código, es decir, tienen una A en segunda posición del codón.

Ejercicio 1.4

En un planeta de otra galaxia existe vida con algunas similitudes a la que existe en nuestro planeta. En particular, también las proteínas parecen ser la base estructural y funcional de la vida, mientras que los ácidos nucleicos son los responsables del mantenimiento y duplicación de la información. Sin embargo, a diferencia de nuestro planeta, los aminoácidos proteinogénicos codificados son 24, mientras que las bases nucleotídicas análogas a las de nuestro ADN son 5. Suponiendo que los mecanismos moleculares en dicho planeta han sido seleccionados para ser altamente eficientes:

a. ¿De qué largo mínimo deberían ser los codones?

b. Si el código genético en ese planeta tuviese algún nivel de redundancia (más de un codón para el mismo aminoácido), ¿de qué largo mínimo serían los codones? En ese caso, ¿cuál sería el número promedio de codones por aminoácido?

c. ¿Cómo se compara esa redundancia promedio con la existente en nuestro planeta?

d. ¿Qué ventaja potencial tiene la redundancia en el código?

Solución

a. Si existen 5 bases nucleotídicas y 24 aminoácidos, entonces con codones de largo 2 tenemos $5 \times 5 = 25$ combinaciones posibles. Si consideramos la existencia de un codón **STOP**, entonces el mínimo es $24 + 1 = 25$ codones, por lo que alcanza con largo de codón 2.

b. Si queremos algo de redundancia vamos a precisar un código genético con largo mayor a 2. El siguiente largo es 3, por lo que ahora tenemos $5 \times 5 \times 5 = 5^3 = 125$ codones. Como hay 24 AA $+1$ **STOP** $= 25$ «necesidades», entonces la redundancia promedio será igual a $\frac{125}{25} = 5$.

c. En nuestro planeta tenemos 20 aminoácidos proteinogénicos codificados directamente, por lo que equivale a 20 AA $+1$ STOP = 21 «necesidades». Sin embargo, al existir solo

4 bases nucleotídicas en nuestro código y el largo de nuestros codones ser de 3, entonces tenemos un total de $4 \times 4 \times 4 = 4^3 = 64$ posibilidades. Por lo tanto, en nuestro planeta, la redundancia promedio es de $\frac{64}{21} \approx 3{,}05$.

d. La redundancia en el código tiene como ventaja que es capaz de absorber determinadas mutaciones (cambio de bases) sin cambiar el aminoácido codificado, lo que lo hace un sistema con cierta robustez o tolerancia a los fallos en la duplicación de la información (mutaciones).

Ejercicio 1.5

Asumiendo que existen 20 aminoácidos (los estándares),

a. ¿Cuántas proteínas diferentes de largo 5 aminoácidos sería posible tener?

b. ¿Cuántas serían posibles de largo n?

c. Considerando que una proteína de largo promedio en procariotas es del orden de los 300 aminoácidos, ¿cuántas proteínas diferentes, con ese largo se podrían codificar?

Solución

a. Se podrían codificar $20 \times 20 \times 20 \times 20 \times 20 = 20^5 = 3.200.000$.

b. De largo n se pueden codificar 20^n.

c. En el caso de $n = 300$, tenemos $20^{300} \approx \infty$ para los fines prácticos.

Parte II: Genética de poblaciones

Capítulo 2

Variación y equilibrio de Hardy-Weinberg

> *Do you mean, Hardy, if I said that two and two make five that you could prove any other proposition you like?». Hardy said, «Yes, I think so». «Well, then, prove that McTaggart is the Pope». «Well,» said Hardy, «if two and two make five, then five is equal to four. If you subtract three, you will find that two is equal to one. McTaggart and the Pope are two; therefore, McTaggart and the Pope are one».*[1]
>
> **Ronald A. Fisher**, 1958

Una de las características evidentes de la vida en nuestro planeta es la heterogeneidad de los organismos, aún dentro del nivel taxonómico más bajo que se nos ocurra para clasificarlos razonablemente (por ejemplo, especie o subespecie). Es decir, aún a simple vista podemos usualmente distinguir esa «falta de uniformidad» como una de las características **aparentemente** distintivas de los seres vivos. Pensemos esto en contraposición con el resultado de los productos industriales, como autos, lámparas LED, televisores, etcétera. Existe en estos objetos artificiales muchísima homogeneidad, más allá de que el uso de todas estas cosas las transforma de alguna manera en únicas (en general luego de un tiempo y más allá de la matrícula o el número de chasis, todos somos capaces de reconocer nuestro propio vehículo, especialmente si en él viajan cuatro niñas pequeñas). Como veremos y analizaremos más adelante, el ***fenotipo***, aquello que podemos observar (usualmente a simple vista en los organismos macroscópicos, luego veremos definiciones más precisas) es el producto de la información contenida en el ***genotipo*** (la información genética, aunque veremos que esto no es muy preciso dicho así) y de su interacción con el **ambiente**, que a los efectos operacionales termina siendo todo lo no-genético.

La importancia de *lo genético* en el fenotipo no deja lugar a dudas: por un lado, los individuos «similares» se aparean entre ellos y no con los «distintos» (tanto sea a nivel de especie, o subespecie, por ejemplo). Por otro lado, en un mismo ambiente coexisten individuos muy distintos (por ejemplo, individuos pertenecientes a especies diferentes), por lo que el ambiente *per se* no estaría determinando una gran parte de los fenotipos. Pese a esta importancia de lo genético, entender los mecanismos que están por detrás de este fenómeno fue un proceso largo y complejo que recién pasada la mitad del siglo XX permitió elaborar una teoría relativamente completa a los distintos niveles involucrados (desde el nivel macro de los organismos, al nivel molecular del ADN, pasando por los niveles de organización intermedia de

[1] Fragmento del artículo de RA Fisher en 1958 titulado «The nature of probability» en la *Centen. Rev.* 2 261-274, refiriéndose al matemático GH Hardy, de quien hablaremos en este capítulo. Una traducción podría ser: «¿Quieres decir, Hardy, que si yo dijera que dos y dos son cinco podrías demostrar cualquier otra proposición que quisieras?». Hardy respondió: «Sí, creo que sí». «Bien, entonces, demuestra que McTaggart es el Papa». «Bien —dijo Hardy—, si dos y dos son cinco, entonces cinco es igual a cuatro. Si restas tres, encontrarás que dos es igual a uno. McTaggart y el Papa son dos; por lo tanto, McTaggart y el Papa son uno».

la información y sus formas de transferencia). Más aún, gran parte de la teoría en genética fue desarrollada sin conocer los mecanismos explícitos subyacentes, lo que llevó a interesantes confrontaciones académicas entre teorías que explicaban mejor o peor una parte de los datos (el debate del inicio del siglo XX entre mendelianos y biométricos, por ejemplo). Esto viene a colación porque el concepto de gen ha ido cambiando con el paso del siglo, así como la interpretación del concepto de alelo, pensada originalmente como una de pocas formas alternativas posibles (estados) de un gen. Desde hace muchos años, la comprensión de las bases moleculares de la herencia (o de algunas, al menos) nos ha permitido entender que la variabilidad subyacente es enorme. La misma comprende la existencia de SNP (que pueden rondar el orden de 1/1000 bases en diferentes especies de mamíferos, por ejemplo), o la existencia de «indels», variantes en el número de copias —CNV—, etcétera. Esto implica que el número de formas alternativas de un gen en la población también puede ser enorme, y solemos recurrir a un concepto más operacional del término alelo: una forma alternativa del gen que tiene consecuencias para nuestro fenotipo.

En un sentido importante la realidad de la genética ha experimentado un cambio trascendental en los últimos decenios: hasta hace relativamente poco (aunque parezca increíble para los lectores más jóvenes) el genotipo real de los individuos era en general desconocido y muchas veces imposible de determinar, excepto a través de experimentos o del análisis de un gran número de descendientes. Sin embargo, con la aparición de las tecnologías de genotipado y de secuenciado masivo (NGS, *Next Generation Sequencing*, por ejemplo) esto ha cambiado definitivamente y hoy en día esto no solo es posible, sino que constituye una alternativa real en mejoramiento genético (por ejemplo, genotipado de reproductores en varios programas de mejoramiento genético). Más allá de la definición que usemos para el concepto de alelo y en qué nivel de variación estemos pensando, es posible analizar la relación esperada entre la frecuencia de los distintos alelos y los genotipos que se forman por su combinación al azar. En este capítulo introduciremos un modelo matemático básico para esto (conocido como equilibrio de Hardy-Weinberg), el cual nos será de utilidad en varios capítulos posteriores. El uso de modelado matemático (y el contraste entre las predicciones de estos modelos y los datos observados en las poblaciones naturales) constituye una de las bases de la genética de poblaciones como disciplina.

OBJETIVOS DEL CAPÍTULO

☐ Introducir el modelo de Hardy-Weinberg. Este es un modelo matemático sencillo que, basado en una serie de premisas claras, nos permite estimar las frecuencias genotípicas esperadas en una población a partir de sus frecuencias alélicas.

☐ Adaptar el modelo básico planteado para modelar y realizar predicciones en distintos escenarios de relevancia biológica, como lo son la existencia de dos sexos distinguibles en la población, el comportamiento de alelos ligados a cromosomas sexuales, la existencia de más de dos alelos en el locus estudiado o la existencia de poliploidías en la especie estudiada.

☐ Discutir aspectos prácticos de la aplicación de este modelo, tales como la repercusión del tamaño muestral al momento de estimar frecuencias alélicas, o la derivación de modelos para la estimación de frecuencias basándose en el conteo de individuos.

2.1 El equilibrio de Hardy-Weinberg

La ley de Hardy-Weinberg, a veces conocida también como de Hardy-Weinberg-Castle por los trabajos de Castle (1903), Hardy (1908) y Weinberg (1908),[2] es posiblemente uno de los resultados mejor conocidos de la genética de poblaciones. La misma permite predecir la composición de genotipos en un locus determinado, una vez alcanzado el equilibrio, a partir de la frecuencia de los alelos cuando se cumplen una serie de condiciones (básicamente, el apareamiento aleatorio entre los individuos de la población). Su base es conceptualmente sencilla: en un individuo diploide, la probabilidad de observar un genotipo en particular en un locus autosómico determinado (es decir, no se encuentra en los cromosomas «sexuales») es el producto de la frecuencia de los gametos que lo formaron. Pongamos un ejemplo sencillo para mostrarlo; supongamos un locus autosómico con dos alelos, A_1 y A_2, con frecuencias p y q. Tenemos $q = 1 - p$, ya que la suma de las frecuencias de los dos alelos debe ser 1 (ya que asumimos que este locus solo tiene 2 alelos). Más aún, supongamos que se trata de una especie diploide hermafrodita, es decir, todos los individuos producen gametos masculinos y femeninos. Para simplificar aún más, asumamos que estos gametos se colocan en dos *pools*: uno de gametos masculinos y otro de gametos femeninos, todos los individuos produciendo exactamente la misma cantidad de gametos. Para simplificar el efecto del tamaño poblacional, asumimos también que dos gametos provenientes del mismo individuo, uno masculino y otro femenino, se pueden combinar. Finalmente, aseguremos que los apareamientos sean completamente al azar, una condición que se conoce como ***panmixia***. Para «crear» los individuos de la nueva generación extraemos un gameto de cada pool. Por lo tanto, la probabilidad de cada genotipo estará dada por la probabilidad de extraer esa combinación, que al tratarse de eventos independientes (el gameto masculino del femenino) es el producto de ambas probabilidades. Es decir

[2] Aunque esta ya había sido claramente establecida por de Vries (1900) durante el redescubrimiento de las leyes de Mendel.

$$\begin{aligned} P(A_1A_1) &= P(A_1) \times P(A_1) = p \times p = p^2 \\ P(A_1A_2) &= 2 \times P(A_1) \times P(A_2) = 2 \times p \times q = 2pq \\ P(A_2A_2) &= P(A_2) \times P(A_2) = q \times q = q^2 \end{aligned} \tag{2.1}$$

que es la expresión más sencilla de la ley de Hardy-Weinberg. Es de notar que el factor 2 que aparece multiplicando la frecuencia de los genotipos heterocigotos (A_1A_2) se debe a que existen dos formas de obtener estos individuos: *a)* sacando un alelo A_1 del pool masculino y un A_2 del pool femenino, y *b)* sacando un alelo A_2 del pool masculino y un A_1 del pool femenino, ambas con la misma probabilidad (pq).

Lo primero que debemos verificar es que estas **frecuencias sumen 1**, ya que no existen otros genotipos posibles (recordemos que $q = 1 - p$):

$$\begin{aligned} p^2 + 2pq + q^2 &= p^2 + 2p(1-p) + (1-p)^2 \\ &= p^2 + 2p - 2p^2 + 1 - 2p + p^2 \\ &= (2p^2 - 2p^2) + (2p - 2p) + 1 \\ &= 0 + 0 + 1 = 1 \end{aligned} \tag{2.2}$$

Una vez verificado esto, el siguiente paso será determinar las nuevas frecuencias de los alelos A_1 y A_2, llamémosles p' y q'. Lo podemos hacer sumando los aportes de alelos de cada genotipo. Los heterocigotos aportan la mitad de su frecuencia para cada tipo de alelo (A_1 y A_2), mientras que los homocigotos aportan toda su frecuencia del respectivo alelo:

$$\begin{aligned} p' &= p^2 + \frac{1}{2} \times 2pq = p^2 + pq = p(p+q) = p(1) = p \\ q' &= q^2 + \frac{1}{2} \times 2pq = q^2 + pq = q(q+p) = q(1) = q \end{aligned} \tag{2.3}$$

Por lo tanto, después de una generación bajo este apareamiento al azar en nuestra especie hermafrodita, las frecuencias alélicas permanecen inalteradas ($p = p'$ y $q = q'$). Obviamente ocurrirá lo mismo para las siguientes generaciones, tanto a nivel de la frecuencia de los genotipos como de los alelos (las frecuencias de los genotipos solo dependen de las frecuencias de los alelos, como vimos en las ecuaciones (2.1) y siguientes), por lo que luego de una generación hemos alcanzado el equilibrio: esta condición se conoce como **equilibrio de Hardy-Weinberg**.

Una forma gráfica de ver estos resultados son los llamados «cuadrados de Punnett»,[3] como el que se aprecia en la figura 2.1. En esta, en el eje horizontal tenemos la frecuencia con la que aparecen los gametos masculinos, mientras que en el vertical aparecen las frecuencias de los gametos femeninos; por lo tanto, el producto de estas frecuencias será el área de los rectángulos y cuadrados correspondientes (de geometría elemental, el área de cuadrados y rectángulos es el producto de dos lados contiguos). Se observa además claramente el origen del factor 2 en los heterocigotos (la suma de las 2 áreas coloreadas en rojo, cada una con valor pq).

[3] Reginald Crundall Punnett, genetista británico que fue quien los desarrolló. Además, Punnett se relaciona directamente con el equilibrio de Hardy-Weinberg, ya que al no poder contrarrestar los argumentos de George Udny Yule acerca del crecimiento sostenido de los alelos dominantes, llevó el problema al matemático Godfrey Harold Hardy, con quien jugaba al cricket; el resto es historia conocida.

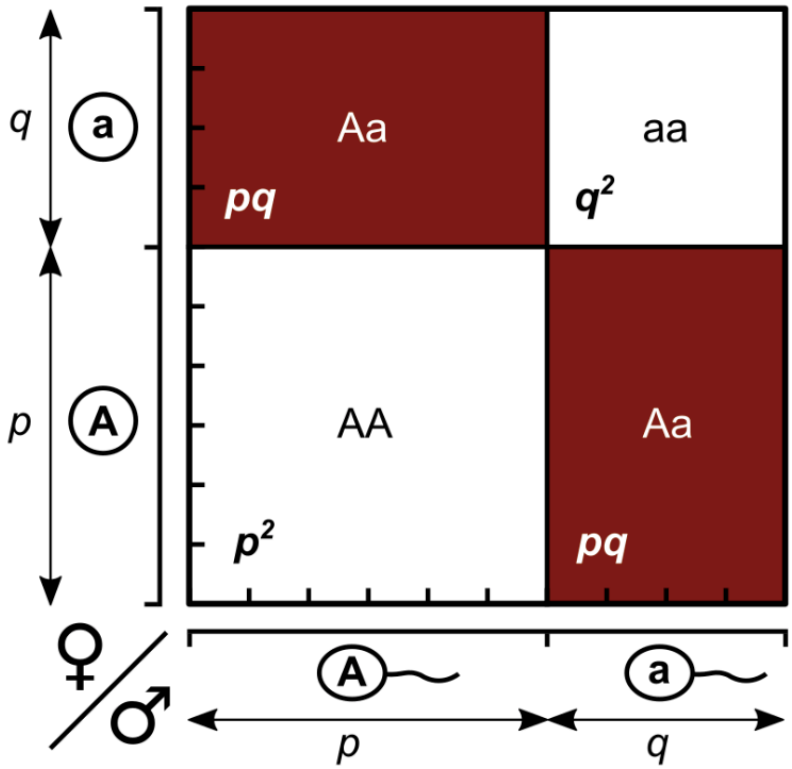


Figura 2.1. Cuadrado de Punnett con frecuencias $p = 0{,}6$ y $q = 0{,}4$. Las áreas de los rectángulos/cuadrados representan las frecuencias de los genotipos. Figura tomada de Wikipedia (CC0), modificada a partir de @pmid26354973

PARA RECORDAR

Hardy-Weinberg en **un locus con dos alelos**, considerando una especie **hermafrodita diploide**

- En el equilibrio, las frecuencias de los genotipos A_1A_1, A_1A_2 y A_2A_2 serán p^2, $2pq$ y q^2, respectivamente.
- Las frecuencias de los alelos (p y q) no cambian luego de una generación de apareamiento y, por lo tanto, como este razonamiento se puede extender a la próxima generación también, no cambiará nunca (mientras no intervengan otras fuerzas evolutivas como la selección o la deriva y el apareamiento sea al azar).
- Las frecuencias de los genotipos quedan completamente determinadas por la frecuencia de los alelos en la generación previa (porque son las frecuencias de los gametos, haploides) y en el caso particular de nuestro modelo de un locus con dos alelos, todo queda determinado por la frecuencia de uno de los dos alelos (p, por ejemplo).
- Además, en el caso particular de nuestro modelo, población hermafrodita, el equilibrio definitivo de las frecuencias de los genotipos se alcanza en una sola generación de apareamientos.

Supuestos que asumimos se cumplen para H-W

1. Todos los genotipos poseen el mismo *fitness*[a] (es decir, ningún genotipo posee una ventaja respecto al número de descendientes que deja).
2. La población debe estar formada por un número infinitamente grande de individuos.
3. El apareamiento al azar debe producirse en toda la población.
4. No debe existir migración que altere las frecuencias.
5. Debe existir un equilibrio mutacional (es decir, las frecuencias de los alelos no varían pese a existir mutaciones).

[a] Esta condición involucra a todos los aspectos que pudieran llevar a diferencias en el éxito reproductivo. Por ejemplo, implica que no existan diferencias ni en la viabilidad de los diferentes genotipos hasta llegar a su etapa reproductiva, ni en su fecundidad.

2.2 Hardy-Weinberg en especies dioicas (dos sexos)

Nuestro modelo previo fue extremadamente sencillo, pero su aplicación se ve reducida a aquellas especies que son hermafroditas. Veamos entonces qué ocurre en especies dioicas (por ejemplo, con dos sexos separados). Supongamos, para mayor generalidad que las frecuencias iniciales en los dos sexos son diferentes, por lo que en la generación inicial tenemos p_M y p_H (y, por lo tanto, $q_M = 1 - p_M$ y $q_H = 1 - p_H$).

Ahora, los diferentes genotipos serán el producto de combinar estos estos alelos con diferentes frecuencias en machos y hembras

$$\begin{aligned} P(A_1A_1) &= P(A_{1M}) \times P(A_{1H}) = p_M \times p_H \\ P(A_1A_2) &= P(A_{1M}) \times P(A_{2H}) + P(A_{2M}) \times P(A_{1H}) = \\ &= p_M \times q_H + q_M \times p_H \\ P(A_2A_2) &= P(A_{2M}) \times P(A_{2H}) = q_M \times q_H \end{aligned} \tag{2.4}$$

Como de estos apareamientos no esperamos diferencias en el número de machos y hembras en cada una de las clases, a partir de ahora tanto machos como hembras tendrán la misma frecuencia de los alelos A_1 y A_2. ¿Pero cuál es esa frecuencia? Para encontrarla, simplemente volveremos a sumar los alelos que aporta cada una de las clases, ponderados por su frecuencia. Es decir, para el alelo A_1, tendremos que la nueva frecuencia p' se puede obtener sumando las contribuciones de $P(A_1A_1)$ y $P(A_1A_2)$

$$p' = P(A_1A_1) + \frac{1}{2}P(A_1A_2) = p_M \times p_H + \frac{1}{2}(p_M \times q_H + q_M \times p_H) \tag{2.5}$$

Multiplicando por 2 ambos lados de la ecuación y separando términos obtenemos

$$\begin{aligned} 2p' &= 2p_M \times p_H + p_M \times q_H + q_M \times p_H \\ 2p' &= p_M \times p_H + p_M \times p_H + p_M \times q_H + q_M \times p_H \end{aligned} \tag{2.6}$$

Reordenando y sacando factor común, tenemos

$$\begin{aligned} 2p' &= (p_M \times p_H + p_M \times q_H) + (p_M \times p_H + q_M \times p_H) \\ 2p' &= p_M(p_H + q_H) + p_H(p_M + q_M) = p_M(1) + p_H(1) \\ 2p' &= p_M + p_H \Leftrightarrow p' = \frac{p_M + p_H}{2} \end{aligned} \tag{2.7}$$

$$p' = \frac{p_M + p_H}{2} \tag{2.8}$$

Es decir, **luego de la primera generación de apareamiento entre machos y hembras con diferentes frecuencias de los alelos A_1 yA_2, la frecuencia de ambos alelos será igual en machos y en hembras y su valor será $p' = \frac{p_M+p_H}{2}$ para el alelo A_1 y $q' = \frac{q_M+q_H}{2}$ para el alelo A_2.**

A este último resultado se puede llegar razonando por simetría (¿hubiera cambiado llamar al revés a los alelos?) o haciendo las cuentas

$$p' = \frac{p_M + p_H}{2} = \frac{(1-q_M)+(1-q_H)}{2} = \\ = \frac{2-q_M-q_H}{2} = 1 - \frac{(q_M+q_H)}{2} \therefore \\ \frac{(q_M+q_H)}{2} = 1 - p' = q' \tag{2.9}$$

El último resultado está justificado, además, porque la suma de los alelos siempre debe dar una frecuencia relativa de 1. Más allá de todas las cuentas que nos han llevado a estos resultados en forma analítica, una alternativa primera sería utilizar el argumento de que el aporte de material genético para formar la siguiente generación debe ser igual en machos que en hembras: es decir, cada sexo aporta la mitad del material genético (autosómico) de la nueva generación. Por lo tanto, la frecuencia de los alelos en la generación de hijos debe ser el promedio de las frecuencias en sus padres, que es el resultado que hemos visto.

Ejemplo 2.1

El genotipado molecular (chips) nos permite conocer las bases (A, C, G, T) particulares que presenta un animal en diferentes posiciones del genoma, desde unas pocas posiciones a centenas de miles (por ejemplo, alrededor de 700 000 en algunos chips). Estos marcadores se conocen como SNP (*Single Nucleotide Polymorphisms*, en inglés) y si bien son relativamente poco informativos como marcadores (se suelen escoger marcadores bialélicos en el diseño), su elevado número los transforma en una herramienta sumamente poderosa. Usando un chip comercial para bovinos en 1000 animales no emparentados de la raza guernsey se encontró que el SNP **rs43375517**, en la región 5'UTR del gen ARL4A y asociado a la producción de leche a 305 días en varios estudios, presentaba las siguientes frecuencias de genotipos: CC 372, CG 476, GG 152. Calcular la frecuencia relativa de ambos alelos, así como la esperanza del número observado de cada genotipo si la población se encontrase en equilibrio de Hardy-Weinberg para este marcador.

Dado que tenemos 1000 individuos genotipados (372+476+152=1000), el número total de alelos en genotipados es 2000 (son individuos diploides). La frecuencia relativa del alelo 'C', llamémosle p, es

$$p = \frac{2 \times 372 + 1 \times 476}{2000} = \frac{744+476}{2000} = \frac{1220}{2000} = 0{,}61 \tag{2.10}$$

por lo que, la frecuencia relativa del alelo 'G' es $q = 1 - p = 0{,}39$. Si la población se encontrase en equilibrio de Hardy-Weinberg, con 1000 individuos genotipados esperaríamos obtener las siguientes frecuencias

$$CC = p^2 \times 1000 = 0{,}61^2 \times 1000 = 0{,}3721 \times 1000 = 372{,}1 \\ CG = 2pq \times 1000 = 2 \times 0{,}61 \times 0{,}39 = 0{,}4758 \times 1000 = 475{,}8 \\ GG = q^2 \times 1000 = 0{,}39^2 \times 1000 = 0{,}1521 \times 1000 = 152{,}1 \tag{2.11}$$

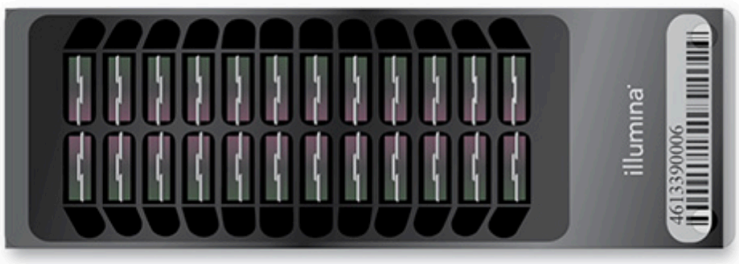


Figura 2.2. Chip BovineSNP50 v3 de genotipado Bovino, de la empresa Illumina, con capacidad para analizar simultáneamente en 24 muestras 53218 SNP altamente informativos (fotografía e información de la página de Illumina).

por lo que más allá de las diferencias atribuibles al redondeo (es imposible genotipar fracciones de individuo, pero incrementando el tamaño de la muestra estos errores pueden reducirse más), tenemos un ajuste casi perfecto a las expectativas del equilibrio Hardy-Weinberg.

2.3 H-W: la frecuencia de heterocigotas en función de la frecuencia alélica

En un locus con dos alelos, ¿a qué frecuencia del alelo p podemos observar la mayor frecuencia de heterocigotas, asumiendo que la población se encuentra en equilibrio de Hardy-Weinberg? La alternativa más sencilla para tener una idea aproximada del comportamiento del número de heterocigotas bajo estos supuestos es graficar la misma como función de p. Tenemos que $2pq = 2p(1-p)$, y podemos graficar esta relación en cualquier software que permita graficar funciones, calculadora, o aún en una planilla electrónica (por ejemplo, generando una serie en el intervalo 0-1). La figura 2.3 es el resultado de graficar esta función, donde se ve que la máxima frecuencia de heterocigotas se encuentra en el entorno de $p = 0{,}5$, o sea en frecuencias intermedias de los dos alelos; la máxima proporción de heterocigotas esperable para una población en equilibrio de HW es 0,5.

Una forma más precisa, complementaria con la anterior, es el estudio analítico de esta función. Es decir, dada la función $f_{Het}(p) = 2pq = 2p(1-p) = 2p - 2p^2$, podemos reconocer sus extremos relativos (máximos y mínimos) como los lugares donde la pendiente de la recta tangente es igual a cero. Lo que es equivalente, estos son los puntos donde la derivada primera de la función se hace cero.

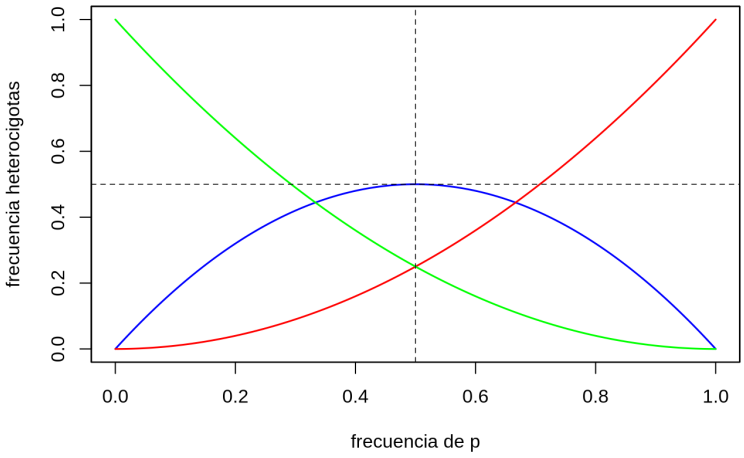


Figura 2.3. Frecuencia de heterocigotas en función de la frecuencia del alelo p en un modelo de un locus con dos alelos, bajo equilibrio Hardy-Weinberg. En azul la frecuencia de los heterocigotas, rojo y verde los homocigotas.

Por lo tanto, derivando la función obtenemos

$$\begin{aligned} f_{Het}(p) = 2pq = 2p(1-p) = 2p - 2p^2 \\ \frac{df_{Het}}{dp} = 2 - 4p \\ 2 - 4p = 0 \Leftrightarrow p = \frac{1}{2} \end{aligned} \tag{2.12}$$

Es decir, la función tiene un solo máximo o mínimo en $p = \frac{1}{2} = 0{,}5$. Para reconocer de cuál de los dos se trata es necesario ver el signo de la derivada segunda (lo que hacemos volviendo a derivar la derivada primera)

$$\frac{d^2 f_{Het}(p)}{dp^2} = -4 \tag{2.13}$$

Como el signo es negativo, se trata de un máximo. Más aún, la forma de función original ($f_{Het} = 2p - 2p^2$) es claramente reconocible como la de una parábola con forma de «U» invertida y cuyas intersecciones con el eje de las abscisas (raíces) son 0 y 1, por lo que por simetría el vértice debe estar en el punto medio (es decir, $\frac{1}{2}$). Sustituyendo, tenemos

$$\begin{aligned} f_{Het_{max}}(p = \frac{1}{2}) = 2p - 2p^2 = 2\frac{1}{2} - 2\left(\frac{1}{2}\right)^2 = \\ 2\frac{1}{2} - 2\frac{1}{4} = 1 - \frac{1}{2} = \frac{1}{2} \end{aligned} \tag{2.14}$$

2.4 El equilibrio de Hardy-Weinberg en cromosomas ligados al sexo

En los razonamientos previos hemos asumido que el locus de interés era autosómico, es decir que no estaba en uno de los cromosomas sexuales. ¿Qué ocurriría si este locus se encuentra en el cromosoma X del sistema **XY** de determinación del sexo?[4]

A diferencia de antes, los machos poseen ahora solo un cromosoma X (son hemicigotas), que es donde está nuestro locus de interés y que por lo tanto tienen el alelo A_1 con probabilidad p_M y el alelo A_2 con probabilidad $q_M = 1 - p_M$. En las hembras tenemos los dos cromosomas X y en ambos los alelos A_1 y A_2 tendrán probabilidad p_H y q_H respectivamente (con $q_H = 1 - p_H$). Las hembras descendientes de este cruzamiento tendrán dos cromosomas X, uno del padre y otro de la madre, y las frecuencias de los genotipos serán equivalentes a los de las ecuaciones (2.4) y siguientes. En cambio, los machos tienen un solo cromosoma X, aportado por la madre (el padre aporta el cromosoma Y) y, por lo tanto, las frecuencias de los alelos A_1 y A_2 en esta nueva generación de machos serán iguales a las frecuencias del respectivo alelo en las hembras de la generación anterior.

En resumen, en la nueva generación de machos y hembras (la distinguimos de la anterior con ') tendremos ahora las siguientes frecuencias

$$\begin{aligned} p'_H &= \frac{p_M + p_H}{2} \\ q'_H &= \frac{q_M + q_H}{2} \\ p'_M &= p_H \\ q'_M &= q_H \end{aligned} \tag{2.15}$$

El mismo razonamiento lo podemos extender para la próxima generación (la distinguimos de la anterior con '), por lo que ahora tendremos

$$\begin{aligned} p''_H &= \frac{p'_M + p'_H}{2} = \frac{p_H + \frac{p_M + p_H}{2}}{2} \\ q''_H &= \frac{q'_M + q'_H}{2} = \frac{q_H + \frac{q_M + q_H}{2}}{2} \\ p''_M &= p'_H = \frac{p_M + p_H}{2} \\ q''_M &= q'_H = \frac{q_M + q_H}{2} \end{aligned} \tag{2.16}$$

O sea, hemos establecido una regla de recurrencia para calcular las frecuencias de p y q a partir de la generación anterior (se puede llegar hasta la generación inicial), pero aún no hemos estudiado el comportamiento a largo plazo de este proceso. Para entenderlo mejor podemos resumir la información, como aparece en el siguiente cuadro, donde además de

[4] En el sistema **XY**, presente en los mamíferos, el gen Y es el que determina el sexo (masculino), siendo los machos heterogaméticos (XY) y las hembras homogaméticas (XX). En otras especies, como las aves, existen otros sistemas como el **ZW**, en el que son las hembras las que son heterogaméticas (ZW), mientras que los machos son homogaméticos (ZZ).

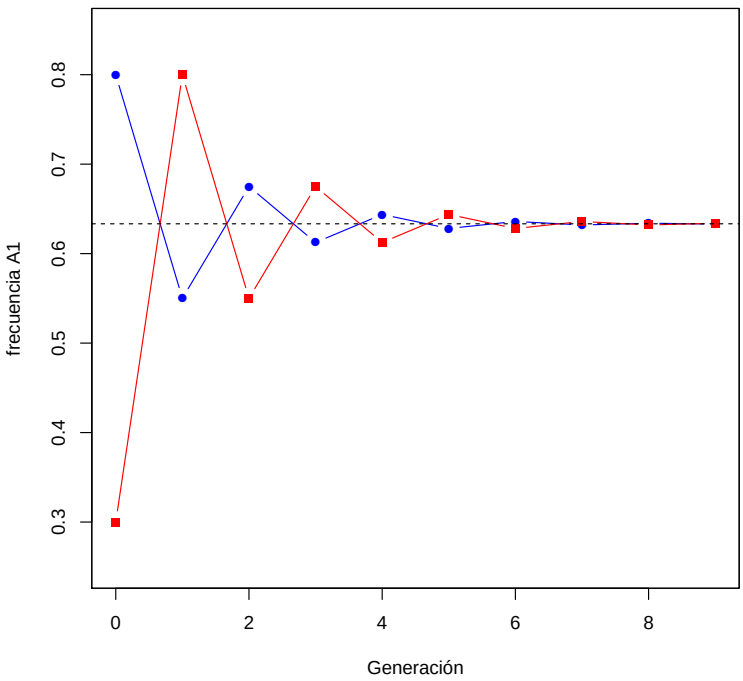


Figura 2.4. Evolución de las frecuencias alélicas en machos (cuadrados rojos) y hembras (círculos azules) en un locus perteneciente al cromosoma X en especies en que el Y determina el sexo. La línea a trazos negros representa la media de la población. Las frecuencias de partida (generación 0) son $p_M = 0{,}3$ y $p_H = 0{,}8$.

las frecuencias del alelo A_1 para hembras y machos en las 3 primeras generaciones también calculamos la frecuencia del alelo en toda la población (media) y la diferencia de frecuencias entre hembras y machos.

Generación	A_{1H}	A_{1M}	Media	Diferencia
0	p_H	p_M	$\frac{2p_H+p_M}{3}$	$p_H - p_M$
1	$\frac{p_M+p_H}{2}$	p_H	$\frac{2p_H+p_M}{3}$	$\frac{p_M-p_H}{2} = -\frac{(p_H-p_M)}{2}$
2	$\frac{p_H+\frac{p_M+p_H}{2}}{2}$	$\frac{p_M+p_H}{2}$	$\frac{2p_H+p_M}{3}$	$\frac{p_H-p_M}{4}$

Es de notar que, para calcular la frecuencia del alelo en la población, debemos ponderar la frecuencia en hembras por 2/3 y la de machos por 1/3, ya que esas son las proporciones de cromosomas X que aportan cada uno de los sexos a la población.

Si nos fijamos en la columna de la media en el cuadro anterior, lo primero que observamos es que la misma no cambia con el tiempo. Es decir, la frecuencia general en la población de los alelos A_1 y A_2 está determinada desde el principio, y su valor es

$$p = \frac{2p_H + p_M}{3} \tag{2.17}$$

Lo segundo que podemos apreciar, si observamos la columna que muestra la diferencia entre los dos sexos, es que el valor absoluto de esta diferencia se va reduciendo a la mitad en cada generación, mientras que el signo de la misma alterna de generación en generación. Con estas observaciones en mano es fácil entender que el proceso tiende igualar las frecuencias entre los dos sexos (asumiendo que parten de valores diferentes en machos y hembras). Si bien teóricamente estas frecuencias nunca llegan a ser iguales, desde el punto de vista práctico la diferencia es despreciable luego de unas pocas generaciones, y en general los procesos estocásticos (por ejemplo, la deriva, que veremos más adelante) tendrán un impacto mucho mayor que estas diferencias.

El comportamiento de estas frecuencias puede observarse gráficamente también, como lo muestra, por ejemplo, la figura 2.4, partiendo en este caso de $p_M = 0{,}3$ y $p_H = 0{,}8$. Se aprecia claramente en esta figura que *a)* la frecuencia en machos es igual a la frecuencia en las hembras de la generación anterior, *b)* la frecuencia en hembras es el promedio de las frecuencias de machos y hembras en la generación anterior, *c)* la media de la población permanece estable (línea a trazos negros) y *d)* la diferencia entre machos y hembras se reduce a la mitad en cada generación.

PARA RECORDAR

Para el equilibrio de Hardy-Weinberg en cromosomas ligados al sexo, asumiendo que el macho es el sexo heterogamético (al revés si la hembra lo es), tenemos:

- La frecuencia del alelo $\boldsymbol{A}_1$ en **machos** (p_M) de la generación t es igual a p_H en la generación $t-1$.
- La frecuencia del alelo $\boldsymbol{A}_1$ en **hembras** (p_H) de la generación t es igual a $\frac{p_H+p_M}{2}$ en la generación $t-1$.

- Este proceso converge relativamente rápido al valor $p = \frac{(2p_H + p_M)}{3}$, que es la frecuencia inicial de alelos p en el **pool de gametos.**

2.5 Tres o más alelos

Hasta ahora hemos manejado los diferentes problemas asumiendo que se trataba de un locus con únicamente dos alelos, lo que en principio permite simplificar algunas cuentas ya que la frecuencia de q queda determinada por la frecuencia de p (por ejemplo, $q = 1 - p$). Sin embargo, la extensión a tres o más alelos es relativamente sencilla y no requiere de cosas nuevas, por más que aprovecharemos la oportunidad para presentar el resultado central del equilibrio de Hardy-Weinberg de una forma diferente.

Si recordamos nuestro modelo de dos alelos, podíamos representar el conjunto de los alelos en la población como la suma de p y q, es decir $(p + q)$. Cuando hacemos el apareamiento en condiciones de panmixia, estamos cruzando este $(p + q)$ con otro similar (porque los alelos salen de la misma población), o lo que es equivalente

$$(p + q)(p + q) = (p + q)^2 = p^2 + 2pq + q^2 \tag{2.18}$$

De lo anterior surge claro que podemos representar las frecuencias genotípicas esperadas en el equilibrio Hardy-Weinberg como el cuadrado del binomio formado por las frecuencias alélicas en la generación previa (frecuencias gaméticas). Con esto en mano, resulta tentador extender este resultado a más de 2 alelos. Supongamos entonces que tenemos un locus con 3 alelos, A_1, A_2 y A_3, con frecuencias respectivas p, q y r. Nuestro «todo» estará ahora conformado por la suma de p, q y r, es decir $(p + q + r)$, y el cruzamiento resultará en

$$(p+q+r)(p+q+r) = (p+q+r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr \tag{2.19}$$

Es decir, si llamamos A_i a los $i = 1,2,...,k$ alelos diferentes, y p_i a las frecuencias correspondientes, la frecuencia de los genotipos homocigotas estará determinada por $p_i p_i = p_i^2$ y la de los heterocigotas por $2p_i p_j, i \neq j$. Obviamente, el razonamiento del binomio y del trinomio se puede extender a un número mayor de alelos.

En general, para k alelos diferentes, el número de genotipos heterocigotas es igual a $k(k-1)/2$. Para ver de dónde se desprende esto, es fácil imaginar un cuadrado de Punnett, ahora con 3 o más alelos. El total de celdas del cuadrado es $k \times k = k^2$, de las cuales, los homocigotas serán las celdas que se encuentren en la diagonal, o sea k celdas. Esto nos deja $k^2 - k = k(k-1)$. Sin embargo, las celdas simétricas respecto a la diagonal representan el mismo genotipo (por ejemplo, $p_{ij} \equiv p_{ji}, \forall i \neq j$; de ahí el factor 2 para los heterocigotas), por lo que al total de celdas que nos quedaban debemos dividirlo entre 2 para tener el número de genotipos heterocigotos y de ahí que este sea igual a $k(k-1)/2$. Finalmente, el número total de genotipos es la suma de homocigotos (k) y heterocigotos ($k(k-1)/2$).

Es decir, el número de genotipos corresponde a
$k + k(k-1)/2 = (2k/2) + k(k-1)/2 = k(k+1)/2$
De acuerdo con lo anterior, la frecuencia total de homocigotos será

$$G = \sum_{i=1}^{k} p_i^2 \tag{2.20}$$

lo que nos deja la posibilidad de definir la heterocigosis del locus como lo que falta para completar la totalidad, o sea

$$H = 1 - G = 1 - \sum_{i=1}^{k} p_i^2 \tag{2.21}$$

Es importante entender que la fórmula anterior no es función de la proporción real de heterocigotos que existen en una población determinada para un locus dado, sino que se trata de la esperanza en función de las frecuencias alélicas, si se cumpliera el equilibrio de Hardy-Weinberg.

PARA RECORDAR

- Para 3 alelos el equilibrio de Hardy-Weinberg es la expansión del cuadrado del trinomio, es decir para los alelos $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{A}_3$, con frecuencias respectivas p, q y r, $p + q + r = 1$,

$$(p + q + r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr \tag{2.22}$$

los tres primeros términos de la derecha corresponden a los homocigotos $\boldsymbol{A}_1\boldsymbol{A}_1$, $\boldsymbol{A}_2\boldsymbol{A}_2$, $\boldsymbol{A}_3\boldsymbol{A}_3$ y el resto a los heterocigotos $\boldsymbol{A}_1\boldsymbol{A}_2$, $\boldsymbol{A}_1\boldsymbol{A}_3$, $\boldsymbol{A}_2\boldsymbol{A}_3$.

- La ecuación anterior es generalizable a un número k de alelos, $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_k$ de acuerdo a

$$(p_1 + p_2 + p_3 ... + p_k)^2 \tag{2.23}$$

- Bajo equilibrio de Hardy-Weinberg, la proporción esperada de **homocigotas** será

$$G = \sum_{i=1}^{k} p_i^2, \tag{2.24}$$

por lo que la heterocigosis del locus quedará como lo que falta para completar la totalidad, o sea

$$H = 1 - G = 1 - \sum_{i=1}^{k} p_i^2 \tag{2.25}$$

2.6 La estimación de frecuencias y el equilibrio (o no)

En una población con un gran número de individuos, tendiendo a infinito, si la población se encuentra en equilibrio de Hardy-Weinberg el número de individuos de los distintos genotipos coincidirá estrechamente con las esperanzas de los mismos a partir de las frecuencias de los alelos. Es decir, la variación estocástica en cada uno de los apareamientos (o en la unión de los gametos) tiende a cancelarse cuando consideramos el gran número de individuos en la población. Sin embargo, aún en esas grandes poblaciones no solemos tener acceso a la información de todos los individuos y sí, solamente, a la información de una **muestra aleatoria** de la misma. El concepto de muestra aleatoria es fundamental ya que será el que nos permitirá extender nuestras conclusiones de la muestra hacia la población, que en última instancia suele ser lo que nos interesa. Al trabajar con una muestra aleatoria relativamente reducida de la población, los números ahora serán mucho menores y, por lo tanto, los efectos del azar se verán incrementados.

Supongamos, para ver el impacto del tamaño de la muestra, que tenemos interés en estimar las frecuencias de los dos alelos en un locus autosomal de una especie diploide en una población determinada. Como se trata de una especie salvaje, difícil de capturar para muestrear, nos contentamos con los 5 individuos que pudimos genotipar para este locus; sin duda es mucho mejor que no tener nada. Como se trata de una especie diploide, si tenemos n individuos vamos a tener $2n$ alelos genotipados (por ejemplo, 10 alelos genotipados). Supongamos ahora que de esos 10 alelos genotipados, 3 fueron del alelo A_1 y 7 del alelo A_2. Es decir, la proporción de A_1 es igual a $p = 3/10 = 0{,}3$, mientras que la del alelo A_2 es de $q = 7/10 = 0{,}7$ (hubiésemos llegado también a este resultado como $q = 1 - p$). Parece razonable usar este valor de p como un **estimador** del «verdadero valor»[5] del parámetro en la población, en este caso, la frecuencia del alelo A_1, tanto que coincide con el estimador por máxima verosimilitud (es decir, escogemos como valor estimado aquel que tiene mayor probabilidad de ocurrir de acuerdo con lo que hemos observado).

Ahora, ¿qué certeza tenemos que este valor que estimamos esté cerca del «verdadero valor»? En principio, como solo tenemos 10 alelos genotipados, nuestras estimaciones irán de 10 % en 10 % (los saltos en la cuenta de alelos son de a 1, pero en 10 total, así que tenemos $\frac{1}{10} = 0{,}1 = 10\,\%$). O sea, la precisión de nuestra estimación es realmente baja.

Otra forma de ver este problema es pensar en cuál sería la probabilidad de observar este conteo de alelos A_1, por ejemplo, dado que conocemos el verdadero valor del parámetro. Si suponemos que los individuos representan un muestreo al azar de los gametos en la población y que los individuos de la muestra fueron seleccionados al azar, entonces la probabilidad de observar un conteo x en particular, dado el número total de alelos genotipados n, se puede describir por la distribución binomial (en nuestro caso de dos alelos)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2.26}$$

[5] Ya veremos en otras partes que el concepto de «verdadero valor» de un parámetro está ligado a una corriente estadística llamada frecuentista y que se contrapone en cierta manera a una visión más subjetiva de la probabilidad, representada en la escuela bayesiana.

con las combinaciones de n elementos tomados de a x elementos iguales a

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{2.27}$$

Por ejemplo, si la frecuencia «real» en la población entera para el alelo A_1 es de $p = 0{,}5$, ¿cuál sería la probabilidad de observar este conteo en 3 alelos de los 10 que genotipamos? De acuerdo con la ecuación (2.26), la probabilidad de obtener 3 conteos en una muestra de 10, con una probabilidad $p = 0{,}5$ en la población, es de

$$f(3) = \binom{10}{3}(0{,}5)^3(1-0{,}5)^{10-3} = \frac{10!}{7!3!}(0{,}5)^3(0{,}5)^7 = \frac{10\times 9\times 8}{6}(0{,}5)^{10} = 0{,}1171875 \sim 11{,}7\,\%. \tag{2.28}$$

Es decir, aun cuando la frecuencia real en la población es de $p = 0{,}5$, en más de 1 de cada 10 muestras que tomemos, solo por azar, observaremos un conteo de 3 alelos. En estos casos, estimaríamos erróneamente $p \approx 0{,}33$. Esto puede parecer poco, pero si nos fijamos en la figura 2.5(a), donde graficamos la probabilidad de observar cada uno de los 11 conteos posibles (de 0 a 10 alelos A_1), resulta claro que la probabilidad de no acertar en el único número correcto (5 alelos) es realmente alta.

De hecho, si calculamos la probabilidad de observar exactamente 5 alelos en 10, dada la frecuencia real de $p = 0{,}5$, esta probabilidad es igual a 0,2460938, por lo que la probabilidad de **no** obtener el «verdadero» valor de p en una muestra de este tamaño es igual a $1 - 0{,}2460938 = 0{,}7539062 \sim 75{,}4\,\%$. Por otro lado, la probabilidad de tener un estimado de $p \leq 0{,}3$ es la suma de las probabilidades de obtener 0, 1, 2 o 3 conteos, y esto es igual a $0{,}171875 \sim 17{,}2\,\%$ (el lector lo puede verificar fácilmente).

Claramente, de lo anterior surge que este tamaño muestral de 5 individuos no nos conducirá a nada muy razonable. ¿Qué ocurriría si conseguimos aumentar el tamaño muestral a 50 individuos? Las probabilidades esperadas de los distintos conteos en la muestra aleatoria se grafican en la figura 2.5(b). Ahora resulta claro que, por un lado, la «granularidad» de las barras es mucho más fina (por ejemplo, la «precisión» de nuestros estimados podría ser mucho mayor) y por el otro, que las estimaciones muy alejadas del «verdadero valor» son bastante improbables. Por ejemplo, ¿cuál sería la probabilidad de tener un estimado de $p \leq 0{,}3$? En este caso se trata de la suma de las probabilidades de todos los conteos, de 0 a 30 ($30/100 = 0{,}3$). En decir, se trata de la **función de distribución** (a veces conocida como probabilidad acumulada, o también CDF en inglés). En este caso, para la distribución binomial, tenemos para un valor k (30 en nuestro caso)

$$Pr(X \leq k) = \sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x} \tag{2.29}$$

Sería algo tedioso calcularlo a mano, pero una primera idea del valor se puede tener observando la figura 2.5(b), donde resulta claro que las probabilidades para conteos de hasta

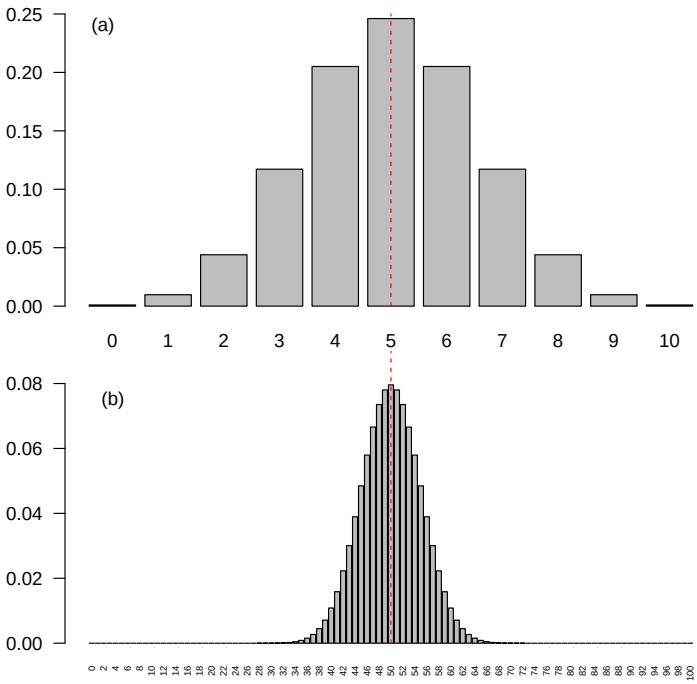


Figura 2.5. Probabilidad de observar un conteo de alelos determinado en una muestra de 5 individuos (a) y en una muestra de 50 individuos (b), dado que la frecuencia del mismo en la población es de 0,5.

30 son despreciables. De hecho, utilizando la función de distribución en cualquier software que la tenga implementada, la probabilidad de observar un conteo de 30 o menos alelos en 100 genotipados es de $3{,}92507 \times 10^{-5}$ ($\sim 0{,}004\,\%$) si la frecuencia del alelo A_1 en la población es $p = 0{,}5$; o sea, prácticamente improbable.

Ahora también bajó la probabilidad de acertar **exactamente** en la estimación de $p = 0{,}5$. ¿Pero qué sucede con la probabilidad de obtener observaciones en un intervalo que nos lleve a una estimación razonablemente cercana a $p = 0{,}5$? Comparemos con nuestra muestra de 5 individuos, donde la probabilidad de observar exactamente 5 alelos era de $\sim 24{,}6\,\%$. El intervalo de frecuencias equivalente en nuestra muestra de 50 individuos corresponde al que contempla observar entre 45 y 54 alelos (probabilidades entre 0,45 y 0,54);[6] este intervalo concentra ahora $\sim 63\,\%$ de la probabilidad. Dicho de otra manera, ahora los resultados se concentrarán (aproximadamente) entre 49 y 51 con una probabilidad similar a la que observábamos en nuestra muestra de 5 individuos. Claramente, nuestras estimaciones se vuelven mucho más precisas y útiles.

Hasta el momento hicimos nuestras estimaciones de las frecuencias alélicas poblacionales (y, por lo tanto, la posibilidad de conocer nuestras expectativas para el equilibrio de Hardy-Weinberg) de una manera bastante sencilla: asumimos que —genómica mediante— podemos acceder al verdadero estado alélico (por ejemplo, ADN) de cada una de las muestras. Sin embargo, en muchos casos esto no es posible. Por ejemplo, puede desconocerse la base molecular de las diferencias fenotípicas observadas. En otros casos, obtener esta información no es práctico, y la única información disponible es la fenotípica. Por lo tanto, en estos casos solo contamos con el conteo de los distintos fenotipos. Esto introduce un nuevo factor en el problema: es necesario conocer el modelo de herencia (mendeliana, codominancia, etcétera) para poder pasar de los fenotipos a los genotipos, y de allí a las frecuencias de los alelos.

Supongamos, para comenzar, que **el modelo de herencia es de codominancia** en un locus autosomal con dos alelos (A_1 y A_2) en una especie diploide. Bajo este modelo, ambos alelos se expresan en el fenotipo y hay una correspondencia biunívoca entre fenotipo y genotipo. En este caso tenemos tres genotipos que son identificables directamente, por lo que podemos estimar las frecuencias relativas a partir del conteo de individuos. Para ello, tendremos en cuenta: *a)* que los homocigotas presentan 2 alelos A_1 y los heterocigotas 1 y *b)* que para n individuos tenemos $2n$ alelos. Así, en una muestra de n individuos la frecuencia relativa (muestral) del alelo A_1 será igual a la suma de 2 veces el número de homocigotas A_1 más el número de heterocigotas, todo dividido entre 2 veces el número de individuos

$$p = \frac{2n_{A_1A_1} + n_{A_1A_2}}{2n}, q = \frac{2n_{A_2A_2} + n_{A_1A_2}}{2n} \qquad n = n_{A_1A_1} + n_{A_1A_2} + n_{A_2A_2} \tag{2.30}$$

En todo caso, si conocemos los genotipos a nivel molecular (ver ejemplo 2.1), la forma de estimar es análoga al caso donde los genotipos corresponden en forma biunívoca con los fenotipos.[7] Tanto si se trata de un modelo de codominancia como de dominancia parcial, lo relevante aquí es la posibilidad de distinguir los heterocigotas de ambos homocigotas. Esto

[6] Si observamos 45/100 alelos en nuestro muestreo, nuestra estimación será $p = 0{,}45$ ($p = 0{,}5$, si redondeamos a una cifra significativa). Lo mismo sucedería si observamos cualquier valor entre 45 y 54 alelos.

[7] En estos casos, observar el fenotipo aporta la misma información que determinar el genotipo.

nos permite simplemente hacer el «conteo» de alelos, ya que la base genética es clara: en diploides, los homocigotas poseen dos alelos del mismo tipo y los heterocigotas poseen uno de cada tipo.

Las cosas empiezan a complicarse un poco más cuando **aparece el efecto de dominancia**. Supongamos, por ejemplo, que en el locus anterior (autosomal en especie diploide, con dos alelos), el alelo A_1 domina completamente al A_2. En este caso, tanto el genotipo A_1A_1 como el A_1A_2 tendrán idéntico fenotipo, mientras que el A_2A_2 tendrá un fenotipo diferente. Esto implica que el conteo de fenotipos ya no permite determinar las frecuencias de los ambos alelos, ya que los conteos del fenotipo A_1 corresponde al de dos genotipos en simultáneo ($G_{11} = A_1A_1$ y $G_{12} = A_1A_2$), ambos con diferente número de alelos A_1 y A_2 (2 y 0 para el homocigota, 1 y 1 para el heterocigota). Dicho de otra forma, sabemos calcular las frecuencias alélicas a partir de los 3 genotipos, pero solo contamos con dos fenotipos: el del alelo dominante y el del alelo recesivo. De hecho, con solo esta información se trata de un problema indeterminado. Para poder salir de esta situación debemos agregar algún tipo de información adicional. Una alternativa sencilla (suponiendo que no podemos conocer los genotipos, lo que resolvería el problema) es **asumir que la población se encuentra en equilibrio de Hardy-Weinberg**. Esta asunción establece una relación explícita entre frecuencias alélicas (lo desconocido) y genotipos. Notemos que la clase genotípica A_2A_2 (homocigota) es la única que corresponde con un único fenotipo. Si la población se encuentra en equilibrio de Hardy-Weinberg, sabemos que la esperanza del número de individuos para este genotipo es igual a nq^2, donde n es el número total de individuos ($n = G_{1.} + G_{22}$). Sabemos además que el número de individuos con el fenotipo dominante está dado por $G_{1.} = G_{11} + G_{12}$. Esto nos permite estimar q (es decir, la frecuencia del alelo A_2),

$$G_{22} = nq^2 \therefore q = \pm\sqrt{\frac{G_{22}}{n}} = \pm\sqrt{\frac{G_{22}}{G_{1.} + G_{22}}} \tag{2.31}$$

De las dos raíces que tiene esta ecuación (una positiva y otra negativa), solo una posee sentido biológico: no existen frecuencias alélicas negativas y, por lo tanto, dicha raíz no es una solución válida.

Es importante recalcar en este punto que la validez de nuestra estimación depende de que la población se encuentre en equilibrio de Hardy-Weinberg, pues de lo contrario no será una estimación que tenga una base teórica.

Cálculos a partir de otras notaciones para genotipos

Algunas veces es posible encontrar otras notaciones para los distintos genotipos que pueden resultar más claras en el caso de un modelo de un gen con dos alelos, particularmente en el caso de codominancia.
Si llamamos P, H y Q respectivamente a las frecuencias relativas de los homocigotos para el alelo cuya frecuencia es p (por ejemplo, los A_1A_1), a los heterocigotos (A_1A_2) y a los homocigotos para el alelo cuya frecuencia es $q = (1 - p)$ (A_2A_2), entonces $P + H + Q = 1$.

Estas son las frecuencias genotípicas relativas **observadas.** De hecho, si el **conteo de individuos** de cada uno de los genotipos es respectivamente N_P, N_H y N_Q, el número total de individuos será $N = N_P + N_H + N_Q$, entonces

$$P = N_P/N, H = N_H/N, Q = N_Q/N \tag{2.32}$$

A partir de esto, en el caso de codominancia, tenemos que mientras que todos los alelos de los genotipos A_1A_1 serán del tipo A_1, solo la mitad de los correspondientes a los heterocigotos (A_1A_2) lo serán A_1, por lo que la frecuencia del alelo A_1 en la muestra será igual a

$$p = P + \frac{1}{2}H \tag{2.33}$$

mientras que, con la misma lógica, la frecuencia del alelo A_2 estará definida por

$$q = Q + \frac{1}{2}H \tag{2.34}$$

Obviamente, $p + q = (P + \frac{1}{2}H) + (Q + \frac{1}{2}H) = P + Q + H = 1$ ya que

$$P + Q + H = \frac{N_P}{N} + \frac{N_Q}{N} + \frac{N_H}{N} = \frac{N_P + N_Q + N_H}{N} = \frac{N}{N} = 1 \tag{2.35}$$

Notemos que en caso de codominancia **no necesitamos invocar el equilibrio de Hardy-Weinberg** si tenemos las frecuencias fenotípicas.

En el caso de dominancia completa, solo vamos a tener dos categorías observables, los genotipos que contienen el alelo dominante (por ejemplo, A_1A_1 y A_1A_2) y el que no los contiene (A_2A_2). Como no podemos discriminar fenotípicamente los A_1A_1 de los A_1A_2, no podemos separar P de H y la única opción es agregar algún supuesto, por ejemplo, invocar el equilibrio de Hardy-Weinberg. En este caso, de acuerdo con lo visto antes, la **frecuencia relativa esperada** de los homocigotos recesivos será q^2 e igualando esto a la **frecuencia relativa observada** Q, tenemos

$$q^2 = Q \therefore q = \sqrt{Q} \tag{2.36}$$

Ejemplo 2.2

La fenilcetonuria (PKU) es un error innato del metabolismo que resulta en una disminución del metabolismo del aminoácido fenilalanina. La enfermedad es resultado de una deficiencia en la enzima fenilalanina hidroxilasa, codificada por el gen PAH. El resultado de esto es la acumulación de fenilalanina a niveles potencialmente tóxicos. Desde el punto de vista del modelo de herencia se comporta como autosómica recesiva, por lo que es necesario heredar dos copias «defectuosas» del gen PAH (a decir verdad, que tengan actividad seriamente disminuida) para que la enfermedad se manifieste. La incidencia es muy variable de población en población, pero dos casos extremos podrían ser el de la población de Finlandia (1 afectado en 200 000 nacimientos) y el de la población de Turquía (1 afectado en 2600

nacimientos). Calcular para ambas poblaciones la frecuencia «del»[8] alelo defectuoso, así como la proporción de portadores (que no manifiestan la enfermedad), asumiendo que las poblaciones se encuentran en equilibrio de Hardy-Weinberg.
Utilizando la ecuación (2.31), tenemos para los dos casos

$$\begin{aligned} q_{Fin} &= \sqrt{\frac{1}{200.000}} = 0{,}002236068 \sim 0{,}22\,\% \\ q_{Tur} &= \sqrt{\frac{1}{2600}} = 0{,}01961161 \sim 1{,}96\,\% \end{aligned} \tag{2.37}$$

Los portadores serán los heterocigotas, por lo que, si asumimos que las poblaciones se encuentran en equilibrio de Hardy-Weinberg, entonces tendremos

$$\begin{aligned} Het_{Fin} &= 2p_{Fin}q_{Fin} = 2 \times (1 - 0{,}002236068) \times 0{,}002236068 = 0{,}004462136 \sim 0{,}45\,\% \\ Het_{Tur} &= 2p_{Tur}q_{Tur} = 2 \times (1 - 0{,}01961161) \times 0{,}01961161 = 0{,}038454 \sim 3{,}85\,\% \end{aligned} \tag{2.38}$$

2.7 El sistema ABO

Hasta el momento nos hemos centrado en el caso de un locus con dos alelos. En general, la relación entre los distintos alelos de un locus suele ser más compleja. En principio, esto se debe a que existen más de dos alelos cuyo impacto en el fenotipo es distinguible. A su vez, las relaciones entre dichos alelos pueden ser variable, implicando muchas veces diferentes modos de herencia (por ejemplo, dominancia y codominancia en el mismo locus). Un ejemplo de esto es el sistema de grupos sanguíneos ABO, el más importante en humanos,[9] cuya herencia (en forma simplificada) puede apreciarse en la figura 2.6.

El locus codifica para antígenos de superficie en los eritrocitos (glóbulos rojos de la sangre), existiendo tres alelos para el mismo. Básicamente, tanto el alelo I^A como el I^B (codificantes para los antígenos **A** y **B**, respectivamente) dominan al alelo i, que no produce antígenos de superficie (**O**). La relación entre el I^A y el I^B es de codominancia, expresándose ambos antígenos en la superficie de los eritrocitos si estos alelos se encuentran presentes al mismo tiempo. En función de las combinaciones alélicas posibles, existirán cuatro fenotipos diferentes detectables por los test inmunológicos de rutina (por ejemplo, test de aglutinación).

[8] Asumiremos que se trata de un solo tipo de alelo defectuoso. Si bien a nivel molecular los alelos defectuosos pueden ser de varios tipos diferentes, con efectos deletéreos, a nivel fenotípico se comportan como si se tratase de un mismo alelo (ver [Ejercicio 2.6]).

[9] Descubierto por el médico e inmunólogo austríaco Karl Landsteiner en 1901, lo que le valió el premio Nobel de Medicina y Fisiología en 1931.

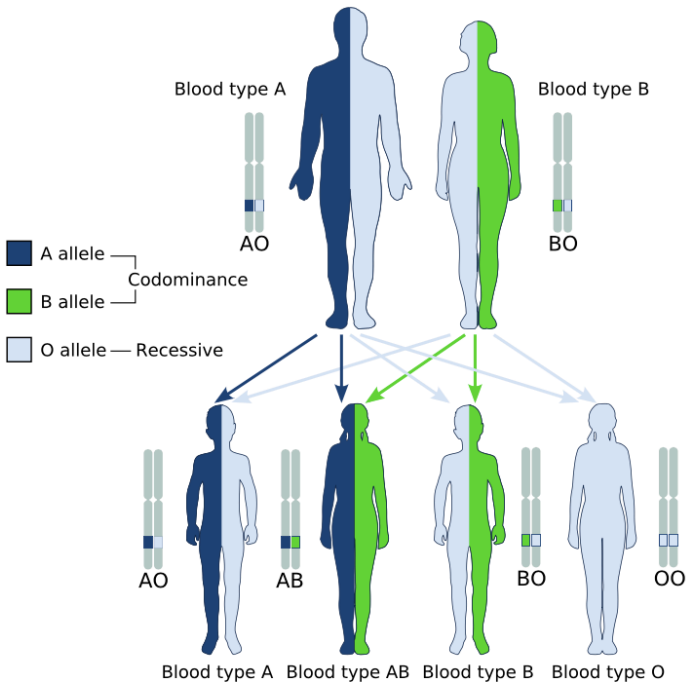


Figura 2.6. Sistema de determinación sanguínea ABO. Los alelos A y B muestran codominancia entre sí, mientras que ambos son dominantes respecto al O. Figura de dominio público tomada de Wikipedia (Archivo: ABO system codominance.svg).

Si denominamos a las frecuencias de los alelos I^A, I^B e i como p, q y r, respectivamente, estos cuatro fenotipos son definidos como

- Fenotipo **A**, formado por los homocigotas del alelo I^A, más los heterocigotos del alelo I^A y el alelo i. La frecuencia esperada bajo equilibrio de Hardy-Weinberg es de $p^2 + 2pr$.
- Fenotipo **B**, formado por los homocigotas del alelo I^B, más los heterocigotos del alelo I^B y el alelo i. La frecuencia esperada bajo equilibrio de Hardy-Weinberg es de $q^2 + 2qr$.
- Fenotipo **AB**, formado por los heterocigotas del alelo I^A y el alelo I^B. La frecuencia esperada bajo equilibrio de Hardy-Weinberg es de $2pq$.
- Fenotipo **O**, formado por los homocigotas del alelo i. La frecuencia esperada bajo equilibrio de Hardy-Weinberg es de r^2.

Notemos que la estimación de las tres frecuencias a partir del conteo de individuos ya no es tan sencilla como en los sistemas previamente analizados. Por un lado, existen en el sistema ABO 3 alelos en lugar de 2. Por el otro, lo que realmente complejiza el panorama es la relación de dominancia-codominancia entre alelos. De hecho, en casi todos los casos resulta imposible inferir el genotipo de un individuo observando su fenotipo; en efecto, esto es solo posible para el fenotipo O (homocigotas del alelo i). En la sección «La estimación de frecuencias en el locus ABO» se desarrolla un método para inferir las frecuencias alélicas a partir del conteo de los cuatro fenotipos detectables por el test de aglutinación.

2.8 ¿Dónde se «esconden» los alelos recesivos?

Muchas mutaciones tienen consecuencias deletéreas o muy graves cuando están presentes en homocigosis, pero no tienen mayores consecuencias para el individuo heterocigota. A modo de ejemplo, se puede pensar en un sistema que requiera de una única copia funcional de una enzima para mantener una actividad metabólica necesaria. Resulta claro que una mutación en homocigosis puede afectar la actividad enzimática (por ejemplo, alterando la capacidad de unión a ligando), alterando así a este sistema metabólico. No obstante, la misma mutación en heterocigosis no afectará la eficiencia del sistema, en tanto seguirá existiendo una copia funcional codificante para la enzima en cuestión.

La diferencia entre el efecto fenotípico de algunos alelos en homocigosis respecto a heterocigosis puede apreciarse también en casos donde una característica evolutivamente favorable se ha transformado con el paso del tiempo en algo negativo (por ejemplo, debido a las implicancias para la domesticación). Este es el caso de los cuernos en el bovino. En las razas bovinas británicas, la ausencia de cuernos (mochos, o *polled* en inglés) se trata de una característica que muestra un evidente modo de herencia dominante. Aun habiéndose podido «mapear» la ubicación de la característica en el genoma bovino, resulta bastante elusivo el mecanismo que genera esta ausencia (Allais-Bonnet et al. 2013; Medugorac et al. 2012). Pese a esto, dado que los individuos con ambos alelos astados (recesivos) presentan un fenotipo astado, resulta bastante tentador evitar que se reproduzcan: intuitivamente se puede presumir que esto permitiría avanzar rápidamente en la eliminación de este alelo de la población, erradicándose así la existencia del fenotipo astado. Sin embargo, si se sigue esta premisa se observa un fenómeno interesante. Luego de una rápida caída en la proporción

de individuos astados, el ritmo de la erradicación comienza a detenerse. Nuestra percepción inicial era apenas una ilusión, y la erradicación del fenotipo comienza a parecer una meta cada vez más lejana. ¿Por qué sucede esto? El principal problema con esta idea es que, al tratarse de un alelo recesivo el que queremos eliminar, resulta imposible distinguir fenotípicamente a los individuos heterocigotas de los homocigotas para el alelo dominante. Para entenderlo, veamos qué ocurre con la relación entre los heterocigotas y los homocigotas cuando nos acercamos a la eliminación del alelo. Sin pérdida de generalidad, llamaremos A_1 al alelo dominante y A_2 al alelo recesivo (con frecuencias p y q, respectivamente). En la situación planteada, tenemos que $q \to 0$ y, por lo tanto, $p \to 1$, por lo que la proporción de heterocigotas respecto a homocigotas estará dada por

$$\frac{2pq}{q^2} = \frac{2p}{q} = \frac{2(1-q)}{q} \approx \frac{2}{q} \therefore \lim_{q \to 0} \frac{2pq}{q^2} = +\infty \tag{2.39}$$

Es decir que, a medida que q tiende a cero, la relación tenderá a infinito. Dicho de otra forma, la enorme mayoría de los alelos recesivos estarán presentes en los individuos heterocigotas, los cuales son fenotípicamente indistinguibles de los homocigotas para el alelo dominante. La figura 2.7 muestra gráficamente esta relación. A medida que nos acercamos hacia la frecuencia de cero para el alelo A_2 (q), la pendiente de la relación se incrementa en valor absoluto (notemos que la derivada primera es igual a $-2/q^2$) y rápidamente la función aparece como vertical.

Como ya se había planteado, observamos que casi todos los alelos provienen de los individuos heterocigotas. A una frecuencia $q = 0{,}02$ (por ejemplo, el 2 % de los alelos son A_2), hay 98 veces más individuos heterocigotas que homocigotas recesivos. Pese a esto, si bien es imposible eliminar los alelos de la población por este mecanismo, sí es posible reducir sustancialmente su incidencia si mantener el criterio de «refugo» no tuviese otras implicancias asociadas.

Ejemplo 2.3

Con los datos del ejemplo 2.2, calcular la relación entre el número de portadores y el número de individuos que padecerán la enfermedad, asumiendo que las poblaciones se encuentran en equilibrio de Hardy-Weinberg.

$$\begin{aligned} q_{Fin} &= \sqrt{\frac{1}{200.000}} = 0{,}002236068 \sim 0{,}22\,\% \\ q_{Tur} &= \sqrt{\frac{1}{2600}} = 0{,}01961161 \sim 1{,}96\,\% \end{aligned} \tag{2.40}$$

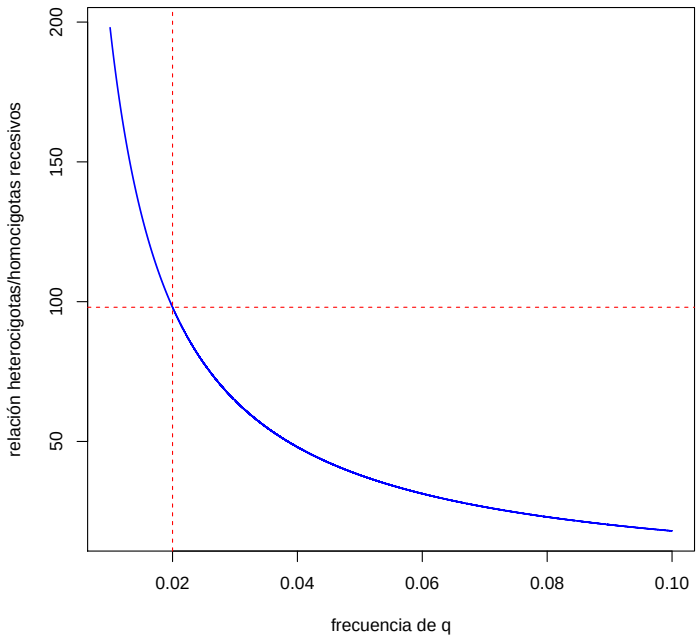


Figura 2.7. Relación del número de heterocigotas a homocigotas recesivos en función de la frecuencia del alelo *q* en un modelo de un locus con dos alelos, bajo equilibrio Hardy-Weinberg

Los portadores serán los heterocigotas, por lo que, si asumimos que las poblaciones se encuentran en equilibrio de Hardy-Weinberg, entonces tendremos

$$\begin{aligned} \frac{Het_{Fin}}{HomR_{Fin}} &= \frac{2p_{Fin}}{q_{Fin}} = 2 \times (1 - 0{,}002236068)/0{,}002236068 = 892{,}43 \\ \frac{Het_{Tur}}{HomR_{Tur}} &= \frac{2p_{Tur}}{q_{Tur}} = 2 \times (1 - 0{,}01961161)/0{,}01961161 = 99{,}98 \end{aligned} \tag{2.41}$$

Es decir, en la población de Turquía tenemos casi 100 portadores por cada enfermo, mientras que en la de Finlandia la relación es de aproximadamente 892 portadores por cada enfermo.

PARA RECORDAR

- La relación entre los heterocigotas y los homocigotas con el alelo recesivo —que sin pérdida de generalidad supondremos es el A_2, con frecuencia q— cuando nos aproximamos a su eliminación (es decir, cuando la frecuencia del alelo A_1 (p) tiende a 1) será

$$\frac{2pq}{q^2} = \frac{2p}{q} = \frac{2(1-q)}{q} \approx \frac{2}{q} \therefore \lim_{q \to 0} \frac{2pq}{q^2} = +\infty \tag{2.42}$$

- Es decir, a medida que q tiende a cero, la relación tenderá a infinito. Dicho de otra forma, **casi todos los alelos recesivos estarán en los individuos heterocigotas**, los que son fenotípicamente imposibles de distinguir de los homocigotas con el alelo dominante.

2.9 Hardy-Weinberg en especies poliploides

El fenómeno de la poliploidía (varias copias del genoma) representa un importante factor evolutivo que lleva a la especiación y es bastante usual en diferentes especies de todos los dominios de la vida, como se observa en la figura 2.8. En particular, la poliploidía es frecuente en muchas de las especies vegetales, y algunas animales, de interés productivo. Ejemplo de esto son las bananas y sandías sin semillas (triploides), el maní y las moras (tetraploides), o el trigo y el boniato (hexaploides). En algunos casos, como el de la frutilla, el número básico de cromosomas (7) es repetido diferente número de veces en las más de 20 especies, existiendo individuos diploides, tetraploides, hexaploides, octaploides y hasta decaploides.

Las frecuencias alélicas esperadas en equilibrio Hardy-Weinberg se pueden calcular a partir de la generalización fórmula que describe el equilibrio para especies diploides. Recordemos del diagrama de Punnett que las frecuencias genotípicas que obtenemos para el caso diploide son el resultado de multiplicar las frecuencias de las combinaciones de alelos correspondientes. Podemos describir este resultado como

$$(f_{\mathbf{A}} + f_{\mathbf{a}})(f_{\mathbf{A}} + f_{\mathbf{a}}) = (p+q)(p+q) = (p+q)^2 \tag{2.43}$$

En el caso diploide, el producto de dos paréntesis $(p + q)$ representa el producto de dos eventos independientes: la generación de gametos. En el caso de los organismos poliploides

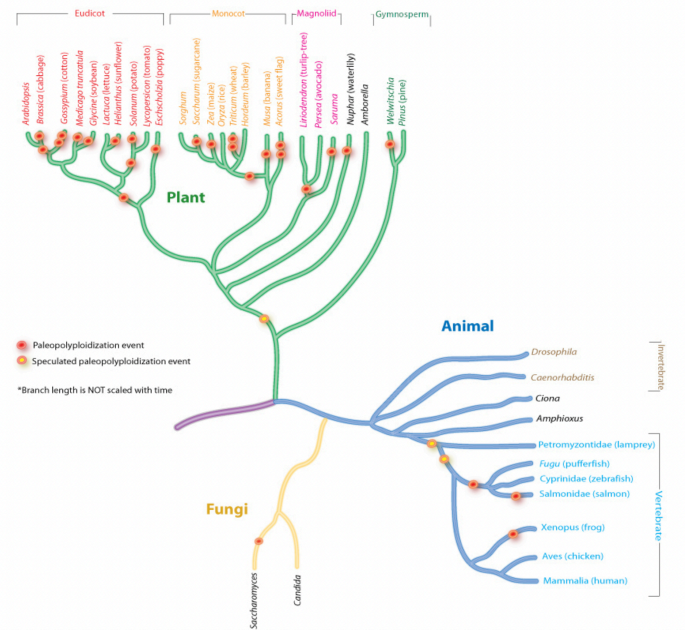


Figura 2.8. Eventos de paleopoliploidías (poliploidías que datan de millones de años) en la evolución, con evidencia fuerte (puntos rojos) y eventos putativos (puntos amarillos). Figura extraída de Wikipedia, creada por Peter Zhang (CC BY-SA 3.0).

(pongamos por caso un tetraploide), existe en primer lugar el proceso de sorteo aleatorio de qué par de alelos terminarán en el mismo gameto, y luego la combinación de gametos para producir el cigoto. Es decir, en el caso del tetraploide hay 4 multiplicaciones de las frecuencias de los alelos **A** y **a**, por lo que ahora la distribución de las frecuencias de los distintos genotipos estará dada por

$$(f_{\mathbf{A}} + f_{\mathbf{a}})(f_{\mathbf{A}} + f_{\mathbf{a}})(f_{\mathbf{A}} + f_{\mathbf{a}})(f_{\mathbf{A}} + f_{\mathbf{a}}) = (p+q)(p+q)(p+q)(p+q) = (p+q)^4 \quad (2.44)$$

Para el caso de un locus con dos alelos en una especie tetraploide, el desarrollo de la fórmula anterior nos lleva a la distribución de las frecuencias que se puede observar en el siguiente cuadro:

Genotipo	Frecuencia
AAAA	p^4
AAAa	$4p^3q$
AAaa	$6p^2q^2$
Aaaa	$4pq^3$
aaaa	q^4

Claramente esto es generalizable a otras ploidías, al menos a las pares (recordar que las ploidías impares suelen tener problemas de segregación), de acuerdo a

$$(f_{\mathbf{A}} + f_{\mathbf{a}}) \times ... \times (f_{\mathbf{A}} + f_{\mathbf{a}}) = (p+q) \times ... \times (p+q) = (p+q)^n \qquad (2.45)$$

donde n corresponde a la ploidía ($n = 2$ para diploides, $n = 4$ para tetraploides, $n = 6$ para hexaploides, etcétera). Obviamente, al tratarse de un locus con dos alelos, los coeficientes que aparecen en el cuadro anterior, y que surgen de la generalización, son los coeficientes del binomio de Newton, o lo que es equivalente del triángulo de Pascal (o de Tartaglia, pese a que el desarrollo de este claramente antecede a ambos).

PARA RECORDAR

Al estimar frecuencias en un sistema de un locus con dos alelos:

- Si suponemos que los individuos representan un muestreo al azar de los gametos en la población y que los individuos de la muestra fueron seleccionados al azar, entonces la probabilidad de observar un conteo x en particular, dado el número total de alelos genotipados n, se puede describir por la distribución binomial: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, con las combinaciones de n elementos tomados de a x elementos dados por $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Equilibrio de Hardy-Weinberg en especies poliploides:

- A modo de ejemplo, en el caso de organismos tetraploides hay **4** multiplicaciones de las frecuencias de los alelos **A** y **a**, por lo que ahora la distribución de las frecuencias de los distintos genotipos será dada por $(p+q)^4$.
- Esto es generalizable a otras ploidías, al menos a las pares de acuerdo a

$$(f_A + f_a) \times \cdots \times (f_A + f_a) = (p+q) \times \cdots \times (p+q) = (p+q)^n$$

2.10 Geometría y genética: los diagramas de de Finetti

Aunque nos pueda parecer algo extraño, existe una fuerte conexión entre genética y geometría. De hecho, como veremos más adelante, Sewall Wright[10] imaginó la evolución como un paisaje adaptativo, con picos y valles por los que se movían las poblaciones (Wright 1932), lo que conecta mentalmente de inmediato a un mundo de formas geométricas. Sin embargo, ya antes de esa fecha, Bruno de Finetti[11] (Finetti 1926) había establecido que gran parte de

[10] Sewall Green Wright (21 de diciembre de 1889 - 3 marzo de 1988) fue un genetista que realizó notables aportes sobre los efectos de la selección y la deriva, así como el análisis de pequeñas poblaciones y el efecto de la consanguinidad. Junto a Ronald A. Fisher y J. B. S. Haldane establecieron la fundación matemática de la genética de poblaciones y de la teoría evolutiva.

[11] Bruno de Finetti (Innsbruck, Austria, 13 de junio de 1906 - Roma, 20 de julio de 1985) fue un probabilista, estadístico y actuario italiano. Entre otros aportes trascendentes, es conocido por sus concepciones «operacionalmente subjetivas» de la probabilidad, que se encuentran dentro de las bases modernas de los enfoques bayesianos.

los mecanismos y resultados de la genética mendeliana se podían obtener de consideraciones geométricas. Uno de los aportes más relevantes en este sentido es la idea de que para un locus con dos alelos en una especie diploide es posible representar todas las combinaciones posibles de las frecuencias genotípicas como los puntos interiores de un triángulo equilátero. Para ello, de Finetti hace uso del resultado del teorema de Viviani, por lo que primero veremos en primer lugar qué establece dicho teorema.

El teorema de Viviani[12] establece que, para triángulos equiláteros, la suma de las distancias desde cualquier punto interno del mismo a los tres lados es igual a la altura del triángulo. Para demostrarlo nos basaremos en el esquema de la figura 2.9, donde dentro del triángulo de vértices **A**, **B** y **C**, con altura h y lado a, tenemos el punto **P**.

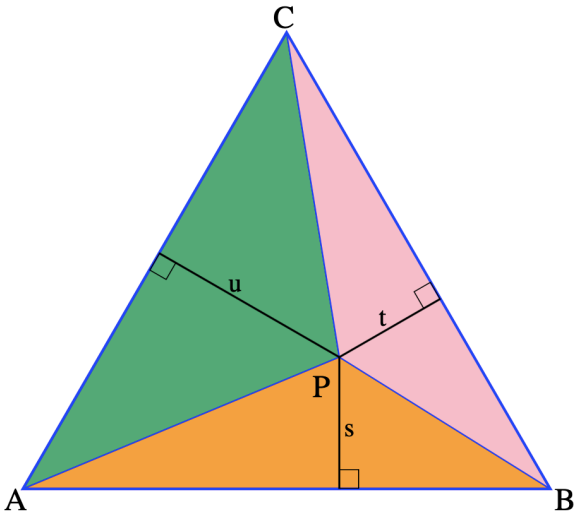


Figura 2.9. Triángulo usado para demostrar el teorema de Viviani. Figura tomada de Wikipedia, CC0 (Archivo: Viviani Theorem.svg).

Llamemos **s**, **t** y **u** a las distancias de **P** a los lados $\overline{\text{AB}}$, $\overline{\text{BC}}$ y $\overline{\text{CA}}$, respectivamente, que son las alturas (por definición) de los tres triángulos internos de vértices **APB**, **BPC** y **CPA**. Estos triángulos, a su vez, completan toda el área del triángulo de vértices **ABC**. Como el área de un triángulo es igual a $\frac{\text{base} \times \text{altura}}{2}$, tenemos entonces que

$$\frac{a.s}{2} + \frac{a.t}{2} + \frac{a.u}{2} = \frac{a.h}{2} \Leftrightarrow \frac{a}{2}(s+t+u) = \frac{a}{2}h \Leftrightarrow s+t+u = h \qquad (2.46)$$

por lo que queda demostrado este teorema. Dicho de otra manera, la significación de este teorema es que para cualquier punto interno a un triángulo equilátero, la suma de distancias a los lados es constante e igual a la altura del triángulo.

Veamos ahora cómo se relaciona este teorema con la genética. Supongamos que tenemos un locus con dos alelos. Asignamos los tres genotipos posibles a los vértices de un triángulo equilátero de altura igual a 1. Sin pérdida de generalidad, pongamos, por ejemplo, los

[12] Vincenzo Viviani (5 de abril de 1622 - 22 de setiembre de 1703), fue un matemático y científico italiano, pupilo de Torricelli y discípulo de Galileo.

homocigotas en la base (es un triángulo equilátero, por lo que todos sus lados son iguales) como muestra la figura 2.10.

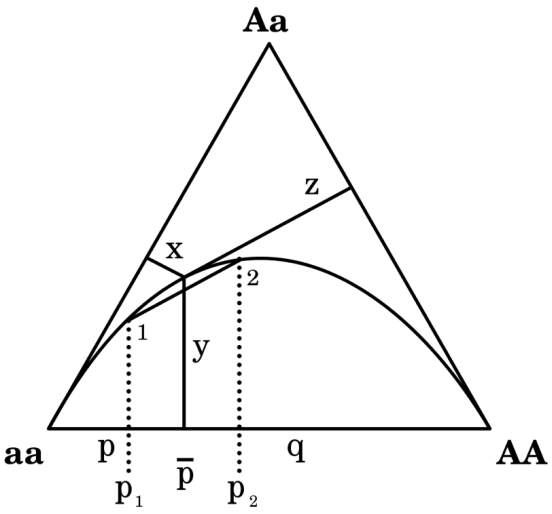


Figura 2.10. Diagrama de de Finetti. Figura tomada de Wikipedia, CC BY-SA 2.5 (Archivo: De Finetti diagram.svg).

Denotamos p a la frecuencia del alelo A y $q = 1 - p$ la frecuencia del alelo a. Ahora, de acuerdo con el teorema de Viviani, sabemos que para cualquier punto interno del triángulo la suma de distancias a los lados es igual a 1 (la altura del triángulo). Pero ¿qué representan estas distancias? Para entender este punto, imaginemos primero que estamos en el vértice **AA** (abajo a la derecha): ahora la distancia de este punto hasta el lado opuesto es igual a 1; lo mismo ocurrirá con los otros dos vértices. Es decir que cuando estamos en algún vértice, la distancia al lado opuesto es igual a 1, o lo que es lo mismo, la frecuencia del correspondiente genotipo es de 1 (todos los individuos son de ese genotipo). Tenemos entonces que cada distancia a un lado del triángulo representa la frecuencia del genotipo en el vértice opuesto. A medida que nos apartamos de los vértices comienzan a aparecer distancias no nulas a los dos otros lados, además de al lado opuesto al vértice. Esto se ve gráficamente en la figura 2.10, donde las distancias **x**, **y** y **z** representan las frecuencias de los genotipos **AA**, **Aa** y **aa**, respectivamente. Que la suma de las frecuencias de los genotipos sea igual a 1 está garantizado por el teorema de Viviani.

Más aún, el punto de intersección de la distancia **y** con el eje formado por los vértices **AA** y **aa** separa a este lado en dos segmentos; el de la izquierda es proporcional a la frecuencia del alelo **A** (p), mientras que el de la derecha es proporcional a la frecuencia del alelo **a** ($p = 1 - q$). La distancia **y** representa la frecuencia de los heterocigotas (**Aa**) y, como ya vimos previamente (sección «H-W: la frecuencia de heterocigotas en función de la frecuencia alélica»), bajo equilibrio de Hardy-Weinberg la misma describe una parábola en función de la frecuencia del alelo **A** (p, o en forma alternativa del otro alelo). Por lo tanto, el equilibrio de Hardy-Weinberg se encuentra representado en la figura 2.10 por la curva que va desde el vértice **aa** al vértice **AA**. Es decir, solo los puntos que se encuentran en esta curva (y, por lo tanto, las distancias asociadas a los lados) representan configuraciones de genotipos que se encuentran en equilibrio de Hardy-Weinberg.

Un ejemplo de la utilidad de lo anterior es la demostración visual de un efecto que veremos más adelante: la unión de (los individuos de) dos poblaciones en equilibrio de Hardy-Weinberg usualmente no produce una población conjunta en equilibrio H-W (antes de la reproducción). Esto suele resultar en la deficiencia en el número de heterocigotos respecto a la frecuencia esperada del promedio de las frecuencias alélicas en las dos poblaciones. Para ver esto en la figura 2.10, supongamos que tenemos dos poblaciones **1** y **2** (del mismo tamaño) que se encuentran en equilibrio H-W y, por lo tanto, son puntos sobre la curva. Las frecuencias del alelo **A** en cada una de estas poblaciones será el segmento de recta izquierdo en el eje entre los vértices **AA** y **aa**, p_1 y p_2, cuyo promedio es $\bar{p}$. Si ahora subimos $\bar{p}$ hasta la curva de equilibrio, vemos que la intersección con el segmento de recta que une los puntos **1** y **2** (el promedio de las frecuencias genotípicas entre estas poblaciones) ocurre antes que con la curva. Lo que es equivalente, la distancia que representa la frecuencia promedio de los heterocigotas es menor a la que representa el valor esperado bajo equilibrio de Hardy-Weinberg.

Un análisis más exhaustivo sobre las posibilidades del diagrama de de Finetti se encuentra en Cannings y Edwards (1968), donde, por ejemplo, los autores muestran que la endocría (*inbreeding*, concepto que veremos también más adelante) produce una parábola por debajo de la que corresponde al equilibrio de H-W, mientras que los procesos de selección producen una serie de «cónicas» (curvas resultantes de cortar un cono variando el ángulo del plano de corte) que también pasan por los vértices **AA** y **aa**.

2.11 La estimación de frecuencias en el locus ABO

Los métodos elaborados para la estimación de las frecuencias hacen uso del concepto de máxima verosimilitud, pero Yasuda (1984) propuso un método sencillo basado en conteo. Supongamos que $T = A + B + AB + O$, es decir el total de individuos de los distintos fenotipos. Trabajaremos con la determinación de p, la frecuencia del alelo I^A, pero para q (la frecuencia del I^B) la situación es idéntica por simetría y basta con sustituir A por B e I^A por I^B. La frecuencia del alelo I^A será igual a la cantidad total de alelos I^A en los fenotipos que tienen este alelo, es decir A (I^AI^A) y AO (I^Ai) que solo expresan el A, más el AB (I^AI^B) que expresa los dos alelos, todo dividido entre $2T$ que es el número total de alelos en los individuos. En el primer caso hay dos alelos I^A por cada individuo, mientras que en los otros dos hay solo un alelo por individuo. Sin embargo, nosotros no podemos distinguir los 3 fenotipos, ya que tanto el A como el AO solo expresan el alelo I^A, por lo que debemos buscar una forma de separar el aporte de cada uno de estos genotipos al conteo. Mientras que el I^Ai aporta un alelo I^A, el I^AI^A aporta dos alelos, por lo que todos los fenotipos A aportarán un alelo, mientras que una cierta fracción (los homocigotos) aportará un segundo alelo también. ¿Cuál es la fracción del fenotipo A que corresponde a los homocigotas I^AI^A? De acuerdo a lo predicho por el equilibrio de Hardy-Weinberg, los homocigotas I^AI^A serán p^2, mientras que los heterocigotas I^Ai serán $2pr$, por lo que la fracción de homocigotas h_A será

$$h_A = \frac{p^2}{p^2 + 2pr} = \frac{p^2}{p(p + 2r)} = \frac{p}{p + 2r} \tag{2.47}$$

Es decir, si sumamos las contribuciones de AB, todos los A y el resto de los A homocigotos, tenemos

$$p = \frac{[AB + A + Ah_A]}{2T} \tag{2.48}$$

Si definimos a partir de los fenotipos observables A y O, $k_A = O/(A + O)$, entonces podemos expresar k_A en términos de lo esperable bajo equilibrio Hardy-Weinberg como

$$\begin{aligned} k_A = \frac{Tr^2}{T(p^2 + 2pr) + Tr^2} = \frac{r^2}{p^2 + 2pr + r^2} = \frac{r^2}{(p+r)^2} = [r/(p+r)]^2 \\ \therefore \sqrt{k_A} = \frac{r}{(p+r)} \end{aligned} \tag{2.49}$$

¿Cuál es la relación entre h_A y k_A? Usando las definiciones de h_A (ecuación (2.47)) y k_A (ecuación (2.49)), tenemos que

$$\begin{aligned} h_A = \frac{p}{p+2r} = \frac{p+r-r}{p+r+r} = \frac{\frac{p+r-r}{p+r}}{\frac{p+r+r}{p+r}} = \frac{\frac{p+r}{p+r} - \frac{r}{p+r}}{\frac{p+r}{p+r} + \frac{r}{p+r}} = \\ = \frac{1 - \frac{r}{p+r}}{1 + \frac{r}{p+r}} = \frac{1 - \sqrt{k_A}}{1 + \sqrt{k_A}} \\ h_A = \frac{1 - \sqrt{k_A}}{1 + \sqrt{k_A}} = \frac{(1 - \sqrt{k_A})}{(1 + \sqrt{k_A})} \frac{(1 - \sqrt{k_A})}{(1 - \sqrt{k_A})} = \\ = \frac{(1 - \sqrt{k_A})^2}{1 - k_A} \end{aligned} \tag{2.50}$$

¡Excelente! Tenemos la relación entre h_A y k_A (expresada de dos formas), pero aún nos falta llegar a una forma de h_A que sea una función de cantidades observables, es decir los fenotipos. Si recordamos la definición de k_A a partir de los fenotipos observables ($k_A = O/(A + O)$) y sustituimos en la ecuación (2.50), tenemos

$$h_A = \frac{(1 - \sqrt{k_A})^2}{1 - k_A} \Leftrightarrow h_A(1 - k_A) = (1 - \sqrt{k_A})^2 \Leftrightarrow h_A(1 - \frac{O}{A+O}) = 1 - 2\sqrt{k_A} + k_A \tag{2.51}$$

Llevando todo a denominador $A+O$, tenemos

$$
\begin{aligned}
h_A(\frac{A+O}{A+O}-\frac{O}{A+O}) = h_A(\frac{A+O-O}{A+O}) = 1-2\sqrt{\frac{O}{A+O}}+\frac{O}{A+O} \Leftrightarrow \\
h_A(\frac{A}{A+O}) = \\
= \frac{A+O}{A+O}-\frac{\sqrt{A+O}}{\sqrt{A+O}}\times 2\frac{\sqrt{O}}{\sqrt{A+O}}+\frac{O}{A+O} \\
h_A(\frac{A}{A+O}) = \frac{(A+O)-2\sqrt{O(A+O)}+O}{A+O} \Leftrightarrow \\
Ah_A(\frac{O}{A+O})\times(A+O) = A+O-2\sqrt{O(A+O)}+O \\
Ah_A = A+O-2\sqrt{O(A+O)}+OAh_A \therefore \\
Ah_A = A+2O-2\sqrt{O(A+O)}
\end{aligned}
\tag{2.52}
$$

Finalmente, si volvemos al principio con este resultado y sustituimos en (2.48), tenemos el estimador basado en conteo

$$
\begin{aligned}
\hat{p} = \frac{[AB+A+Ah_A]}{2T} = \frac{[AB+A+(A+2O-2\sqrt{O(A+O)})]}{2T} \therefore \\
\hat{p} = \frac{[(1/2)AB+A+O-\sqrt{O(A+O)}]}{T}
\end{aligned}
\tag{2.53}
$$

Análogamente, por simetría, para la estimación de q tenemos

$$
\hat{q} = \frac{[(1/2)AB+B+O-\sqrt{O(B+O)}]}{T} \tag{2.54}
$$

y, por lo tanto, $\hat{r} = 1-\hat{p}-\hat{q}$.

Sencillo, ¿no?

Ejemplo 2.4

Entre donantes de sangre de la ciudad de Ankara (Turquía) se determinó que de 1500 donantes sanos se observaron 669 del grupo A, 484 del grupo O, 232 del grupo B y 115 del grupo AB. Utilizando los estimadores que propuso (Yasuda 1984) estimar las frecuencias de los alelos I^A, I^B e i, así como el número esperado de los distintos fenotipos si la población estuviese en equilibrio de Hardy-Weinberg.

$$\begin{aligned}\hat{p} &= \frac{\frac{1}{2}AB + A + O - \sqrt{O(A+O)}}{T} \\ &= \frac{(115/2 + 669 + 484 - \sqrt{484(669+484)})}{1500} \\ &= 0{,}3089808\end{aligned}$$

$$\begin{aligned}\hat{q} &= \frac{\frac{1}{2}AB + B + O - \sqrt{O(B+O)}}{T} \\ &= \frac{(115/2 + 232 + 484 - \sqrt{484(232+484)})}{1500} \\ &= 0{,}1232134\end{aligned}$$

$$\begin{aligned}\hat{r} &= 1 - \hat{p} - \hat{q} = 1 - 0{,}3089808 - 0{,}1232134 \\ &= 0{,}5678058\end{aligned}$$

De acuerdo con las frecuencias estimadas previamente, el número esperado de cada fenotipo sería

$$E(A) = (p^2 + 2pr)T = (0{,}3089808^2 + 2 \times 0{,}3089808 \times 0{,}5678058) \times 1500 = 669{,}527 \sim 670$$

$$E(B) = (q^2 + 2qr)T = (0{,}1232134^2 + 2 \times 0{,}1232134 \times 0{,}5678058) \times 1500 = 232{,}6562 \sim 233$$

$$E(AB) = 2pqT = 2 \times 0{,}3089808 \times 0{,}1232134 \times 1500 = 114{,}2117 \sim 114$$

$$E(O) = r^2T = 0{,}5678058^2 \times 1500 = 483{,}6051 = 484$$

Esto coincide estrechamente con los números observados, por lo que **no podemos descartar que la población se encuentre en equilibrio Hardy-Weinberg** (en general no podemos resolver a favor de una hipótesis alternativa basados en la estadística, a lo sumo podemos afirmar que la misma es completamente improbable dados los supuestos del test que aplicamos).

2.12 Conclusión

En este capítulo planteamos un modelo matemático básico, el equilibrio de Hardy-Weinberg, para realizar predicciones sobre las frecuencias genotípicas observadas en una población partiendo de las frecuencias alélicas existentes en ella. A su vez, observamos cómo el modelo es fácilmente ajustable a distintas situaciones de interés biológico. Las asunciones del modelo pueden resultar muy poco realistas (¿qué población natural tiene tamaño infinito?, ¿dónde se vio que los individuos se apareen al azar?). Si bien esto puede parecer una debilidad del modelo, en capítulos posteriores veremos cómo el equilibrio de Hardy-Weinberg puede utilizarse como una hipótesis nula para contrastar si es viable plantear que una población natural se comporta o no bajo alguna de las premisas del modelo. Gracias a su simpleza, el modelo resulta una excelente puerta de entrada al uso de la matemática para el estudio de la genética de poblaciones, aproximación de la que nos seguiremos nutriendo en lo que resta de este libro.

2.13 Actividades

2.13.1 Control de lectura

1. Varias de las asunciones necesarias para que se cumpla el modelo de equilibrio de Hardy-Weinberg no parecen ser realistas. ¿Cuál es la utilidad del modelo, si esto es así? Plantea al menos un ejemplo donde se viole alguna de las asunciones del modelo, y discute la utilidad del modelo en dicho caso.
2. Describe qué sucede con las frecuencias alélicas esperadas en una población cuando existen diferentes proporciones de individuos machos y hembras.
3. A medida que aumenta el número de alelos presentes en una población en equilibrio de Hardy-Weinberg, aumenta necesariamente el número de genotipos heterocigotas en ella. En estos casos, ¿cómo se puede calcular de forma sencilla la heterocigosidad esperada en la población, sin necesidad de sumar todas las frecuencias de genotipos heterocigotas?
4. Si un locus se ve asociado a un cromosoma sexual, ¿se comporta igual que un locus autosómico? Describe qué dinámica esperas para las frecuencias alélicas en machos y hembras, así como el valor esperado para la frecuencia alélica a largo plazo en la población.
5. Las frecuencias alélicas deben ser estimadas a través de muestreos cuando se realiza un estudio poblacional. *a)* ¿En qué medida afectan los tamaños muestrales en este contexto? *b)* Nombra al menos un método que permita la determinación de los alelos presentes en un individuo muestreado.

2.13.2 ¿Verdadero o falso?

1. Si se cambian las frecuencias alélicas p y q de una población que se encuentra en equilibrio Hardy-Weinberg, estas se ajustarán en la próxima generación y volverán a su valor inicial (lo cual implica que las frecuencias genotípicas permanecerán invariables). A esta propiedad es que alude la palabra «equilibrio».
2. El modelo de equilibrio de Hardy-Weinberg predice que un alelo que sea letal en homocigosis no será eliminado de una población si el fenotipo heterocigota es indistinguible del resto de los fenotipos homocigotas.
3. La suma de las frecuencias genotípicas en una población que segrega un locus bialélico (con frecuencias p y q) siempre cumple con la ecuación $p^2 + 2pq + q^2 = 1$.
4. Para un locus bialélico, la máxima heterocigosidad esperada corresponde al caso donde ambas frecuencias alélicas son iguales.
5. Si dos poblaciones se encuentran bajo equilibrio de Hardy-Weinberg, una población constituida por ambas también lo hará.

Solución

1. **Falso.** Un cambio en las frecuencias alélicas llevará a un cambio de las frecuencias genotípicas esperadas bajo equilibrio de Hardy-Weinberg. La palabra «equilibrio» refiere al hecho de que, una vez establecidas estas frecuencias genotípicas, estas permanecerán iguales si las condiciones no cambian.

2. **Verdadero.** A medida que la frecuencia del alelo disminuye, el modelo prevé una prevalencia del alelo en los individuos heterocigotas (los cuales no son eliminados, ya

que su fenotipo no sufre los efectos de cargar con el alelo). El alelo por tanto no es erradicado de la población.

3. **Falso.** Es correcto que la suma de las frecuencias genotípicas en una población debe equivaler a 1, ya que la suma de todas las frecuencias equivale al total de los casos posibles. No obstante, la ecuación $p^2+2pq+q^2=1$ solo será correcta si la población se encuentra bajo equilibrio de Hardy-Weinberg, ya que en dicho caso las frecuencias de genotipos homocigotas y heterocigotas se encuentran reflejadas en la ecuación.

4. **Verdadero.** En efecto, este caso implica que las frecuencias alélicas sean $p=q=0{,}5$, con una hetercigosidad esperada de $2\times 0{,}5\times 0{,}5=0{,}5$. Véase la sección «Ejercicios», donde se propone demostrar esta propiedad mencionada en el texto del capítulo.

5. **Falso.** La población total no se encontrará bajo equilibrio de Hardy-Weinberg. La causa de fondo es que los individuos de cada una de las poblaciones no se aparean con los de la otra (por ejemplo, no existe panmixia en la población total). Un cálculo sencillo muestra que existirá un déficit de heterocigotas (lo cual es observable en los diagramas de de Finetti). Ahondaremos en este fenómeno en el capítulo «Apareamientos no-aleatorios».

2.13.3 Ejercicios

Ejercicio 2.1

Un locus A posee dos alelos conocidos (A_1 y A_2), los cuales segregan en dos poblaciones independientes de organismos diploides. Las poblaciones tienen tamaños $N_1=N$ y $N_2=2N$, respectivamente. Las frecuencias genotípicas en las poblaciones son

$$\begin{cases} f_1(A_1A_1)=0{,}36;\ f_1(A_1A_2)=0{,}48;\ f_1(A_2A_2)=0{,}16 \\ f_2(A_1A_1)=0{,}09;\ f_2(A_1A_2)=0{,}42;\ f_2(A_2A_2)=0{,}49 \end{cases}$$

Se asume que en ambos casos el locus segrega acorde a lo esperado bajo equilibrio de Hardy-Weinberg. Calcula las frecuencias genotípicas esperadas si se unen las poblaciones.

Solución

Para poder calcular las frecuencias genotípicas esperadas bajo equilibrio de Hardy-Weinberg, debemos calcular las nuevas frecuencias alélicas en la población resultante de la unión de ambas poblaciones, las cuales llamaremos $f_T(A_1)$ y $f_T(A_2)$. Una vez obtenidos estos valores, podemos proceder a calcular las frecuencias genotípicas esperadas.

La frecuencia alélica $f_T(A_1)$ corresponde a

$$f_T(A_1)=\frac{\text{núm. alelos } A_1}{\text{núm. total de alelos}}=\frac{(2N_1\times f_1(A_1))+(2N_2\times f_2(A_1))}{2N_1+2N_2}$$

donde $f_1(A_1)=f_1(A_1A_1)+\frac{1}{2}\times f_1(A_1A_2)$ y $f_2(A_1)=f_2(A_1A_1)+\frac{1}{2}\times f_2(A_1A_2)$

Sustituyendo, tenemos

$$f_T(A_1) = \frac{(2N_1 \times [f_1(A_1A_1) + \frac{1}{2} \times f_1(A_1A_2)]) + (2N_2 \times [f_2(A_1A_1) + \frac{1}{2} \times f_2(A_1A_2)])}{2N_1 + 2N_2}$$

$$f_T(A_1) = \frac{(2N \times [f_1(A_1A_1) + \frac{1}{2} \times f_1(A_1A_2)]) + (2(2N) \times [f_2(A_1A_1) + \frac{1}{2} \times f_2(A_1A_2)])}{2N + 2(2N)}$$

$$f_T(A_1) = \frac{(2N \times [0{,}36 + \frac{1}{2} \times 0{,}48]) + (2(2N) \times [0{,}09 + \frac{1}{2} \times 0{,}42])}{2N + 2(2N)} \tag{2.55}$$

$$f_T(A_1) = 0{,}4$$

Por analogía, se puede calcular $f_T(A_2)$. A su vez, se puede considerar que $f_T(A_1) + f_T(A_2) = 1 \Rightarrow f_T(A_2) = 1 - f_T(A_1) = 1 - 0{,}4 = 0{,}6$

De esto se desprende que

$$\begin{cases} f_T(A_1A_1) = 0{,}4^2 = 0{,}16 \\ f_T(A_1A_2) = 2 \times 0{,}4 \times 0{,}6 = 0{,}48 \\ f_T(A_2A_2) = 0{,}6^2 = 0{,}36 \end{cases}$$

Ejercicio 2.2

Dos poblaciones independientes de igual tamaño (N) se encuentran divididas por una barrera geográfica que impide la reproducción entre organismos de ambos lugares. Las poblaciones segregan un locus A con tres alelos conocidos (A_1, A_2 y A_3), acorde a lo esperado en equilibrio de Hardy-Weinberg.

Las poblaciones presentan las siguientes frecuencias genotípicas:

$$\begin{cases} f_1(A1) = 0{,}6; f_1(A_2) = 0{,}4; f_1(A_3) = 0 \\ f_2(A1) = 0; f_2(A_2) = 0{,}3; f_2(A_3) = 0{,}7 \end{cases}$$

En un momento dado desaparece la barrera geográfica que dividen a las poblaciones, pasándose a un régimen de panmixia para la población resultante.

a. Calcula el promedio de la heterocigosidad de las poblaciones, previa a la ruptura de la barrera geográfica.

b. Calcula la heterocigosidad en la población resultante de la unión de las poblaciones, asumiendo que esta cumple con las asunciones del equilibrio Hardy-Weinberg.

c. ¿Son iguales los valores obtenidos en a y b? Vincula tu respuesta a las asunciones del equilibrio Hardy-Weinberg.

Solución

a. Calculamos la heterocigosidad esperada en cada caso:

$$H_1 = 2 \times 0{,}6 \times 0{,}4 = 0{,}48$$

$$H_2 = 2 \times 0{,}7 \times 0{,}3 = 0{,}42$$

de esto se desprende

$$\overline{H} = \frac{H_1 + H_2}{2} = \frac{0{,}48 + 0{,}42}{2} = 0{,}45$$

b. Primero debemos recalcular las frecuencias alélicas ($f_T(A_i)$, para cada alelo i), a efectos de poder estimar frecuencias genotípicas bajo equilibrio de Hardy-Weinberg en la población total.

Tenemos que

$$\begin{cases} f_T(A_1) = \frac{2N \times 0{,}6 + 2N \times 0}{2N+2N} = \frac{0{,}6}{2} \\ f_T(A_2) = \frac{2N \times 0{,}4 + 2N \times 0{,}3}{2N+2N} = \frac{0{,}4+0{,}3}{2} = \frac{0{,}7}{2} \\ f_T(A_3) = \frac{2N \times 0{,}7 + 2N \times 0}{2N+2N} = \frac{0{,}7}{2} \end{cases} \tag{2.56}$$

Notemos que las frecuencias alélicas suman 1.

La heterocigosidad esperada se puede calcular a partir de la homocigosidad, notando que

$$H = 1 - \sum_{i=1}^{i=3} A_i^2$$

por lo que

$$H_T = 1 - (0{,}3^2 + 0{,}35^2 + 0{,}35^2)$$

$$H_T = 0{,}665$$

c. La heterocigosidad esperada bajo equilibrio Hardy-Weinberg para la población en conjunto es mayor al promedio de heterocigosidad entre ambas. Esto es de esperar (el lector puede referirse a la subsección 2.10 «Geometría y genética: los diagramas de de Finetti» para una demostración visual del fenómeno). ¿A qué se debe esto? Una forma de aproximarse al problema es pensar desde el inicio en la población en su totalidad: bajo equilibrio de Hardy-Weinberg se asume panmixia, y los heterocigotas son producto del apareamiento al azar en la población. Cuando la población se estructura en dos o más grupos que solo se reproducen entre sí (en nuestro caso, las poblaciones iniciales) se rompe este régimen de panmixia general, disminuyendo el número de heterocigotas producidos. Una forma clara de ver esto en el ejemplo planteado es que, cuando se consideran dos poblaciones, no se generan individuos heterocigotas A_1A_3.

Ejercicio 2.3

Un locus A presenta seis alelos que segregan acorde a lo esperado por el equilibrio Hardy-Weinberg. Las proporciones alélicas son las siguientes:

$$\begin{cases} f(A_1) = 0{,}15 \\ f(A_2) = 0{,}25 \\ f(A_3) = 0{,}3 \\ f(A_4) = 0{,}09 \\ f(A_5) = 0{,}2 \\ f(A_6) = 0{,}01 \end{cases}$$

a. Calcula la heterocigosidad esperada para el locus.

b. Durante un período las condiciones ambientales cambian, y se pierden todos los individuos con alelos A_4, A_5 y A_6. Una vez reestablecido el equilibrio de Hardy-Weinberg, las proporciones alélicas son ahora

$$\begin{cases} f(A_1) = 0{,}1 \\ f(A_2) = 0{,}7 \\ f(A_3) = 0{,}2 \end{cases}$$

¿La heterocigosidad aumentó o disminuyó en la población? ¿Qué tanto respecto a lo obtenido en la primera parte del ejercicio?

Solución

a. Recordemos que la frecuencia total de heterocigotas (heterocigosidad) se puede calcular a partir de la frecuencia total de homocigotas, en tanto la suma de ambas frecuencias debe sumar uno (ya que un organismo es o bien homocigota o bien heterocigota).

Tenemos por tanto que

$$H = 1 - \sum_{i=1}^{i=6} A_i^2$$

por lo que

$$H = 1 - (0{,}15^2 + 0{,}25^2 + 0{,}3^2 + 0{,}09^2 + 0{,}2^2 + 0{,}01^2)$$
$$H = 0{,}7768$$

b. Aplicando la misma lógica, tenemos que en el nuevo escenario

$$H = 1 - (0{,}1^2 + 0{,}7^2 + 0{,}2^2) = 0{,}46$$

La heterocigosidad disminuyó. Aplicando un cociente, podemos expresar qué fracción del primer valor representa este nuevo valor de heterocigosidad. Llamaremos a estas heterocigosidades H_1 y H_2. Tenemos que

$$\frac{H_2}{H_1} = \frac{0{,}46}{0{,}7768} \approx 0{,}59$$

Vemos que la heterocigosidad disminuyó en el segundo escenario a casi el 60 % del valor planteado en primera instancia.

Ejercicio 2.4

Dos loci bialélicos, A y B, se encuentran segregando de forma independiente en una población de organismos diploides. Las frecuencias alélicas en la población son

$$\begin{cases} f(A_1) = 0{,}2 \\ f(A_2) = 0{,}8 \\ f(B_1) = 0{,}4 \\ f(B_2) = 0{,}6 \end{cases}$$

a. Realiza un diagrama de Punnett mostrando los posibles genotipos en la población.

b. Calcula las frecuencias genotípicas esperadas bajo equilibrio de Hardy-Weinberg.

Solución

a. A continuación se muestra el cuadro de 16 combinaciones genotípicas posibles, coloreadas según el genotipo resultante:

	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1	A_1A_1/B_1B_1	A_1A_1/B_1B_2	A_1A_2/B_1B_1	A_1A_2/B_1B_2
A_1B_2	A_1A_1/B_2B_1	A_1A_1/B_2B_2	A_1A_2/B_2B_1	A_1A_2/B_2B_2
A_2B_1	A_2A_1/B_1B_1	A_2A_1/B_1B_2	A_2A_2/B_1B_1	A_2A_2/A_1B_2
A_2B_2	A_2A_1/B_2B_1	A_2A_1/B_2B_2	A_2A_2/B_2B_1	A_2A_2/B_2B_2

b. Los loci segregan de forma independiente, por lo que la presencia de un alelo de A no altera la probabilidad de encontrar un determinado alelo de B. Se calculan las frecuencias genotípicas para cada locus, y la multiplicación de estos equivale a la frecuencia de organismos que portarán los genotipos correspondientes en simultáneo

$$\begin{cases} f(A_1A_1) = 0{,}2^2; f(A_1A_2) = 2 \times 0{,}2 \times 0{,}8; f(A_2A_2) = 0{,}8^2 \\ f(B_1B_1) = 0{,}4^2; f(B_1B_2) = 2 \times 0{,}4 \times 0{,}6; f(B_2B_2) = 0{,}6^2 \end{cases}$$

por lo que

$$\begin{cases} f(A_1A_1; B_1B_1) = 0{,}2^2 \times 0{,}4^2 = 0{,}0064 \\ f(A_1A_1; B_1B_2) = 0{,}2^2 \times (2 \times 0{,}4 \times 0{,}6) = 0{,}0192 \\ f(A_1A_1; B_2B_2) = 0{,}2^2 \times 0{,}6^2 = 0{,}0144 \\ f(A_1A_2; B_1B_1) = (2 \times 0{,}2 \times 0{,}8) \times 0{,}4^2 = 0{,}0512 \\ f(A_1A_2; B_1B_2) = (2 \times 0{,}2 \times 0{,}8) \times (2 \times 0{,}4 \times 0{,}6) = 0{,}1536 \\ f(A_1A_2; B_2B_2) = (2 \times 0{,}2 \times 0{,}8) \times 0{,}6^2 = 0{,}1152 \\ f(A_2A_2; B_1B_1) = 0{,}8^2 \times 0{,}4^2 = 0{,}1024 \\ f(A_2A_2; B_1B_2) = 0{,}8^2 \times (2 \times 0{,}4 \times 0{,}6) = 0{,}3072 \\ f(A_2A_2; B_2B_2) = 0{,}8^2 \times 0{,}6^2 = 0{,}2304 \end{cases}$$

Ejercicio 2.5

Un factor de transcripción (denotado por F) regula la expresión de un gen codificante para una enzima (denotada por E) vinculada a la coloración en una especie diploide. Se asume que ambos loci segregan de forma independiente.

Los siguientes alelos son conocidos para estos loci - Alelo F_a: alelo codificante para una copia funcional del factor de transcripción. - Alelo F_b: alelo codificante para una copia disfuncional del factor de transcripción. - Alelo E_a: alelo codificante para una copia funcional de la enzima. - Alelo E_b: alelo codificante para una copia disfuncional de la enzima.

La presencia de una copia funcional para el factor de transcripción basta para que el sistema sea funcional. En cambio, la coloración de los organismos depende del número de copias funcionales para la enzima E: dos copias funcionales resultan en coloración roja, una copia funcional resulta en coloración rosada y la ausencia de copias funcionales resulta en ausencia de color.

a. Considera una población de tamaño N, donde se tienen las siguientes frecuencias alélicas: $f(F_a) = 0{,}5$; $f(E_a) = 1$. Describe la proporción de individuos que presentan coloración roja, rosada y nula en la población si se asume que esta se encuentra en equilibrio de Hardy-Weinberg.

b. Considera una población de tamaño N, donde se tienen las siguientes frecuencias alélicas: $f(F_a) = 1$; $f(E_a) = 0{,}5$. Describe la proporción de individuos que presentan coloración roja, rosada y nula en la población si se asume que esta se encuentra en equilibrio de Hardy-Weinberg.

c. Considera la población resultante de unificar las poblaciones anteriormente mencionadas. Describe la proporción de individuos que presentan coloración roja, rosada y nula en la población si se asume que esta se encuentra en equilibrio de Hardy-Weinberg.

Solución

a. La proporción de individuos que mostraran algún tipo de coloración corresponde a aquellos que presentan alguna copia funcional para el factor de transcripción F. La proporción de individuos homocigotas F_bF_b para el alelo disfuncional está dada por $f(F_b)^2 = 0{,}5^2 = 0{,}25$. Por lo tanto, la proporción de individuos que presentan coloración equivale a $1 - 0{,}25 = 0{,}75$. De estos, la totalidad portarán dos alelos E_a, lo que resulta en coloración roja. Por lo tanto, se esperan las siguientes proporciones de coloración: $f(\text{Rojo}) = 0{,}75$, $f(\text{Rosado}) = 0$; $f(\text{Albino}) = 0{,}25$.

b. La totalidad de los organismos serán homocigotas F_aF_a para el locus F, por lo cual la coloración dependerá exclusivamente de las frecuencias esperadas de individuos E_aE_a, E_aE_b y E_bE_b para el locus E. Tenemos que

$$\begin{cases} f(\text{Rojo}) = f(E_aE_a) = f(E_a)^2 = 0{,}5^2 = 0{,}25 \\ f(\text{Rosado}) = f(E_aE_b) + f(E_bE_a) = 2 \times f(E_a) \times f(E_b) = 2 \times 0{,}5 \times 0{,}5 = 0{,}5 \\ f(\text{Albino}) = f(E_bE_b) = f(E_b)^2 = 0{,}5^2 = 0{,}25 \end{cases}$$

c) El fenotipo de coloración dependerá tanto de los alelos presentes en el locus E como en el locus F. Al combinar las poblaciones, las frecuencias alélicas (y genotípicas) cambiarán y, por lo tanto, cambiará la proporción esperada para cada fenotipo. Los loci segregan de forma independiente, por lo cual podemos obtener las frecuencias genotípicas esperadas para cada locus por separado, y luego multiplicar las frecuencias resultantes para obtener las frecuencias esperadas en conjunto, las cuales correlacionan con los fenotipos esperados. Para ello, primero necesitamos establecer las nuevas frecuencias alélicas $f(F_a);\ f(F_b);\ f(E_a);\ f(E_b)$ para la población resultante.

Observemos qué pasa para un caso particular, el alelo F_a

$$f(F_a) = \frac{\text{núm. alelos } F_a}{2N + 2N} = \frac{(2N \times f_1(F_a) + 2(N \times f_2(F_a)))}{4N} = \frac{2 \times 0{,}5 + 2 \times 1}{4} = 0{,}75 \tag{2.57}$$

Por analogía, se calculan las nuevas frecuencias alélicas en la población en el resto de los casos

$$\begin{cases} f(F_b) = 0{,}25 \\ f(E_a) = 0{,}75 \\ f(E_b) = 0{,}25 \end{cases}$$

Por lo tanto, las frecuencias genotípicas esperadas en equilibrio Hardy-Weinberg para cada locus, son

$$\begin{cases} f(F_aF_a) = 0{,}75^2; f(F_aF_b) = 2 \times 0{,}75 \times 0{,}25; f(F_bF_b) = 0{,}25^2 \\ f(E_aE_a) = 0{,}75^2; f(E_aE_b) = 2 \times 0{,}75 \times 0{,}25; f(E_bE_b) = 0{,}25^2 \end{cases}$$

Considerando los dos genotipos al mismo tiempo, podemos dar las frecuencias fenotípicas esperadas para coloración.

$$\begin{cases} f(\text{Rojo}) = f(F_aF_a) \times f(E_aE_a) = 0{,}75^2 \times 0{,}75^2 \\ f(\text{Albino}) = f(F_bF_b;\ E.E.) + f(F.F.; E_bE_b) - f(F_bF_b;\ E_bE_b) = \\ 0{,}25^2 + 0{,}25^2 - (0{,}25^2 \times 0{,}25^2) \end{cases}$$

en donde en el último caso restamos $f(F_bF_b;\ E_bE_b)$ para no contar este genotipo dos veces.

Para facilitar los cálculos, notemos que los individuos que no sean rojos o albinos deberán ser necesariamente rosados, por la configuración del sistema. Esto nos da

$$f(\text{Rosado}) = 1 - (0{,}75^2 \times 0{,}75^2) - (0{,}25^2 + 0{,}25^2 - (0{,}25^2 \times 0{,}25^2))$$
$$f(\text{Rosado}) = 0{,}5625$$

Se puede verificar que esta frecuencia es equivalente a sumar las frecuencias de los genotipos que expresan este color.

Σ Ejercicio 2.6

Demuestra: (a) que para un locus con dos alelos en una especie diploide en equilibrio Hardy-Weinberg, la diferencia de frecuencias alélicas es igual a la diferencia de frecuencia entre los homocigotos respectivos y (b) que esta diferencia es igual a $2p - 1$.

Solución

Supongamos que llamamos p y $q = 1 - p$ a las frecuencias de los dos alelos. Para la parte (a) debemos demostrar que su diferencia es igual a la diferencia entre los genotipos homocigotos, es decir, $p - q = p^2 - q^2$. Para ello utilizaremos el hecho de que multiplicar por 1 ambos lados de la ecuación y el hecho de que $p + q = 1$

$$(p - q) = 1(p - q) = (p + q)(p - q) = p^2 + pq - pq - q^2 = p^2 - q^2 \tag{2.58}$$

En tanto, para la parte (b) utilizaremos que $q = 1 - p$, por lo que

$$p - q = p - (1 - p) = p - 1 + p = 2p - 1 \tag{2.59}$$

Ejercicio 2.7

En general, si bien en muchas enfermedades humanas y animales que se clasifican como de herencia mendeliana se considera que se trata de un solo alelo recesivo, a nivel molecular suelen coexistir varios alelos deletéreos que producen similar fenotipo. Demuestra que, si en un locus existen dos alelos deletéreos recesivos respecto al alelo normal, bajo equilibrio de Hardy-Weinberg los resultados respecto al número de portadores y de enfermos son equivalentes a considerarlos como un solo alelo defectuoso.

Solución

Supongamos que llamamos A_1 al alelo normal y A_2 y A_3 a los alelos defectuosos, con frecuencias respectivas p, q y r, $p+q+r=1$. Llamemos $u=q+r$ a la frecuencia de los dos alelos defectuosos sumados o, lo que es lo mismo, a la frecuencia de alelos defectuosos (como si fuera uno solo). De acuerdo con lo visto para más de un alelo, las frecuencias esperadas bajo equilibrio Hardy-Weinberg serán

$$(p+q+r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr \tag{2.60}$$

De estos genotipos, los enfermos serán aquellos que tengan 2 alelos defectuosos, sean cuales sean, o sea A_2A_2, A_3A_3 y A_2A3, es decir

$$q^2 + r^2 + 2qr = (q+r)^2 = u^2 \tag{2.61}$$

Es decir, demostramos que el número de enfermos esperados es equivalente a considerarlo como un solo alelo. Ahora, el número de portadores será la suma de los genotipos que tienen un solo alelo defectuoso y el otro normal, o sea A_1A_2 y A_1A_3, es decir

$$2pq + 2pr = 2p(q+r) = 2pu \tag{2.62}$$

con lo que queda demostrado que también la frecuencia de portadores es equivalente a considerarlo un solo alelo defectuoso.

Ejercicio 2.8

En muchos casos las ecuaciones que derivamos se simplifican enormemente debido a la restricción $p + q = 1$. Demuestra que las siguientes relaciones son válidas (adaptado de Hartl y Clark (2007)):

$$\begin{aligned} p^2 - q^2 &= p - q \\ p^2 + pq &= p \\ p - q &= 1 - 2q \\ p^2 + q^2 &= 1 - 2pq \\ (p - q)^2 &= 1 - 4pq \end{aligned} \tag{2.63}$$

Solución

Haciendo uso de que $p + q = 1$, por lo que también $(p + q)^2 = p^2 + 2pq + q^2 = 1$ (ya que lo que está dentro del paréntesis vale 1 y $1^2 = 1$), tenemos que

$$\begin{aligned} p^2 - q^2 &= (p + q)(p - q) \\ &= (1)(p - q) \\ &= p - q \end{aligned}$$

$$\begin{aligned} p^2 + pq &= p(p + q) \\ &= p(1) \\ &= p \end{aligned}$$

$$\begin{aligned} p - q &= (1 - q) - q \\ &= 1 - q - q \\ &= 1 - 2q \end{aligned}$$

$$\begin{aligned} p^2 + q^2 &= 1 - 2pq \\ &\Leftrightarrow p^2 + q^2 + 2pq = 1 \\ &\Leftrightarrow (p + q)^2 = 1 \\ &\Leftrightarrow 1^2 = 1 \end{aligned}$$

$$\begin{aligned} (p - q)^2 &= p^2 - 2pq + q^2 + 2pq - 2pq \\ &= p^2 + 2pq + q^2 - 2pq - 2pq \\ &= (p + q)^2 - 4pq \\ &= 1 - 4pq \end{aligned}$$

Ejercicio 2.9

Demuestra que bajo equilibrio de Hardy-Weinberg, en un locus con dos alelos, 4 veces el producto de las frecuencias relativas de los homocigotos es igual al cuadrado de la frecuencia relativa de los heterocigotos.

Solución

Las frecuencias de los homocigotos bajo equilibrio de HW son p^2 y q^2, mientras que la de los heterocigotos es $2pq$. Por lo tanto, debemos ver si $4p^2q^2 \stackrel{?}{=} (2pq)^2 \Leftrightarrow 4p^2q^2 = 4p^2q^2$, que en efecto se cumple.

Capítulo 3

Deriva genética

Quantum mechanics is certainly imposing. But an inner voice tells me that it is not yet the real thing. The theory says a lot, but does not really bring us closer to the secret of the 'Old One.' I, at any rate, am convinced that He does not throw dice.[1]

Albert Einstein

Einstein, stop telling God what to do with His dice.[2]

Niels Bohr

En el capítulo anterior («Variación y equilibrio de Hardy-Weinberg») comenzamos a entender los procesos que hay por detrás del mantenimiento de la variabilidad genética en las poblaciones de distintas especies. En particular, a partir de las reglas mecanísticas de la herencia mendeliana[3] vimos que una vez alcanzado el equilibrio, si no aparecen otras fuerzas, las frecuencias alélicas y genotípicas se mantienen constantes generación tras generación. Para esto hemos asumido varias cosas («Supuestos que asumimos se cumplen para H-W»), algunas más razonables que otras. En particular, una de las cosas que asumimos es que el tamaño de nuestra población es prácticamente infinito (con esto queremos decir un número realmente muy alto de individuos), una situación que no siempre se cumple en la práctica. Más aún, esto nos lleva al concepto de población en nuestro contexto: el conjunto de individuos que potencialmente podrían aparearse en igualdad de condiciones con cualquier otro de la misma población para dejar descendencia fértil. Se trata de un concepto fácil de comprender, pero en la práctica suele ser bastante más complejo delimitar sus alcances. Pensemos, por ejemplo, en las poblaciones de corriedale (ovinos), angus (bovino de carne) o de holando (bovino de leche) del Uruguay. ¿Se cumplen los supuestos del equilibrio de Hardy-Weinberg? ¿Cuál sería mi población si el rebaño o rodeo es cerrado (por ejemplo, produce sus propios reemplazos)?

Habiendo definido de forma operativa lo que consideramos una población, rápidamente podemos llegar a ver que en muchas poblaciones asumir un tamaño casi infinito para las estas es algo absurdo. Veremos que en las poblaciones donde se deja de lado esta asunción empieza a jugar una de las fuerzas siempre presentes en biología: **el azar**.

[1] *The Born-Einstein Letters*, by Max and Hedwig Born. La traducción podría ser algo así como «La mecánica cuántica es ciertamente imponente. Pero una voz interior me dice que aún no es la cosa real. La teoría dice mucho, pero no nos acerca realmente al secreto del 'Viejo'. En cualquier caso, estoy convencido de que Él no juega a los dados».

[2] Respuesta atribuida a Niels Bohr en los intercambios orales entre ambos, aunque no existe una referencia exacta. La traducción podría ser algo así como «Einstein, deja de decirle a Dios qué hacer con sus dados».

[3] Gregor Mendel (20 de julio de 1822 - 6 de enero de 1884) fue un meteorólogo, biólogo y matemático, nacido en Heizendorf bei Odrau, Imperio Austro-Húngaro (hoy República Checa). Considerado el padre de la genética moderna, realizó diversos experimentos con plantas que lo llevaron a postular algunas leyes, decenas de años antes de conocerse las bases moleculares de la herencia.

OBJETIVOS DEL CAPÍTULO

☐ Comprender cómo el azar puede ser una fuerza evolutiva de relevancia en poblaciones de tamaño finito.

☐ Modelar este fenómeno a través de modelos matemáticos; esto nos permitirá derivar predicciones respecto a qué se espera de la evolución de las frecuencias alélicas en un conjunto de poblaciones iguales. En particular, veremos el modelo de Wright-Fisher y la aplicación de cadenas de Markov en este problema.

☐ Describir la evolución de la homocigosidad en poblaciones de tamaño finito conforme pasan las generaciones. Se presentará el coeficiente de fijación y su tendencia según el modelo de Wright-Fisher.

☐ Presentar el concepto de tamaño poblacional efectivo, el cual permite tratar a una población que se desvía de lo ideal con las herramientas matemáticas presentadas hasta ahora en distintos contextos biológicos.

☐ Ver la aproximación por difusión, una herramienta matemática que permite integrar los procesos azarosos de cambio en las frecuencias alélicas con los procesos direccionados. Este modelado conjunto es más realista, en tanto en las poblaciones naturales es de esperar que los dos fenómenos tengan influencia.

☐ Presentar el modelo de coalescente, un modelo clave en genética de poblaciones. Este nos permitirá indagar en la historia de una muestra de alelos, yendo hacia atrás en el tiempo (hasta identificar su último ancestro común en una población), estimando los tiempos esperados en el proceso.

3.1 El rol de los procesos estocásticos en la genética

Para estudiar el alcance del azar en nuestro contexto propongamos un experimento conceptual sencillo, que cada persona puede reproducir en su casa. Tomemos una jarra o vaso grande y coloquemos 20 bolitas de vidrio azules y 20 bolitas de vidrio rojas, **idénticas excepto por el color**. Si, como es de esperar, no tienes tantas bolitas de diferentes colores, alcanza con hacer bolitas de papel coloreado o cosas similares, aunque nosotros nos seguiremos refiriendo en nuestro ejemplo a las bolitas. Por lo tanto, el estado inicial de nuestro experimento (la jarra) es

	Azul	Roja
Cantidad	20	20
Proporción	$\frac{20}{40}$	$\frac{20}{40}$

Si sacas una bolita sin mirar el color, ¿qué esperas que ocurra? ¿Será roja o será azul? La extracción de una bolita de nuestra jarra se trata de un fenómeno aleatorio (no podemos saber el resultado con certeza), aunque en algunos casos podemos hacer conjeturas razonables. Por ejemplo, en nuestra jarra hay 20 bolitas azules y 20 rojas ($20 + 20 = 40$ bolitas en total), por lo que la proporción de azules es $p_{azul} = \frac{20}{40} = \frac{1}{2} = 0{,}5$ y lo mismo para las rojas $p_{roja} = \frac{20}{40} = \frac{1}{2} = 0{,}5$.[4] Por lo tanto, como elegimos al azar qué bolita tomamos y no existe nada (en principio) que lleve a tener preferencias por el color al elegirlas (recordar que las estamos tomando sin verlas) la probabilidad de tomar unas u otras será esencialmente su proporción. Como la proporción de azules es igual a la de rojas ($\frac{1}{2}$ en cada caso), lo único que podría conjeturar razonablemente es que hay tanta probabilidad de que aparezca una roja como una azul. Además, si **reponemos** la bolita elegida en la jarra la situación vuelve a ser como al principio (20 rojas y 20 azules), por lo que podemos repetir la lógica tantas veces como queramos.

Supongamos ahora que la bolita que sacamos antes fue azul y que **en lugar de reponerla la dejamos afuera del juego** (la puedes colocar en una caja). Antes de extraer la siguiente bolita el estado del sistema es

	Azul	Roja
Cantidad	19	20
Proporción	$\frac{19}{39}$	$\frac{20}{39}$

¿Cuál será la probabilidad, en una nueva extracción de la jarra, de sacar una bolita azul? ¿Será nuevamente equiprobable sacar una bolita roja que una azul? Para responder a estas preguntas debemos volver a hacer las cuentas. La probabilidad de sacar una bolita azul en la segunda extracción, **dado que la primera fue azul y que no la repusimos** es de $p_{azul} = \frac{19}{39} \sim 0{,}4872$, y la probabilidad de sacar una roja será por lo tanto $p_{roja} = \frac{20}{39} \sim 0{,}5128$ (notemos que ahora, en lugar de 40 bolitas, solo teníamos 39 antes de sacar la segunda). La diferencia de probabilidades respecto al experimento original es relativamente pequeña, pero aún así real: $p_{azul} = 0{,}5$ en la primera extracción, mientras que $p_{azul} \sim 0{,}4872$ en la segunda.

Supongamos que, una vez más, en la segunda extracción volvió a aparecer una bolita azul, algo nada extraño ya que existían aún en la jarra casi tantas bolitas rojas como azules. Nuevamente, la dejamos afuera de la jarra (la guardamos en la caja, junto a la primera). El estado actual de nuestro sistema es, por lo tanto:

	Azul	Roja
Cantidad	18	20
Proporción	$\frac{18}{38}$	$\frac{20}{38}$

Ahora, si volvemos a hacer las cuentas previas a una tercera extracción, tenemos que $p_{azul} = \frac{18}{38} \sim 0{,}47372$, y la probabilidad de sacar una roja será por lo tanto $p_{roja} = \frac{20}{38} \sim 0{,}5263$ (siempre la suma de $p_{azul} + p_{roja} = 1$). Claramente, si empezamos (por azar) a extraer más bolitas azules que rojas (o más rojas que azules) **y no reponemos**,

[4] Obviamente, habríamos llegado a lo mismo sabiendo que $p_{azul} + p_{roja} = 1 \therefore p_{roja} = 1 - p_{azul} = 1 - \frac{1}{2} = \frac{1}{2}$.

la probabilidad de volver a sacar una del mismo color que las que venían saliendo tiende a bajar.

Vayamos para atrás en el tiempo (afortunadamente podemos hacerlo porque se trata de un experimento conceptual). Supongamos en cambio que luego de haber sacado la primera bolita azul, la segunda hubiese sido roja y, por supuesto, no repusimos ninguna de las dos. En este escenario la probabilidad de sacar una bolita azul en la tercera extracción hubiese sido $p_{azul} = \frac{19}{38} = \frac{1}{2} = 0{,}5$, y la de sacar una roja hubiese sido también $p_{roja} = \frac{19}{38} = \frac{1}{2} = 0{,}5$. En otras palabras, como volvimos a igualar las proporciones respecto al estado inicial de la jarra (cuando había 40 bolitas), la probabilidad de extraer alguno de los dos colores hubiese sido la misma.

Hasta ahora habíamos parado nuestro experimento luego de la extracción de dos bolitas. Veamos un experimento real, pero sigámoslo hasta que no queden más bolitas en la jarra. Las extracciones han seguido el orden que aparece en la primera columna («bolita») del cuadro 3.1. Es fundamental entender que este es el resultado de **un** experimento en particular. Si repetimos el experimento el resultado (orden de salida de las bolitas) posiblemente sea muy distinto. Tenemos solo 39 filas y no 40, porque la última extracción ya no es aleatoria y porque además no tiene sentido calcular las proporciones luego de esa última extracción (ya no quedan más bolitas).

Tabla 3.1. La realización de un experimento en particular. Partimos de 20 bolitas azules y 20 bolitas rojas. En la columna «bolita» aparece qué bolita sacamos (sin reposición) y en las otras columnas cuántas nos quedan luego de esa extracción de cada color, así como su proporción

bolita	rojas	azules	p. roja	p. azul
azul	20	19	0,5128	0,4872
roja	19	19	0,5000	0,5000
roja	18	19	0,4865	0,5135
azul	18	18	0,5000	0,5000
azul	18	17	0,5143	0,4857
azul	18	16	0,5294	0,4706
roja	17	16	0,5152	0,4848
roja	16	16	0,5000	0,5000
azul	16	15	0,5161	0,4839
roja	15	15	0,5000	0,5000
roja	14	15	0,4828	0,5172
azul	14	14	0,5000	0,5000
roja	13	14	0,4815	0,5185
azul	13	13	0,5000	0,5000
roja	12	13	0,4800	0,5200
roja	11	13	0,4583	0,5417
azul	11	12	0,4783	0,5217
azul	11	11	0,5000	0,5000
roja	10	11	0,4762	0,5238
roja	9	11	0,4500	0,5500
roja	8	11	0,4211	0,5789
roja	7	11	0,3889	0,6111

azul	7	10	0,4118	0,5882
roja	6	10	0,3750	0,6250
roja	5	10	0,3333	0,6667
azul	5	9	0,3571	0,6429
azul	5	8	0,3846	0,6154
azul	5	7	0,4167	0,5833
azul	5	6	0,4545	0,5455
azul	5	5	0,5000	0,5000
roja	4	5	0,4444	0,5556
roja	3	5	0,3750	0,6250
azul	3	4	0,4286	0,5714
roja	2	4	0,3333	0,6667
azul	2	3	0,4000	0,6000
azul	2	2	0,5000	0,5000
azul	2	1	0,6667	0,3333
azul	2	0	1,0000	0,0000
roja	1	0	1,0000	0,0000

El resultado de graficar las proporciones de bolitas azules remanentes (probabilidades de extraer una bolita azul) podemos verlo en la figura 3.1. Claramente, empezando desde una proporción de $p_{azul} = \frac{1}{2}$, como tenemos muchas bolitas, al comienzo las variaciones en las proporciones son relativamente pequeñas en cada extracción. A medida que van quedando cada vez menos bolitas (porque se trata de un **muestreo sin reposición**), las oscilaciones son cada vez mayores hasta que llegamos a una de las dos **barreras absorbentes** que tiene nuestro experimento: las proporciones de 0 (0 %) y de 1 (100 %). **Se les denomina barreras absorbentes porque luego de llegar a estos valores no podremos escaparnos hacia otros** (insistiremos en este concepto en breve). Es decir, luego de que llegamos a la proporción de 0 bolitas azules (y, por lo tanto, una proporción de 0), no hay forma de incrementar esa proporción. Lo mismo para la proporción de 1, cuando todas las bolitas que quedan son azules: la proporción se mantendrá así hasta el final del experimento.

Claramente, el inicio del experimento es igual al del anterior (en los dos partimos de $p_{azul} = \frac{1}{2}$), y el mismo debe terminar en una proporción de bolitas azules de 0 o de 1, cuando no queden más bolitas azules o cuando todas las que queden sean azules (esa es la razón también por la que no ploteamos hasta después de la última extracción).

¿Qué ocurriría con nuestro experimento si hubiésemos tomado otro punto de partida (por ejemplo, distinto número de bolitas azules que rojas), o si hubiésemos comenzado desde un número mayor o menor de bolitas? Afortunadamente, es muy fácil realizar estos experimentos en una computadora y analizar los resultados, al menos para tener una primera impresión del comportamiento cualitativo del sistema. En la figura 3.2 se observa qué ocurre al variar el número de bolitas en nuestra jarra (columna de la izquierda: 20 bolitas, columna de la derecha: 120 bolitas) y variando la frecuencia (proporción) de bolitas azules al comienzo (fila superior: $p_{azul} = \frac{1}{2}$, fila inferior: $p_{azul} = \frac{3}{4}$).

A partir de la figura 3.2 podemos empezar a sacar algunas conclusiones de este experimento aleatorio que nos servirán de referencia en breve. La primera conclusión se relaciona con algo que ya habíamos visto: **a medida que quedan menos bolitas en la jarra, ya sea porque se comenzó con pocas o porque el experimento se acerca a su fin, las variaciones en**

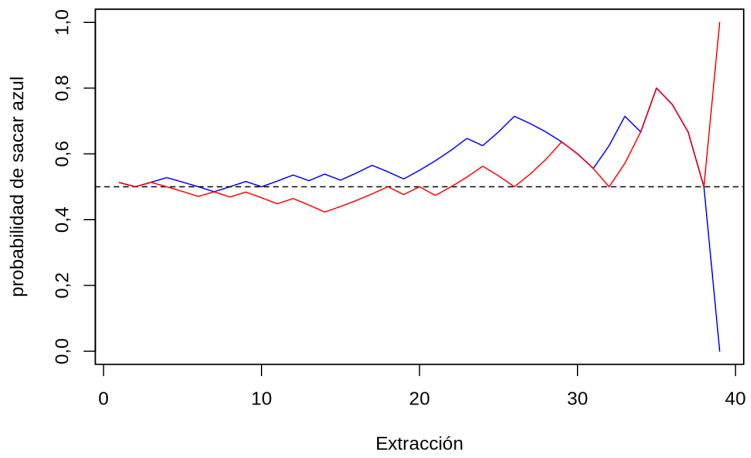


Figura 3.1. Probabilidad de sacar una bolita azul en la próxima extracción, para un nuevo experimento (en rojo), con las mismas condiciones iniciales.

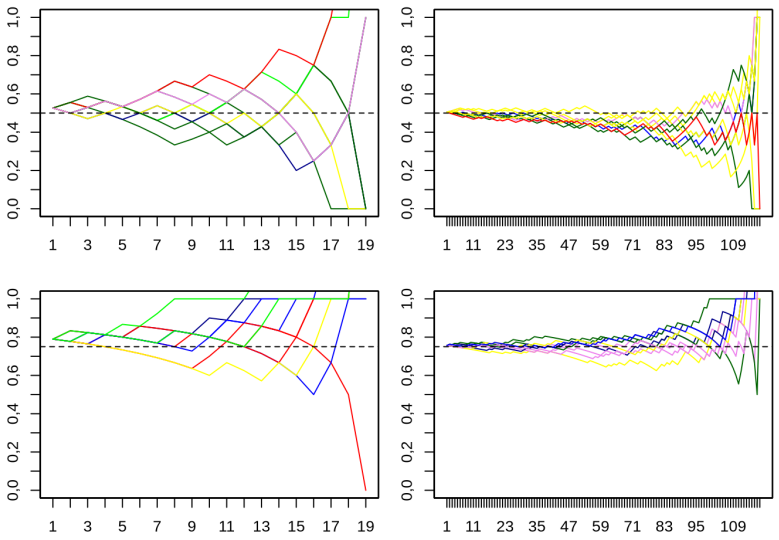


Figura 3.2. Probabilidad de sacar una bolita azul en la próxima extracción, para 10 experimentos, para 4 condiciones: p=1/2 (arriba) y p=3/4 abajo, en 20 (izquierda) y 120 (derecha) bolitas.

frecuencias suelen ser más drásticas. Esto se puede observar, por ejemplo, en la figura a la derecha: al principio las variaciones entre distintas réplicas del experimento apenas se apartan de la línea a trazos, pero a medida que quedan menos bolitas (aumenta el número de experimentos que ya realizamos) las proporciones en cada experimento empiezan a ser cada vez más variables (existe mayor dispersión en torno a la línea a trazos). La segunda conclusión refiere al punto de partida del experimento: **a medida que nos alejamos del punto medio (igual proporción de bolitas azules que rojas), mayor es la probabilidad de que la última bolita que quede en la jarra sea una bolita del tipo con mayor proporción inicial**. De hecho, la probabilidad de que la última bolita sea de uno de los dos colores, si realizamos repetidas veces el mismo experimento, es la probabilidad inicial del tipo al que pertenece dicha bolita. Podemos ver esto claramente a partir de un ejemplo. Supongamos que tenemos al inicio 10 bolitas en la jarra, 9 de ellas azules y una roja. Para que la roja sea la última en salir, todas las anteriores extracciones deben dar como resultado bolitas azules. La probabilidad de que la bolita roja sobreviva a la primera extracción (o sea, que la primera bolita sea azul) es $\frac{9}{10}$ (ya que hay 9 bolitas azules y solo una roja). La de la segunda, asumiendo que la primera fue azul y que por lo tanto la bolita roja sobrevivió a la primera tirada, es de $\frac{8}{9}$. La de la tercera será entonces $\frac{7}{8}$, y así sucesivamente hasta que queden dos bolitas: una roja y otra azul. En este caso, la probabilidad de elegir nuevamente una bolita azul es de $\frac{1}{2}$. Considerando estos eventos en conjunto (multiplicando sus respectivas probabilidades), tenemos

$$P(\text{última} = \text{roja}) = \frac{9}{10} \times \frac{8}{9} \times \frac{7}{8} \times ... \times \frac{1}{2} = \frac{\prod_1^9}{\prod_2^{10}} = \frac{9!}{10!} = \frac{1}{10} \tag{3.1}$$

que es la proporción de bolitas rojas al inicio.[5]

De hecho, podemos generalizar fácilmente el razonamiento anterior a otra proporción de bolitas rojas y azules. Supongamos ahora que tenemos como punto de partida 4 bolitas rojas y 6 azules. ¿Cuál es la probabilidad de que la última bolita sea de color rojo? Llamemos n al número total de bolitas ($n = 10$, en nuestro experimento) y k al número de bolitas rojas ($k = 4$, ídem). El total de posibles órdenes en que salen **todas** las bolitas es igual a las permutaciones de n elementos, dado por $\mathbb{P}(n) = n!$ Para que la última bolita sea roja, separamos una de ellas del conjunto, por lo que ahora solo tenemos $n - 1 = 9$ bolitas. Notemos que el orden de salida de las mismas es irrelevante, por lo que hay $\mathbb{P}(n-1) = (n-1)!$ posibilidades (órdenes de salida) de que esto suceda. Por lo tanto, **dado que tomamos una bolita roja en particular**, la probabilidad de que esa salga última será $\frac{\mathbb{P}(n-1)}{\mathbb{P}(n)} = \frac{(n-1)!}{n!} = \frac{\prod_1^{n-1}}{\prod_1^n} = \frac{1}{n}$
Sin embargo, hay k bolitas rojas **diferentes** y para cualquiera de ellas este razonamiento es válido, por lo que debemos multiplicar la probabilidad anterior por k (todos son eventos complementarios). Considerando esto, la probabilidad de que la última bolita sea roja es

$$\mathbb{P}(\text{última} = \text{roja}) = k\frac{\mathbb{P}(n-1)}{\mathbb{P}(n)} = k\frac{(n-1)!}{n!} = k\frac{\prod_1^{n-1}}{\prod_1^n} = \frac{k}{n} \tag{3.2}$$

[5] Recordemos que $\prod_a^b$ simboliza a la productoria de a a b. Esta operación es análoga a la sumatoria, pero en lugar de calcularse la suma de los números entre a y b, se considera el producto de estos números. A modo de ejemplo, $\prod_1^5 = 1 \times 2 \times 3 \times 4 \times 5$.

que es la probabilidad inicial de sacar una bolita roja (k/n, $4/10$ en nuestro caso) o, lo que es lo mismo, la proporción inicial de bolitas rojas. Claramente, este resultado está de acuerdo con nuestra intuición inicial, ya que en principio no hay nada especial en las bolitas rojas o en las azules y, por lo tanto, sería de esperar que cualquier resultado que sea con base en este agrupamiento (azules versus rojas) sea una función de la proporción de cada una de ellas. La utilidad de todo este razonamiento lo veremos en la siguiente sección («El modelo de Wright-Fisher»).

PARA RECORDAR

- La importancia de las **proporciones de partida** puesto que estas determinarán las probabilidades de obtener cierto resultado.
- Las **barreras absorbentes** serán aquellos valores que no podremos sobrepasar, y representarán los «límites» de los que no podremos escapar hacia otros valores.

3.2 El modelo de Wright-Fisher

A esta altura, ya habrás comenzado a imaginarte la relación entre los experimentos aleatorios de la sección anterior y nuestro modelo de poblaciones con un número relativamente pequeño de individuos. Si bien es posible imaginarse y modelar todo el proceso de gametogénesis, apareamientos al azar para constituir los genotipos de la próxima generación y repetir esto tantas veces como queramos, Sewall Wright[6] y Ronald Fisher[7] imaginaron un procedimiento mucho más sencillo que permite arribar a las mismas conclusiones con mucho menos esfuerzo de cálculo.

La primera suposición que hace el modelo de Wright-Fisher es que, más allá del número de individuos, el **pool de gametos** (el conjunto de todos los gametos en la población) es infinito. Si bien esto puede parecer extraño, es importante recordar que el número de espermatozoides (gametos masculinos) que producen los machos es virtualmente enorme, y el número de óvulos que produce cada hembra es de órdenes de magnitud superior que su descendencia; teniendo esto en cuenta, este no parece un supuesto arriesgado. Los otros supuestos incluyen: *i)* ausencia de selección, *ii)* ausencia de mutaciones, *iii)* ausencia de migración, *iv)* tiempos de generación no superpuestos y *v)* apareamiento aleatorio (panmixia). Si bien no suena realista que se cumplan todas estas condiciones en la vida real, el apartamiento de las mismas suele ser lo suficientemente menor como para considerar el modelo como una buena aproximación inicial. El proceso que plantean Wright-Fisher para el modelo de un gen con dos alelos es el siguiente: para una población de N individuos de una especie diploide, dada la frecuencia de uno de los alelos, se muestrean al azar $2N$ alelos del **pool infinito**. Esta será nuestra nueva población. La lógica del modelo Wright-Fisher se ve graficada para

[6] Sewall Green Wright (21 de diciembre de 1889 - 3 marzo de 1988), fue un genetista que realizó notables aportes sobre los efectos de la selección y la deriva, así como el análisis de pequeñas poblaciones y el efecto de la consanguinidad. Junto a Ronald A. Fisher y J. B. S. Haldane establecieron la fundación matemática de la genética de poblaciones y de la teoría evolutiva.

[7] Sir Ronald Aylmer Fisher (17 de febrero de 1890 – 29 de julio de 1962) fue un matemático, estadístico y biólogo evolutivo inglés. A partir de los datos colectados en la Estación Experimental de Rothamsted desarrolló las bases estadísticas para el diseño de experimentos; desarrolló también el análisis de varianza y fue unos de los tres pioneros (junto a Sewall Wright y John Burdon Sanderson Haldane) que participaron de la fundación de la teoría en genética de poblaciones, el neodarwinismo y la síntesis moderna de la evolución.

un caso con N individuos en la figura 3.3, donde además se hace explícita la analogía con el experimento mental de bolitas que desarrollamos anteriormente.

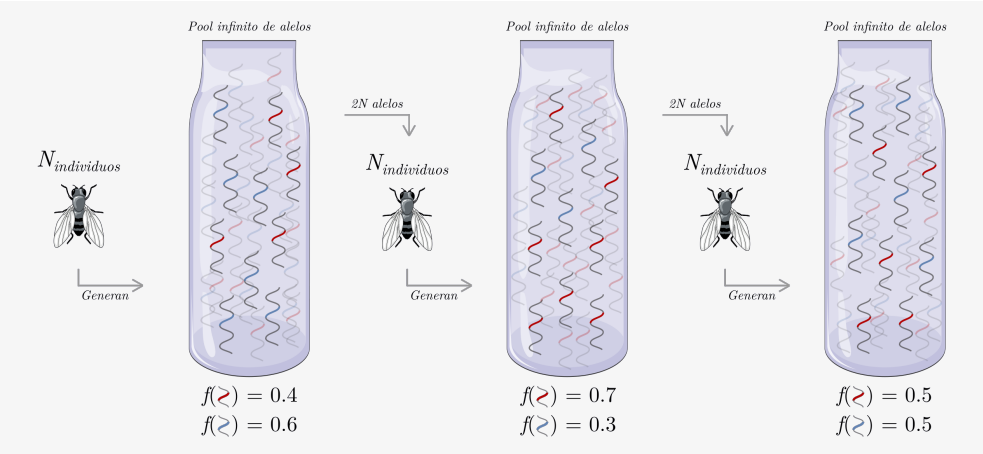


Figura 3.3. Representación esquemática del modelo de Wright-Fisher. Una población finita de N individuos (de la especie *D. melanogaster*, en este ejemplo) generan en cada generación un pool infinito de alelos que mantiene las frecuencias alélicas de la generación. La próxima generación se compone muestreando 2N alelos con reposición, los cuales conformarán a los N individuos de la generación. Imagen creada con elementos gráficos tomados de Bioicons (https://bioicons.com/).

Podemos repetir el procedimiento tantas veces como queramos para estudiar el comportamiento a lo largo del tiempo de nuestro modelo. Ahora, suponemos que ya te habrás preguntado cómo hacemos para muestrear de un pool infinito a partir de un número limitado de alelos disponibles ($2N$). El truco es muy sencillo, la diferencia entre un muestreo **con reposición** y un muestreo **sin reposición** (el que hicimos en la sección precedente) es que mientras que en este último las frecuencias van variando a medida que vamos extrayendo muestras, en el primero (**con reposición**) las frecuencias permanecen incambiadas definitivamente. Claramente, cuando el muestreo es de a una bolita (o alelo) esto se asemeja a muestrear de un pool infinito de bolitas (o alelos).

Veamos cómo funciona en la práctica. Supongamos que tenemos una población de $N = 5$ individuos. Supongamos además de que se trata de un locus con dos alelos, A representado por bolitas azules y a representado por las bolitas rojas. Para simplificar y que se fije la idea del procedimiento vamos a comenzar con una frecuencia inicial de $p_A = p_{azul} = \frac{1}{2} = p_{roja} = p_a$. Para que nuestro procedimiento funcione claramente vamos a tener ahora dos jarras (una inicialmente con igual proporción de bolitas azules y rojas y la otra vacía) y una caja con al menos $2N$ bolitas de cada color, o sea que en nuestro ejemplo, al menos 10 bolitas azules y 10 bolitas rojas.

La jarra vacía va a representar una nueva generación de nuestra población. Para completarla procederemos extrayendo al azar de la **jarra llena** una bolita y de acuerdo con su color sacamos **de la caja** una bolita del mismo color y la colocamos en la **jarra vacía**, devolviendo la bolita que sacamos a la jarra llena. Repetimos este procedimiento $2N$ veces, al cabo del cual tendremos $2N$ bolitas en nuestra antigua **jarra vacía** (la nueva generación de la población). Para completar el procedimiento y estar prontos para una nueva generación, vaciamos el contenido de la anterior **jarra llena** en la caja.

En este punto, podemos analizar el contenido de la **nueva jarra llena** (la población «hija» de la inicial), contando cuántas bolitas de cada color obtuvimos. Ahora, ¿podemos predecir (de alguna manera) el comportamiento de este experimento en una generación? Si llamamos (arbitrariamente) un «éxito» al hecho de sacar una bolita de color azul, entonces la extracción de cada bolita será una variable aleatoria con distribución de Bernoulli[8] y probabilidad de éxito p_{azul} (y, por lo tanto, probabilidad de fracaso, bolita roja, $p_{roja} = 1 - p$). Si consideramos el resultado de las $2N$ extracciones independientes y con la misma probabilidad (recordemos que es un muestreo **con reposición**) como un conjunto, entonces la distribución del número de bolitas azules (éxitos) y rojas (fracasos) entre las $2N$ tenga una distribución binomial:

$$P(i \text{ bolitas azules}) = \binom{2N}{i} p^i (1-p)^{2N-i},$$
$$\binom{2N}{i} = \frac{2N!}{i!(2N-i)!}$$

Es decir, si bien el número de bolitas en la nueva jarra también es de $2N = 10$, ahora el número de bolitas azules (copias del alelo A) que llamamos i (y, por lo tanto, el de bolitas rojas, o alelo **a** igual a $2N - i$) es una variable aleatoria y, por lo tanto, con resultado incierto de experimento en experimento. En general, si una población tiene i copias del alelo A y $2N - i$ copias del alelo a, la probabilidad de transición de pasar de i copias a j copias en el modelo de Wright-Fisher se encuentra dada por la siguiente ecuación

$$T_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j} = \frac{(2N)!}{j!(2N-j)!} p^j q^{2N-j} \tag{3.3}$$

Esta ecuación es bien sencilla de explicar. La probabilidad de éxito (alelo A o bolita azul) varía de generación en generación y, por lo tanto, para la generación actual de gametos es igual a la proporción de alelos A (i) en el total ($2N$), por lo que para cada generación de gametos $p = \frac{i}{2N}$ Concomitantemente, la probabilidad de fracaso (alelo **a** o bolita roja) es igual a la proporción de alelos **a** en el total, o sea $\frac{2N-i}{2N}$. Ahora, al número de éxitos **en la siguiente generación** lo llamamos j (la cantidad de alelos **A** que tienen la probabilidad T_{ij}) y, por lo tanto, la cantidad de alelos **a** será $2N - j$.

Hasta ahora nos hemos referido a una población en particular y como vimos su evolución es completamente al azar. En la figura 3.4 podemos ver un ejemplo de la evolución del alelo A en una población de 5 individuos diploides. El punto de partida es de $p_A = \frac{1}{2}$ y el aspecto de «dientes de sierra» se debe a que lo más fino que puede ser el movimiento es de $\frac{1}{2N} = \frac{1}{10} = 0{,}1$ en nuestro caso. Por otro lado, tanto $p_A = 0$ como $p_A = 1$ son **barreras absorbentes** ya que $p_A = 0$ quiere decir que no quedan más alelos A (bolitas azules) en la población, mientras que $p_A = 1$ quiere decir que todos los alelos que quedan en la población son A, es decir que este alelo se fijó en la población (y, por lo tanto, en ambos casos la frecuencia no variará más en el tiempo; recordar que uno de los supuestos fue que no hay mutación).

Afortunadamente a partir de unas pocas líneas de código podemos explorar a nuestro antojo el comportamiento en varias poblaciones, conjunto que se llama ***ensemble*** (del francés, el

[8] Jacob Bernoulli (6 de enero de 1655 - 16 de agosto de 1705) fue un matemático Suizo que realizó diversos descubrimientos importantes y uno de los promotores tempranos del cálculo integral. Miembro de las academias reales de París y Berlín, la distribución lleva su nombre en homenaje a sus contribuciones.

conjunto de todos los individuos de todas las poblaciones; véase figura 3.6), variando el número de individuos en ellas, así como la proporción inicial del alelo A. El resultado de probar el comportamiento para poblaciones de 5 y 50 individuos, con proporción inicial de alelo de $p = \frac{1}{2}$ y $p = \frac{3}{4}$ lo podemos ver en la figura 3.5. Claramente podemos apreciar que las poblaciones con mayor número de individuos tienen un comportamiento menos errático. Por otro lado, en las poblaciones que arrancan más cerca de una de las **barreras absorbentes** la variación es menor en relación con las poblaciones con el mismo número de individuos que arrancan desde la mitad ($p = \frac{1}{2}$).

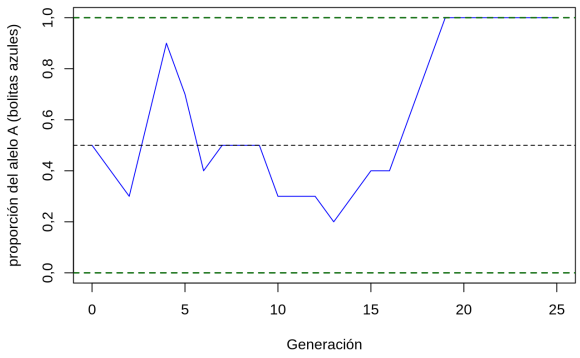


Figura 3.4. Proporción del alelo A (bolitas azules) en una población de 5 individuos diploides. Notemos que se trata de una realización en particular de este experimento y que con seguridad si repetimos el experimento el resultado será diferente.

Este comportamiento no debería resultarnos extraño. Los muestreos que estamos realizando en cada población, a cada generación, provienen de distribuciones binomiales. Si recuerdas de tus cursos de matemáticas y estadística, para una variable $i \sim Binom(2N,p)$ se tiene valor esperado $E(i) = 2Np$ y varianza $\sigma^2(i) = 2Np(1-p) = 2Np - 2Np^2$, por lo que esperamos que mientras las poblaciones se encuentren segregando, la esperanza (la media) entre la distintas poblaciones sea de Np o, lo que es lo mismo, la proporción del alelo A sea igual a $p_A = \frac{E(i)}{2N} = \frac{2Np}{2N} = p$ (es decir, en promedio las líneas de poblaciones oscilarán en torno a la frecuencia inicial). Más aún, como vimos antes en («H-W: la frecuencia de heterocigotas en función de la frecuencia alélica»), $p(1-p)$ tiene un máximo en $p = \frac{1}{2}$, por lo que la varianza $\sigma^2(i) = 2Np(1-p)$ también tendrá un máximo en $p = \frac{1}{2}$ (porque $p(1-p)$ está multiplicado por $2N$ que es necesariamente una constante positiva); en otras palabras, la varianza **dentro de cada población** será máxima cerca de la frecuencia $p = \frac{1}{2}$.

¿Cómo se condice esto con el comportamiento más variable en poblaciones pequeñas? Si recuerdas de esos mismos cursos, la varianza en la frecuencia por **generación** en el ensemble viene dada por

$$\begin{aligned} \sigma^2(p) = \sigma^2(i/(2N_e)) = \frac{1}{(2N_e)^2} \times \sigma^2(i) \therefore \\ \sigma^2(p) = \frac{2N_e p(1-p)}{4N_e^2} = \frac{p(1-p)}{2N_e} \Rightarrow \end{aligned} \tag{3.4}$$

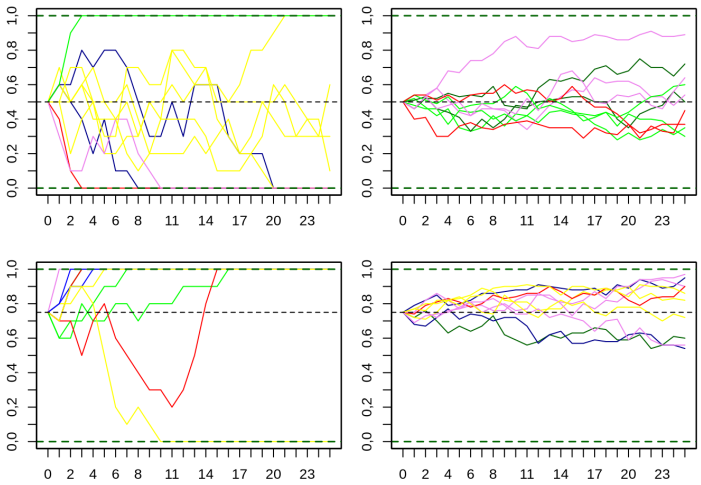


Figura 3.5. Frecuencia del alelo A (bolitas azules) en 10 poblaciones de 5 individuos diploides (a la izquierda) y de 50 individuos diploides (derecha), para frecuencias iniciales de p=1/2 (arriba) y p=3/4 (abajo).

$$\sigma^2(p) = \frac{pq}{2N_e} = \frac{p - p^2}{2N_e} \tag{3.5}$$

Es decir que al aumentar el tamaño de cada población ($2N$) la variación **entre poblaciones** disminuye. Finalmente, la variación **entre poblaciones** para las poblaciones que aún se encuentran segregando (o sea, que no llegaron a las **barreras absorbentes**) se incrementa con el tiempo de manera aproximadamente lineal, por lo que para tiempo t en las líneas que continúan segregando esperamos una varianza de

$$\sigma_t^2(p) = t\frac{pq}{2N} = t\frac{(p - p^2)}{2N}. \tag{3.6}$$

Una última observación es la que hace al número de individuos N de las distintas poblaciones. Hasta ahora asumimos que cada uno de nuestros individuos era completamente independiente de los otros desde el punto de vista genético (es decir, asumimos condiciones ideales) y, por lo tanto, no tuvimos en cuenta el efecto de que se puedan ver emparentadas en nuestros cálculos. En realidad, en la práctica, bajo diferentes criterios (apareamientos dirigidos, diferencias en el número de progenie esperada por pareja, etcétera), las poblaciones se suelen comportar como si el número de individuos fuese menor y a ese número se lo conoce como **tamaño efectivo** de la población y se lo representa con el símbolo N_e. Esto lo veremos con más detalle en la sección «Tamaño efectivo poblacional» y de nuevo en el capítulo «Apareamientos no-aleatorios».

En resumen, a partir del modelo sencillo de Wright-Fisher pudimos entender que **en ausencia de otras fuerzas evolutivas como la selección, migración o mutación, las frecuencias alélicas pueden variar en forma aleatoria y que con el transcurso del tiempo cada alelo puede fijarse en alguna población mientras que desaparece en otras.** Este fenómeno es lo que se conoce como **deriva genética** o **deriva aleatoria**, este último término en razón de que el cambio en las frecuencias no tiene una dirección y que, por lo tanto, no es predecible el sentido para una población en particular.

Ejemplo 3.1

Una población de 1000 individuos es dividida en cuatro subpoblaciones, cada una con un tamaño poblacional efectivo de 250 individuos. Las poblaciones se encuentran segregando un locus bialélico, siguiendo todos los supuestos esperados para encontrarse en equilibrio de Hardy-Weinberg (salvo el de tamaño poblacional infinito). Luego de un tiempo, se observan las siguientes frecuencias alélicas (p_i) para el alelo A_1 en el ensemble de subpoblaciones: $p_1 = 0{,}39;\ p_2 = 0{,}61;\ p_3 = 0{,}57;\ p_4 = 0{,}42$.

Estima el número de generaciones transcurrido desde la formación del ensemble de subpoblaciones.

A partir de las frecuencias alélicas observadas en el ensemble, se estima una frecuencia promedio $\bar{p} = \frac{0{,}39+0{,}61+0{,}57+0{,}42}{4} = 0{,}4975$. A su vez, la varianza observada para la frecuencia alélica en el ensemble es de $\sigma_t^2(p) = \frac{1}{n}\sum_{i=1}^{i=n}(p_i - \bar{p})^2 = \frac{(0{,}39-0{,}4975)^2+(0{,}61-0{,}4975)^2+(0{,}57-0{,}4975)^2+(0{,}42-0{,}4975)^2}{4} \Rightarrow \sigma_t^2(p) = 0{,}00886875$.

Dado que la varianza para la frecuencia alélica p aumenta linealmente con el tiempo generacional t según la ecuación $\sigma_t^2(p) = t \times \frac{pq}{2N_e}$, podemos usar nuestro estimado de $\bar{p}$ y el tamaño poblacional efectivo $N_e = 250$ para calcular el número de generaciones transcurrido desde la generación del ensemble:

$$\sigma_t^2(p) = t \times \frac{pq}{2N_e} \Rightarrow t = \frac{\sigma_t^2(p) \times 2N_e}{\bar{p}\bar{q}}$$

$$t = \frac{0{,}00886875 \times 2 \times 250}{0{,}4975 \times (1 - 0{,}4975)}$$

por lo que inferimos que se estima pasaron $t \approx 18$ generaciones desde la formación del ensemble poblacional.

PARA RECORDAR

- El modelo de **Wright-Fisher** asume que de un **pool infinito** de gametos, con las frecuencias de los alelos proporcionales a lo muestreado para constituir la generación anterior, muestreamos nuevamente $2N$ alelos, que constituirá a la nueva población.
- La suposición de un **pool infinito** de gametos es equivalente, desde el punto de vista matemático, a un muestreo con reposición.
- Al generarse la nueva población a partir de $2N$ muestreos independientes de Bernoulli, la distribución de la nueva población será una Binomial con probabilidad p y tamaño $2N$.
- La varianza de las frecuencias del alelo $A(p)$ entre poblaciones, **por generación**, debida al efecto de la deriva genética es

$$\sigma^2(p) = \frac{pq}{2N_e} = \frac{p - p^2}{2N_e}$$

siendo p y q (con $q = 1 - p$) las frecuencias de los alelos y N_e el tamaño efectivo de las poblaciones que forman el **ensemble**. Como la varianza crece linealmente con el tiempo (en generaciones), tenemos también que para el tiempo t para las líneas que continúan segregando:

$$\sigma_t^2(p) = t\frac{pq}{2N_e} = t\frac{(p - p^2)}{2N_e}$$

3.3 La subdivisión poblacional y la evolución de las frecuencias alélicas

La sección anterior marcó un cambio importante en nuestras perspectivas. Desde analizar el comportamiento de una o dos poblaciones, a partir de la sección anterior empezamos a apreciar que el azar puede jugar un rol muy importante en la evolución de las frecuencias alélicas en poblaciones relativamente pequeñas y nos empezamos a interesar en su visión en conjunto.

El debate acerca del rol de la deriva genética versus la selección en los procesos de especiación (la aparición de nuevas especies) ha sido acalorado (véase, por ejemplo, Charlesworth, Lande y Slatkin 1982). Aunque el debate continúa, pese a reconocer un rol no despreciable a la deriva genética, los actores principales de esta comedia que es la vida parecen ser la mutación (como generador de variabilidad) y la selección, a través de sus diversas formas de acción (algo que veremos en el capítulo «Selección natural»).

Pero veamos cómo podría estar funcionando la deriva genética en los procesos que llevan a la especiación. En efecto, imaginemos un locus con dos alelos que sean neutros respecto a los ambientes (por lo tanto, no hay selección). Asumamos también que la tasa mutacional es suficientemente baja como para que resulte improbable la aparición de los mismos alelos. Supongamos también que se trata de una especie con poca movilidad esperada, o sea la relación entre su área potencial de apareamiento y el rango geográfico de la especie (pueden ser especies animales, plantas, hongos, por ejemplo). Si la dispersión a largas distancias en

esta especie suele ser algo relativamente raro, por ejemplo, semillas llevadas por el viento, es posible imaginar que el ensemble de poblaciones está formado por pequeños **parches**, cada uno constituyendo una población diferente y que dada la baja movilidad de la especie no es probable que intercambien material genético con otros parches (ver representación en la figura 3.6).

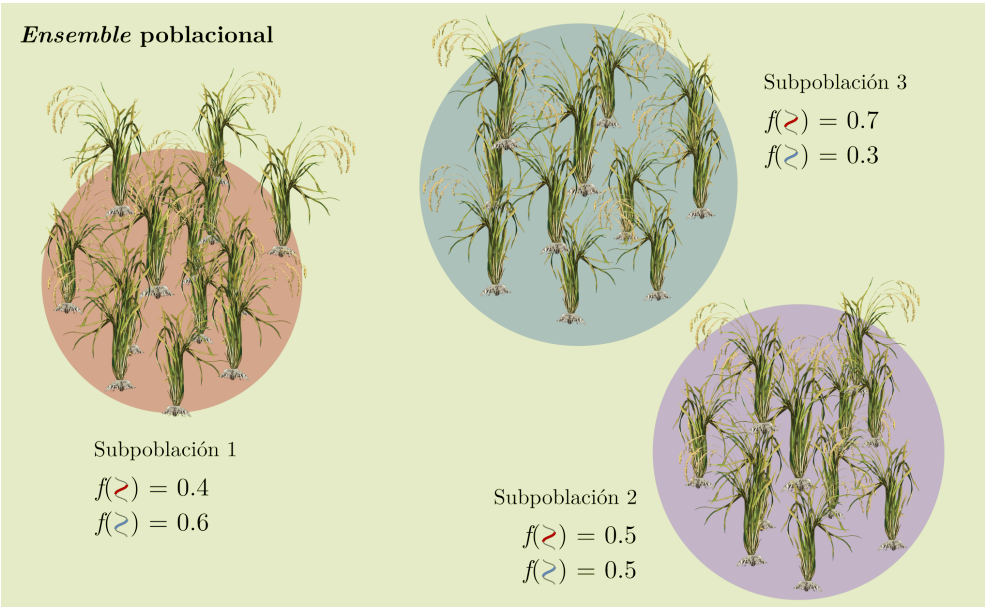


Figura 3.6. Ejemplo de ensemble poblacional. Tres poblaciones de arroz (*Oryza sativa*) se encuentran distribuidas en parches, a distancias tales que los individuos de cada subpoblación solo se reproducen entre sí. Por efecto de la deriva genética, es esperable que las frecuencias alélicas de las subpoblaciones del ensemble diverjan con el tiempo, lo cual se ejemplifica para un locus bialélico (alelos «rojo» y «azul», en la figura). Imagen creada con elementos gráficos tomados de bioicons (https://bioicons.com/).

Suponiendo además que el tamaño efectivo (N_e) es relativamente pequeño, algo muy razonable teniendo en cuenta que los individuos deben reproducirse en su entorno cercano, estamos entonces en una situación como alguna de las representadas en la figura 3.5. Es decir, mientras que en algunos parches la frecuencia del alelo A crecerá hacia la fijación de este, en otras poblaciones posiblemente se pierda (o lo que es lo mismo, quede fijado el alelo a). Pero la deriva genética es un fenómeno que ocurre simultáneamente en todo el genoma (obviamente que en sitios **no ligados** del genoma el fenómeno es independiente) y, por lo tanto, aun las poblaciones que terminen con el alelo A fijado serán diferentes en otros loci, donde la deriva ocurre en forma independiente. Todo este proceso, dado simplemente por la «reglas» de la reproducción, producirá un conjunto de poblaciones cada vez más disímiles entre ellas (y cada vez más parecidas dentro), lo que es un buen comienzo para la especiación.

Un fenómeno usual en muchas especies, que magnifica el rol de la deriva genética, es lo que se conoce como **cuello de botella poblacional** (o *genetic bottleneck*, en inglés). En efecto, en muchas especies es frecuente encontrarse con eventos que llevan a la casi extinción de alguna población en particular, en un momento dado del tiempo, luego de lo cual la población comienza a recuperarse e incrementar el número de individuos. Sin embargo, las huellas de

esta momentánea reducción drástica son normalmente imborrables en el genoma. Como en el momento de máxima restricción solo unos pocos individuos sobreviven (constituyendo la base reproductiva de la población a futuro), al ser N_e muy pequeño la varianza por generación será enorme, además de que la frecuencia inicial de las variantes será posiblemente **sesgada** respecto a la misma población antes del cuello de botella. Esto último nos lleva directamente a otro fenómeno estrechamente relacionado: el **efecto fundador.**

El **efecto fundador** (figura 3.7) se produce cuando un número relativamente pequeño de individuos **funda una nueva población** y por efecto de la frecuencia sesgada en esa pequeña muestra más el efecto de la endogamia (o *inbreeding*, el efecto del apareamiento entre parientes que veremos en otro capítulo más adelante) casi obligada por el número reducido de individuos, llevan a medida que la población va creciendo a una fuerte desviación de las frecuencias alélicas respecto a la población original.

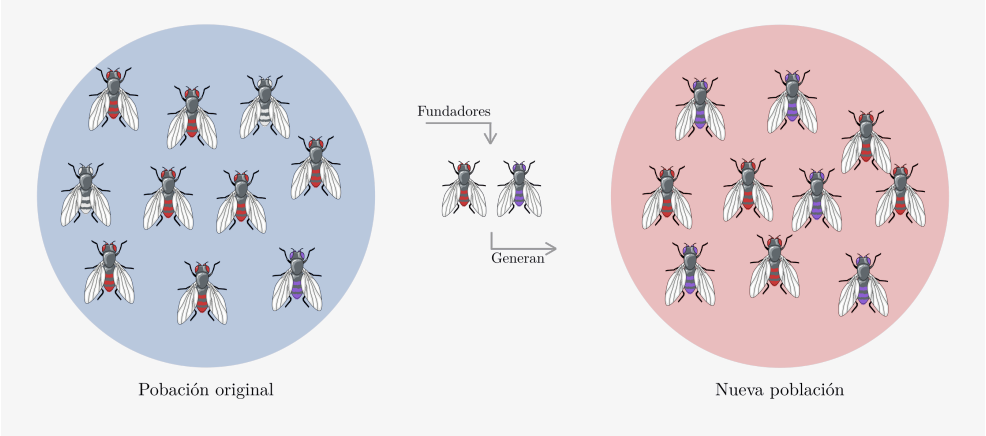


Figura 3.7. Efecto fundador ejemplificado para una población de la mosca de la fruta (*D. melanogaster*). Cuando una nueva población es formada a partir de una muestra pequeña de individuos, las frecuencias alélicas presentes entre los individuos fundadores tendrán un fuerte impacto en el devenir de las frecuencias alélicas de la nueva población. Debido al efecto de muestreo, las frecuencias alélicas entre los miembros fundadores pueden apartarse considerablemente de las presentes en la población original. En el ejemplo vemos una población original de moscas que presentan diferentes fenotipos para el color de ojo (rojo, blanco y violeta), los cuales se asumen se deben a las variantes alélicas que presenta cada individuo en un locus. La muestra de individuos fundadores se encuentra fuertemente sesgada (la componen en igual proporción individuos con ojos rojos y violetas), lo cual lleva a que la nueva población fundada sea radicalmente diferente a la original. Imagen creada con elementos gráficos tomados de bioicons (https://bioicons.com/).

Un ejemplo notable entre poblaciones humanas lo constituye el pueblo de **Cândido Godói**, un municipio de Rio Grande do Sul (Brasil) de unos 6 mil habitantes. El sobrenombre del pueblo es «terra dos gêmeos» (tierra de los gemelos), ya que 1 de cada 10 mujeres ha tenido mellizos/gemelos, una cifra muy superior al 1,8 % de Rio Grande do Sul (y en algún estudio, cerca de la mitad fueron monocigóticos) (figura 3.8).

Si bien se especuló que el origen de este fenómeno sería una serie de experimentos del doctor nazi Josef Mengele, la causa más probable es el estrecho parentesco de sus habitantes, unido al efecto fundador como origen de la colonia (Tagliani-Ribeiro et al. 2011). En particular, en este estudio los autores encuentran una fuerte asociación entre la presencia del

Figura 3.8. Mellizos/gemelos de Cândido Godói (imagen sacada de la web, sin crédito de autor).

alelo P72 en el gen TP53 y el «riesgo» de embarazo gemelar. También los autores reportan una significancia menor para la asociación entre el alelo T del gen MDM4 y en conjunto proponen que ambos alelos (en dos genes distintos) estarían actuando mediante la reducción del mecanismo de apoptosis inducida por p53.

Homocigosidad y heterocigosidad

En las secciones anteriores nos hemos manejado considerando exclusivamente la evolución en el tiempo de las frecuencias alélicas en las distintas poblaciones que conforman nuestro **ensemble** y dejamos de lado el análisis de lo que ocurriría dentro y entre ellas desde el punto de vista de los genotipos. Consideremos, como una suposición *a priori* razonable, que dentro de cada población se mantiene el apareamiento al azar y que más allá del número posiblemente pequeño de individuos dentro de ellas que hacen fluctuar las frecuencias alélicas, podemos esperar que estas se comporten **aproximadamente** según lo esperado en el equilibrio de Hardy-Weinberg de generación en generación. Para graficar nuestra situación supongamos que tenemos una gran población, con una frecuencia del alelo A igual a $p = \frac{1}{2} = 0{,}5$, que en el momento inicial de nuestra historia se divide en 4 poblaciones idénticas en tamaño (cada una de $1/4$ del tamaño de la grande), como lo representamos en la figura 3.9. Cada una de estas poblaciones, a medida que pase el tiempo, por efecto exclusivo de la deriva genética se irá apartando de la frecuencia original $p = 0{,}5$ y la varianza de frecuencias entre ellas irá creciendo en el tiempo, como vimos antes. Es de destacar que nuestra representación en la figura 3.9 es apenas una configuración posible de infinitas, ya que la evolución de las 4 poblaciones será aleatoria, pero esta configuración en particular, sin pérdida de generalidad, nos ayudará a ilustrar el fenómeno general.

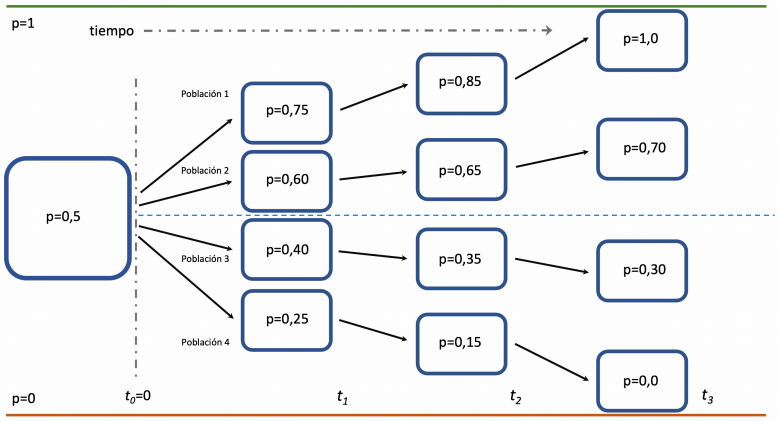


Figura 3.9. Evolución por deriva genética de las frecuencias alélicas en 4 subpoblaciones que parten de la misma frecuencia (el tiempo no está a escala). Las líneas superior e inferior representan las **barreras absorbentes** de fijación y pérdida del alelo. A medida que corre el tiempo las poblaciones se apartan de la frecuencia alélica inicial, aunque se mantiene la frecuencia promedio como el valor inicial.

Si pensamos en la población original, antes de la subdivisión y asumimos que se encontraba en equilibrio de Hardy-Weinberg, ya que $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$, tenemos que la composición de los tres genotipos era la siguiente:

$$AA = p^2 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$
$$Aa = 2pq = 2 \times \frac{1}{2}\frac{1}{2} = 2 \times \frac{1}{4} = \frac{1}{2}$$
$$aa = q^2 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

Luego, con el paso del tiempo, llegamos a un determinado momento t_1 en que las frecuencias alélicas ya han divergido algo. En la tabla 3.2 podemos ver lo que ocurre a nivel de los genotipos en cada una de las 4 poblaciones, así como la media de las 4 poblaciones (que como son todas iguales en tamaño es directamente igual al promedio de los valores). En la última columna tenemos también la frecuencia de los genotipos homocigotos, es decir AA y aa. Claramente, la frecuencia alélica promedio no ha cambiado, era de $\frac{1}{2}$ en $t_0 = 0$ y sigue siendo $\frac{1}{2}$ en t_1. Sin embargo, dentro de las poblaciones el cambio de frecuencia, sumado al equilibrio Hardy-Weinberg para cada una de ellas se refleja en diferencias importantes entre poblaciones en la proporción de homocigotos. Más aún, el promedio de los homocigotos (el promedio de la suma AA y aa en cada población) se vio incrementado, desde $fr(AA) + fr(aa) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} = 0{,}5$ a 0,5725. Este hecho no es menor: sin cambiar el promedio de las frecuencias alélicas y manteniendo el equilibrio de Hardy-Weinberg dentro de las poblaciones, si las volviésemos a agrupar al solo efecto de contar los genotipos, el

ensemble ya no se encuentra en el equilibrio de Hardy-Weinberg (porque el equilibrio para $p = 0{,}5$ es de $G = \frac{1}{2} = 0{,}5$).[9]

Tabla 3.2. Evolución de los genotipos y frecuencias en t_1

	p	AA	Aa	aa	Homocigotos
Población 1	0,75	0,5625	0,3750	0,0625	0,6250
Población 2	0,60	0,3600	0,4800	0,1600	0,5200
Población 3	0,40	0,1600	0,4800	0,3600	0,5200
Población 4	0,25	0,0625	0,3750	0,5625	0,6250
Media	0,50	0,2862	0,4275	0,2862	0,5725

Veamos qué ocurre entonces a medida que nos alejamos del punto inicial en el tiempo. Cuando llegamos al punto t_2, las frecuencias en las poblaciones divergen un poco más, aunque ninguna población ha llegado aún a fijar alguno de los dos alelos. La distribución de los genotipos en ellas, de acuerdo con las nuevas frecuencias alélicas, se puede apreciar en la tabla 3.3. Claramente volvemos a ver un incremento en promedio de homocigotos en la población, que ahora alcanzó el valor 0,6450.

Tabla 3.3. Evolución de los genotipos y frecuencias en t_2

	p	AA	Aa	aa	Homocigotos
Población 1	0,85	0,7225	0,255	0,0225	0,745
Población 2	0,65	0,4225	0,455	0,1225	0,545
Población 3	0,35	0,1225	0,455	0,4225	0,545
Población 4	0,15	0,0225	0,255	0,7225	0,745
Media	0,50	0,3225	0,355	0,3225	0,645

Para ser breves y confirmar la tendencia observada (además de comenzar a imaginar hacia dónde nos lleva esta), veamos qué ocurre en el instante t_3. Llegado este momento, en la *Población 1* se ha fijado el alelo A ($p = 1$), mientras que en la *Población 4* se ha fijado el alelo a (o lo que es equivalente, se ha perdido el A, $p = 0$). La distribución de genotipos se observa en la tabla 3.4.

Tabla 3.4. Evolución de los genotipos y frecuencias en t_3

	p	AA	Aa	aa	Homocigotos
Población 1	1,0	1,000	0,00	0,000	1,00
Población 2	0,7	0,490	0,42	0,090	0,58
Población 3	0,3	0,090	0,42	0,490	0,58
Población 4	0,0	0,000	0,00	1,000	1,00
Media	0,5	0,395	0,21	0,395	0,79

Ahora, además de haber aumentado nuevamente la proporción de homocigotos respecto al punto anterior en el tiempo, también tenemos un fenómeno nuevo: dos poblaciones

[9] Notemos que se utiliza el símbolo G para denotar al conjunto de genotipos homocigotos, como en la sección «Tres o más alelos» del capítulo anterior.

solo tienen (y tendrán a futuro) genotipos homocigotos, ya que se ha fijado uno de los dos alelos. Todo esto ocurre sin cambiar la frecuencia alélica media del ensemble. Más aún, si recordamos el comportamiento de las frecuencias alélicas en el tiempo, tarde o temprano las poblaciones terminan cayendo en alguna de las dos «trampas» que son las **barreras absorbentes**, los estados de fijación o pérdida (fijación del otro alelo), por lo que este comportamiento de la *Población 1* y *Población 4* es también esperable para las otras dos poblaciones (veremos este fenómeno en detalle más adelante). Llegado ese momento ya no quedarán heterocigotos en el ensemble, pese a que la frecuencia alélica se mantiene como al inicio.

Para entender mejor y poder cuantificar este fenómeno de la reducción del número de heterocigotos, debemos pasar a entender primero algunos conceptos sobre los que ahondaremos más adelante (véase capítulo «Apareamientos no-aleatorios»). Como vimos previamente, el concepto de alelo ha ido variando en el tiempo a medida que nuestro conocimiento sobre la biología molecular y la genética han llevado a entender las bases de la variación genética. Sin embargo, hay un aspecto en que las cosas no han cambiado mucho: el origen de la similaridad o diferencia entre alelos. En ausencia de mutación,[10] cuando un alelo es diferente de otro en la presente generación, nos resulta obvio de dónde viene la diferencia: dos alelos son diferentes en esta generación porque provienen de diferentes alelos en la generación previa, y hay poca distinción más que agregar. Sin embargo, cuando dos alelos son «iguales» (decimos que es el mismo alelo) en la presente generación, podemos ser más precisos sobre el origen de esta similitud: los dos alelos «iguales» en la presente generación pueden provenir de dos alelos «iguales» en la generación previa, pero que no sean copias del mismo ADN que los generó, o pueden ser copias del mismo ADN (gametos del mismo individuo y del mismo cromosoma), en cuyo caso los llamamos **idénticos por ascendencia**. Este concepto es fundamental, así que asegúrate de haberlo entendido. En ambos casos, cuando los dos alelos son «iguales» independientemente del origen, decimos que son **idénticos en estado**. Claramente, excepto por mutaciones recurrentes en el tiempo, si nos retrotraemos en el pasado lo suficiente, todos los alelos **idénticos en estado** deberían venir del mismo alelo original y, por lo tanto, serían de alguna manera **idénticos por ascendencia**, por lo que el marco de referencia es fundamental para decidir a partir de cuándo se comienza a hacer esta distinción.

La probabilidad de que **los dos alelos en un individuo** sean **idénticos por ascendencia** se simboliza comúnmente con la letra F, a partir de Sewall Wright (1922), al que llamó **coeficiente de fijación**. Nosotros vamos a utilizar un subíndice para denotar el tiempo desde la generación 0, que consideramos la generación inicial. Supongamos que queremos calcular la probabilidad de que dos alelos sean **idénticos por ascendencia** en la generación t, es decir F_t.

Como se ilustra en la figura 3.10, existen dos posibilidades mutuamente excluyentes para que dos alelos sean **idénticos por ascendencia** en la generación t: *a)* que ambos provengan del mismo alelo en la generación $t-1$, un evento cuya probabilidad es $\frac{1}{2N}$ (ya que hay $2N$ alelos en la población y, por lo tanto, la probabilidad de volver a muestrear el mismo alelo es $\frac{1}{2N}$) y en ese caso $F=1$, ya que al venir del mismo alelo este es **idéntico por ascendencia** a sí mismo con probabilidad de 1, y *b)* que los alelos provengan de distintas copias en la generación $t-1$, y en ese caso la probabilidad de que sean **idénticos por ascendencia** es

[10] La persona interesada en el modelado e implicancias de la relación entre mutación y deriva genética puede referirse a la bibliografía de referencia, donde suele tratarse dicho tema. A lo largo de este capítulo se deja este aspecto de lado, a efectos de simplificar el análisis y hacer foco en el fenómeno de la deriva genética.

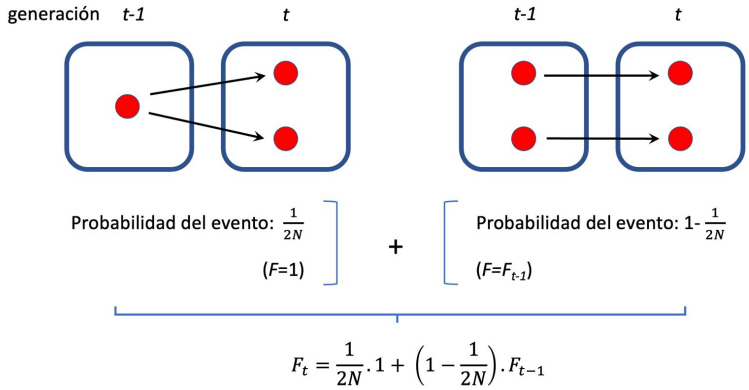


Figura 3.10. Ilustración del razonamiento detrás de la recursión para F en una población de tamaño finito. Existen dos alternativas disjuntas (de ahí la suma) para que un alelo sea idéntico por ascendencia en la generación t: que los alelos provengan del mismo alelo en la generación *t-1*, cuya probabilidad es $\frac{1}{2N}$ (y en ese caso $F = 1$, por definición) o que los alelos provengan de distintas copias en la generación *t-1*, y en cuyo caso la probabilidad de que sean idénticos por ascendencia es por definición F_{t-1}. Por lo tanto, la probabilidad de que dos alelos sean idénticos por ascendencia en la generación t, considerando a la generación $t-1$, es de $F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}$. Aplicando este razonamiento a las sucesivas generaciones anteriores se puede obtener, por recursión, la probabilidad de que dos alelos sean idénticos por ascendencia a tiempo t (tomando como referencia la población a tiempo $t = 0$).

por definición F_{t-1}. Como se trata de dos eventos disjuntos (puede pasar uno u otro, pero no los dos a la vez), podemos sumar sus probabilidades para tener

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1} \tag{3.7}$$

Se puede calcular así la probabilidad de dos alelos sean **idénticos por ascendencia** en una generación, asumiendo que se conoce dicha probabilidad para la generación anterior. Veamos ahora si podemos llegar a una expresión general para la probabilidad de que dos alelos sean **idénticos por ascendencia** en una generación en función del número de generaciones transcurridas desde un momento de referencia ($t = 0$) y la probabilidad inicial F_0 en dicho momento. Si a ambos lados de la ecuación (3.7) multiplicamos por -1 y le sumamos 1[11] tenemos

$$1 - F_t = 1 - \frac{1}{2N} - (1 - \frac{1}{2N})F_{t-1} 1 - F_t = (1 - \frac{1}{2N})(1 - F_{t-1}) \tag{3.8}$$

[11] Esta operación matemática puede resultar arbitraria. ¿Por qué haríamos esto? ¿Existe algún razonamiento biológico que dé sustento a la maniobra? Notemos que $(1 - F_t)$ es la probabilidad de que dos alelos *no* sean idénticos por ascendencia en el tiempo t. Por lo tanto, la utilización de esta maniobra matemática simplifica los cálculos, pero además puede ser vista como un intento de resolver el problema planteándolo desde una óptica biológica «complementaria». Si resolvemos qué sucede con la probabilidad de que dos alelos *no* sean idénticos por ascendencia, automáticamente resolvemos qué sucede con la probabilidad de que sí lo sean, por ser estos eventos disjuntos. La relación entre F y la desviación de la heterocigosidad esperada es analizada más adelante en este capítulo.

Pero con la misma lógica,

$$1 - F_{t-1} = \left(1 - \frac{1}{2N}\right)(1 - F_{t-2}) \tag{3.9}$$

y sustituyendo (3.9) en (3.8), tenemos

$$1 - F_t = (1 - \frac{1}{2N})(1 - \frac{1}{2N})(1 - F_{t-2}) = (1 - \frac{1}{2N})^2(1 - F_{t-2}) \tag{3.10}$$

Prosiguiendo con la recursión hasta llegar a la generación 0 (generación inicial), como transcurren t generaciones entre la 0 y la actual, tenemos que la ecuación (3.10) se generaliza a

$$1 - F_t = \left(1 - \frac{1}{2N}\right)^t (1 - F_0) \tag{3.11}$$

Más aún, si asignamos arbitrariamente $F_0 = 0$ por falta de conocimiento acerca de la probabilidad de que dos alelos sean idénticos por ascendencia en la generación inicial, entonces la ecuación (3.11) se transforma en

$$1 - F_t = \left(1 - \frac{1}{2N}\right)^t \tag{3.12}$$

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t \tag{3.13}$$

Veamos ahora cómo se relaciona todo esto con el problema de la subdivisión poblacional planteado en la figura 3.9. Por definición, F es también la reducción en heterocigosis (o el incremento en homocigosis) respecto a la expectativa en equilibrio de Hardy-Weinberg (y sus frecuencias genotípicas) en poblaciones finitas (porque es la probabilidad de que dos alelos sean idénticos por ascendencia en un individuo, por lo tanto, homocigoto). Entonces, el paralelismo de $1 - F$ es respecto a los heterocigotos conservados bajo apareamiento aleatorio. Si consideramos que la frecuencia esperada de heterocigotas en una población de tamaño finito es una función de p y q, pero también de esa probabilidad de que los alelos no sean idénticos por ascendencia, entonces si llamamos H_t a esa frecuencia, tenemos,

$$H_t = 2pq(1 - F_{t-1})$$
$$\therefore 1 - F_{t-1} = \frac{H_t}{2pq}$$

El $t-1$ en el término $(1-F_{t-1})$ se corresponde con el hecho de que el efecto de la reducción de los pares de alelos ***no*** **idénticos por ascendencia** se ve en la generación anterior a la

presente. Notemos que $H_t = 2pq(1 - F_{t-1})$, lo cual equivale a decir $1 - F_{t-1} = \frac{H}{2pq}$. Esto nos lleva a

$$1 - F_t = \left(1 - \frac{1}{2N}\right)(1 - F_{t-1}) = \left(1 - \frac{1}{2N}\right)\frac{H_t}{2pq} =$$
$$= \left(1 - \frac{1}{2N}\right)^2 \frac{H_{t-1}}{2pq} = \left(1 - \frac{1}{2N}\right)^3 \frac{H_{t-2}}{2pq}$$
$$\therefore \left(1 - \frac{1}{2N}\right)\frac{H_t}{2pq} = \left(1 - \frac{1}{2N}\right)^{t+1} \frac{H_{t-t}}{2pq}$$
$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 = \left(1 - \frac{1}{2N}\right)^t [2pq(1 - F_{-1})]$$

Notemos que en el desarrollo anterior se continuó la recursión planteada hasta alcanzar el tiempo $t = t$ (de ahí que el exponente que acompaña a $H_{t-t} = H_0$ sea $t+1$). Luego, como en ambos lados de la ecuación se tiene $2pq$ en el denominador, los términos se cancelan, y como en el lado izquierdo se tiene $1 - \frac{1}{2N}$, al dividir ambos lados por este término el exponente $t + 1$ en el lado derecho pasa a ser t.

En resumen, luego de este desarrollo por recursión se llega a

$$H_t = (1 - \frac{1}{2N})^t H_0$$

Utilizando la aproximación $(1 - a)^b \approx (e^{-a})^b \approx e^{-ab}$ cuando $a \to 0$ (en nuestro caso, $a = \frac{1}{2N};\ b = t$), tenemos

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 \approx H_0 e^{\frac{-t}{2N}} \tag{3.14}$$

Es decir, pese a que en cada una de las poblaciones se mantiene el equilibrio de Hardy-Weinberg, como se puede apreciar en la figura 3.11, la proporción de individuos heterocigotos en el conjunto de ellas, o mejor dicho el promedio en el **ensemble**, decaerá de forma geométrica respecto al valor inicial a lo largo del tiempo, ya que la proporción en cada generación es multiplicada por el factor $\left(1 - \frac{1}{2N}\right)$ al pasar a la siguiente. La aproximación exponencial funciona en general razonablemente bien, y permite resolver algunos problemas de forma más sencilla. El efecto de la reducción en la proporción de heterocigotos es realmente drástico cuando los tamaños poblacionales son extremadamente pequeños, por ejemplo, $N = 5$ o $N = 10$ individuos en la figura 3.11, pero son apenas notables cuando el número de individuos es 50 o más (notemos que para 100 individuos el punto de reducción a 50 % de la heterocigosis inicial ni siquiera aparece en la gráfica, ya que llevaría más de 138 generaciones para alcanzarlo).

Ejemplo 3.2

La ecuación (3.14), en particular la aproximación exponencial, nos permite estimar fácilmente los tiempos hasta determinado nivel de reducción de la frecuencia inicial de heterocigotos.

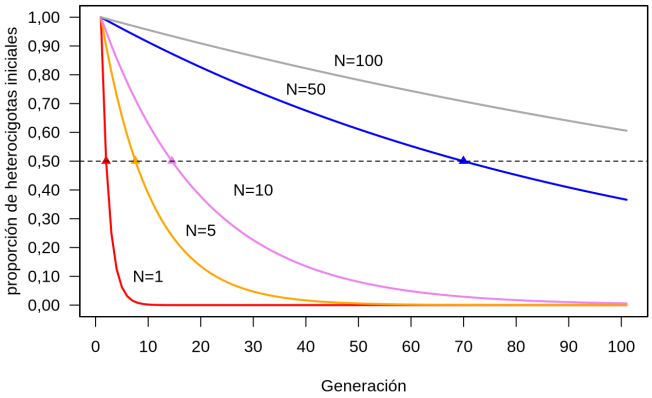


Figura 3.11. Reducción esperada en el número inicial de heterocigotas H_0 de acuerdo con el tamaño N de las poblaciones y al número de generaciones transcurridas. La línea negra horizontal marca la reducción a la mitad del valor inicial y los puntos sobre las curvas resaltan el número de generaciones requeridas para esa reducción a la mitad.

Por ejemplo, para alcanzar la mitad del valor inicial (H_0), con poblaciones de N individuos podemos hacer $H_t = \frac{H_0}{2}$. Luego, de acuerdo con la ecuación (3.14) tenemos que

$$H_t = \frac{H_0}{2} = H_0 e^{\frac{-t}{2N}}$$
$$\therefore \frac{1}{2} = e^{\frac{-t}{2N}}$$

Aplicando el logaritmo a ambos lados de la ecuación, tenemos

$$\ln\left(\frac{1}{2}\right) = \ln\left(e^{\frac{-t}{2N}}\right) = \frac{-t}{2N}$$

y finalmente

$$t = -2N \ln\left(\frac{1}{2}\right) \approx 1{,}39N$$

Es decir, en $1{,}39N$ generaciones. Si lo hubiésemos hecho sin la aproximación, el procedimiento es análogo y llegaríamos a

$$\frac{H_0}{2} = \left(1 - \frac{1}{2N}\right)^t H_0 \therefore t = \frac{\ln\left(\frac{1}{2}\right)}{\ln\left(1 - \frac{1}{2N}\right)}$$

que, si bien es la solución exacta, no es lineal en el número de individuos en cada población.

Resumiendo, en esta sección vimos cómo la subdivisión poblacional produce el efecto de disminuir el número de heterocigotas respecto a los esperado para las poblaciones del **ensemble** agregadas en un solo grupo y bajo condiciones de **panmixia** (es decir, con libertad de aparearse con cualquier otro individuo de la población, sin restricciones genéticas, conductuales o ambientales). Esa disminución de la proporción de heterocigotos esperados ocurre a una tasa constante por generación (en nuestro modelo) y es de $\left(1-\frac{1}{2N}\right)$, por lo que la reducción total a lo largo de t generaciones será del orden de $\left(1-\frac{1}{2N}\right)^t$. Mientras que en la sección «El modelo de Wright-Fisher» construimos un modelo de evolución de las frecuencias alélicas sin preocuparnos de los genotipos, en esta sección vimos las implicancias para la evolución de los genotipos que tiene dicho modelo. En la sección «Cadenas de Markov» veremos de otra forma una conexión entre los dos fenómenos utilizando una aproximación fundamental para comprender los procesos estocásticos en general.

PARA RECORDAR

- La recursión para el **coeficiente de fijación** F en la generación t, F_t es igual a

$$1-F_t=\left(1-\frac{1}{2N}\right)^t(1-F_0)$$

- Cuando $F_0=0$, es decir que asumimos que en la **generación inicial** la probabilidad de que dos alelos en un locus de un individuo diploide es cero, entonces la recursión anterior se simplifica a:

$$F_t=1-\left(1-\frac{1}{2N}\right)^t$$

- El número de heterocigotos en un **ensemble** de poblaciones de tamaño finito N se ve reducido de generación en generación respecto a lo esperado para el equilibrio de Hardy-Weinberg. Esa reducción es a la tasa de $\left(1-\frac{1}{2N}\right)$ por generación, por lo que la frecuencia de heterocigotos en la generación t se calcula de acuerdo con la siguiente ecuación:

$$H_t=\left(1-\frac{1}{2N}\right)^t H_0\approx H_0 e^{\frac{-t}{2N}}$$

3.4 Cadenas de Markov

Si recuerdas la ecuación (3.3), establecía la probabilidad de transición de un número i de copias a un número j (en un total de $2N$ alelos) en una generación, y esta era igual a

$$T_{ij}=\binom{2N}{j}\left(\frac{i}{2N}\right)^j\left(\frac{2N-i}{2N}\right)^{2N-j}=\frac{(2N)!}{j!(2N-j)!}p^j q^{2N-j} \tag{3.15}$$

Visto de otra forma, como i puede ir desde 0 hasta $2N$ y lo mismo para j, podemos poner estas probabilidades de transición en una matriz cuadrada con $2N+1$ filas y $2N+1$ columnas (el $+1$ es debido a que un conteo de 0 para cada alelo es posible). Por ejemplo, calculemos la matriz para una población con 2 individuos diploides. Para ello vamos a calcular de forma

explícita las probabilidades de transición de unas pocas celdas de la matriz y el resto sigue la misma lógica. En particular, vamos a calcular algunas celdas de la fila correspondientes a 2 alelos A en la generación actual (o sea, $p = \frac{2}{4} = \frac{1}{2} = 0{,}5$). Estas celdas corresponden a la probabilidad de pasar de un número $i = 2$ de alelos A a un número j determinado en la próxima generación.

De acuerdo con la ecuación (3.15), la probabilidad de transición de 2 copias del alelo A ($i = 2$) a 1 copia del mismo ($j = 1$) es de

$$T_{2,1} = \binom{4}{1}\left(\frac{2}{4}\right)^1\left(\frac{4-2}{4}\right)^{4-1} = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = 4\left(\frac{1}{2}\right)^4 = 0{,}25 \qquad (3.16)$$

Asimismo, la probabilidad de transición de 2 copias del alelo A ($i = 2$) a 3 copias del mismo ($j = 3$) es de

$$T_{2,3} = \binom{4}{3}\left(\frac{2}{4}\right)^3\left(\frac{4-2}{4}\right)^{4-3} = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 = 4\left(\frac{1}{2}\right)^4 = 0{,}25 \qquad (3.17)$$

De la misma manera, la probabilidad de transición de 2 copias del alelo A ($i = 2$) a **4** copias del mismo ($j = 4$) es de

$$T_{2,4} = \binom{4}{4}\left(\frac{2}{4}\right)^4\left(\frac{4-2}{4}\right)^{4-4} = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = 1\left(\frac{1}{2}\right)^4 = 0{,}0625 \qquad (3.18)$$

En la tabla 3.5 puedes ver el resultado de calcular las probabilidades de transición para todos los pares de estados posibles (pares entrada-salida, o generación actual-próxima generación $0 \leq i \leq 2N$, $0 \leq j \leq 2N$). Llamaremos a esta matriz T, por ser la matriz de probabilidad de **transiciones** de estado. En las filas tenemos los posibles estados en la generación actual, definidos como el número de alelos A en la generación actual, mientras que en las columnas tenemos el número de alelos A en la generación próxima. Se trata de probabilidades, ya que lo visto hasta ahora es que el apareamiento de acuerdo con el modelo de Wright-Fisher generará poblaciones descendientes de forma aleatoria, con una probabilidad dada por esta matriz de transición. Es de notar que los estados actuales de **0** y **4** alelos A se corresponden con lo que definimos como **barreras absorbentes** de nuestro sistema, pues una vez alcanzado alguno de estos dos estados será imposible abandonarlos (estamos asumiendo la ausencia de mutación). Dicho de otra forma, una vez que se pierde el alelo A en una población (0 copias del mismo), o que el mismo queda fijado en la población ($2N$ copias, en nuestro ejemplo $2N = 4$ copias), en las próximas generaciones esto no va a cambiar.

Tabla 3.5. Matriz de transición para un locus con dos alelos y dos individuos diploides

	a 0 alelos A	a 1 alelos A	a 2 alelos A	a 3 alelos A	a 4 alelos A
de 0 alelos A	1,0000	0,0000	0,0000	0,0000	0,0000
de 1 alelos A	0,3164	0,4219	0,2109	0,0469	0,0039
de 2 alelos A	0,0625	0,2500	0,3750	0,2500	0,0625
de 3 alelos A	0,0039	0,0469	0,2109	0,4219	0,3164
de 4 alelos A	0,0000	0,0000	0,0000	0,0000	1,0000

¡Excelente! Tenemos computada nuestra matriz de transiciones. Si observas con cuidado verás que la suma de cada fila es igual a 1, es decir $\sum_{j=0}^{2N} p_{i,j} = 1$. Esto quiere decir que las poblaciones que parten de i alelos en la generación padre deben distribuirse entre todas las posibilidades j, pero que en última instancia, al tratarse de un probabilidad, la suma de todas las posibilidades disjuntas debe ser igual a 1. Veamos ahora cómo la podemos usar para calcular la evolución de las frecuencias alélicas en un **ensemble** de poblaciones, todas del mismo tamaño y que parten de la misma frecuencia alélica. Supongamos que todas estas poblaciones parten en tiempo $t = 0$ de $i = 2$ alelos **A**. Representaremos las frecuencias de poblaciones en las que hay i alelos con un vector de largo $2N + 1$; la primera posición en el vector representará la frecuencia de poblaciones con 0 alelos **A**, la segunda la frecuencia de poblaciones con 1 alelo **A**, la tercera con 2 alelos **AA**, y así hasta la posición $2n + 1 = 5$ en nuestro caso. Como todas las poblaciones arrancan desde $i = 2$, nuestro vector de frecuencias será $\mathbf{I}_0 = (0; 0; 1; 0; 0)$. Ahora, si multiplicamos este **vector fila** $\mathbf{I}_0$ por nuestra matriz $\mathbf{T}$ de transiciones (si precisas recordar cómo se multiplican matrices y vectores, recurre al [Apéndice A: Conceptos Matemáticos Básicos]), obtendremos un nuevo vector de probabilidad de que las distintas poblaciones del **ensemble** posean un número determinado de alelos **A**, dado por $\mathbf{I}_1 = \mathbf{I}_0 * \mathbf{T}$ (representado en la tabla 3.6).

Tabla 3.6. Probabilidades de obtener un número determinado de alelos luego del primer ciclo

0 alelos A	1 alelos A	2 alelos A	3 alelos A	4 alelos A
0,0625	0,25	0,375	0,25	0,0625

Es decir, si bien partimos de que todas las poblaciones poseían 2 alelos A, por efecto del azar (**deriva genética** le llamamos a este efecto), ahora un 6,25 % de ellas poseerán 0 alelos A (se habrá perdido el mismo), un 25 % 1 alelo A, un 37,5 % 2 alelos A, otro 25 % 3 alelos A y finalmente, 6,25 % de ellas tendrán todos los 4 alelos A (habrán fijado ese alelo). Ahora, la misma matriz de transición que aplicamos para pasar de la generación $t = 0 \rightarrow t = 1$ la podemos aplicar al vector de probabilidades (frecuencias relativas) de poblaciones en la generación $t = 1$ para pasar a la $t = 2$.[12] Es decir que ahora tenemos $\mathbf{I}_2 = \mathbf{I}_1 \times \mathbf{T}$, cuyo resultado podemos apreciar en la tabla 3.7.

Tabla 3.7. Probabilidades de obtener un número determinado de alelos luego del segundo ciclo

0 alelos A	1 alelos A	2 alelos A	3 alelos A	4 alelos A
0,166	0,2109	0,2461	0,2109	0,166

Podemos repetir este procedimiento de generación en generación. Cuando un proceso estocástico depende solamente del estado inmediatamente anterior en el tiempo, como en nuestro caso, decimos que se trata de un proceso Markoviano. Una cadena de Markov[13] es una secuencia de eventos (usualmente en el tiempo) en la que cada resultado depende

[12] Notemos que las probabilidades de transición entre estados no cambian con el paso del tiempo; formalmente decimos que se trata de una cadena de Markov *homogénea* o *estacionaria* en el tiempo. En el siguiente párrafo se ahonda en el concepto de *cadena de Markov*.

[13] Andrey Andreyevich Markov (14 de junio de 1856 – 20 de julio de 1922) fue un matemático ruso, ampliamente conocido por sus trabajos en procesos estocásticos.

exclusivamente del resultado precedente. Dicho de otra manera, para predecir el siguiente resultado del proceso (futuro) solo la información del presente estado es relevante y el pasado (los estados anteriores del sistema) no tienen relevancia. A esto último se lo conoce como **propiedad markoviana** (o de **falta de memoria**). En nuestro caso, como la distribución de probabilidades de cada nueva generación $t = k$ es determinada multiplicando el resultado de la anterior ($\mathbf{I}_{t=(k-1)}$) por la matriz de transición ($\mathbf{T}$), claramente podemos ver por qué se trata de una **cadena** (de Markov). Si comparamos la evolución de nuestro sistema, tenemos ahora que desde la generación $t = 0$ hasta la generación $t = 5$, las frecuencias han ido cambiando de la manera que se aprecia en la tabla 3.8.

Tabla 3.8. Evolución de la distribución del número de alelos (probabilidades) desde la generación inicial (t_0) hasta la generación 5 (t_5)

t	0 alelos A	1 alelos A	2 alelos A	3 alelos A	4 alelos A
0	0,0000	0,0000	1,0000	0,0000	0,0000
1	0,0625	0,2500	0,3750	0,2500	0,0625
2	0,1660	0,2109	0,2461	0,2109	0,1660
3	0,2490	0,1604	0,1813	0,1604	0,2490
4	0,3117	0,1205	0,1356	0,1205	0,3117
5	0,3587	0,0904	0,1017	0,0904	0,3587

Otra forma de verlo, más gráfica, aparece en la figura 3.12. Claramente se aprecia que mientras que las probabilidades de encontrarse en los estados intermedios ($i = 1$ a $i = 3$, en gris) decrece, la de los estados correspondientes a las **barreras absorbentes** (rojo: pérdida, azul: fijación) crece en la misma medida. La pregunta sería ¿hacia dónde va este movimiento? o, dicho de otra forma, ¿qué ocurre si seguimos de generación en generación?

Antes de abordar esta cuestión, es interesante notar que $\mathbf{I}_2 = \mathbf{I}_1 \times \mathbf{T} = (\mathbf{I}_0 \times \mathbf{T}) \times \mathbf{T} = \mathbf{I}_0 \times (\mathbf{T} \times \mathbf{T}) = \mathbf{I}_0 \times \mathbf{T}^2$.[14] Claramente, lo anterior se puede generalizar a $\mathbf{I}_{t=k} = \mathbf{I}_0 \times \mathbf{T}^{(k)}$. En palabras, para calcular la distribución de probabilidades de los estados (número de alelos **A** en nuestro **ensemble** de poblaciones) en cualquier generación k (donde consideramos $t = k$, $k > 0$) basta con multiplicar la distribución en la generación inicial (o sea en $t = 0$, que llamamos $\mathbf{I}_0$) por la matriz de transición de probabilidades elevada a la potencia k.

Volvamos entonces a la pregunta anterior: ¿hacia dónde va el movimiento de las frecuencias en las distintas poblaciones a largo plazo? Para abordar esta cuestión, tendremos en cuenta el razonamiento de que $\mathbf{I}_{t=k} = \mathbf{I}_0 \times \mathbf{T}^{(k)}$. Resulta interesante ver cómo se comporta $\mathbf{T}^{(k)}$ cuando $k \to \infty$, es decir cuando el número de generaciones tiende a infinito. Una aproximación numérica sencilla consiste en elevar $\mathbf{T}$ a un número suficientemente grande (por ejemplo, 50 generaciones) y ver la forma que adquiere la matriz, como aparece en la tabla 3.9.

[14] Encontrarás más información sobre productos matriciales en el apéndice correspondiente a álgebra matricial.

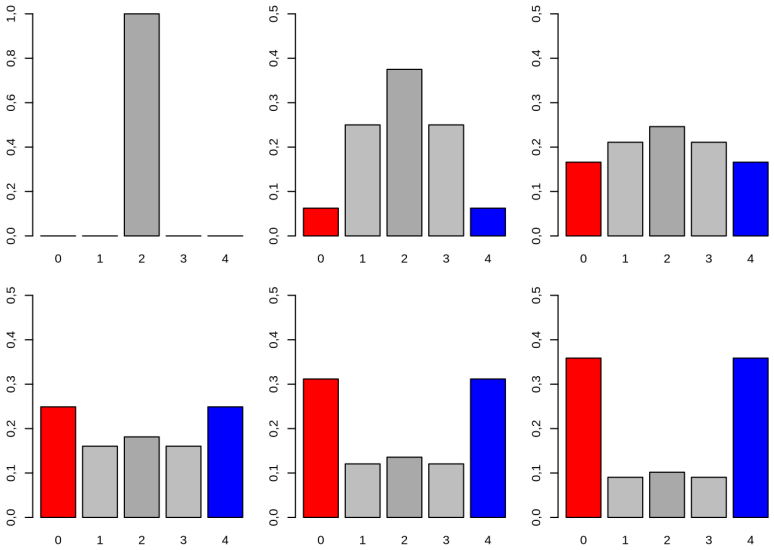


Figura 3.12. Evolución de la probabilidad de encontrar a las poblaciones en un estado alélico determinado desde la generación $t = 0$ a $t = 5$ (desde el costado superior izquierdo al inferior derecho). En rojo el estado de 0 alelos A (pérdida del mismo), en azul 4 alelos A (fijación del mismo) y en gris oscuro 2 alelos (nuestro punto de partida en todas las poblaciones). Notemos la diferente escala de la generación $t = 0$ al resto.

Tabla 3.9. Probabilidad de terminar en cualquiera de los estados posibles, a partir del estado inicial

	a 0 alelos A	a 1 alelos A	a 2 alelos A	a 3 alelos A	a 4 alelos A
de 0 alelos A	1,00	0	0	0	0,00
de 1 alelos A	0,75	0	0	0	0,25
de 2 alelos A	0,50	0	0	0	0,50
de 3 alelos A	0,25	0	0	0	0,75
de 4 alelos A	0,00	0	0	0	1,00

Si miramos las columnas de la tabla 3.9 podemos apreciar dos situaciones claramente distintas; en las columnas 2, 3 y 4 (que corresponden a pasar de una generación a la siguiente a los estados de 1, 2 y 3 alelos A respectivamente), todas las celdas son 0, es decir más allá del punto de arranque (el vector de frecuencias o probabilidades $\mathbf{I}_0$), cuando el tiempo tiende a infinito (en principio lo estaríamos viendo con solo 50 generaciones) estos estados alélicos estarán vacíos (ninguna población tendrá 1, 2 o 3 alelos A), en nuestro modelo de dos individuos diploides. En cambio, las columnas 1 y 5, que se corresponden con la fijación del alelo a (pérdida del alelo A) y la fijación del alelo A respectivamente plantean una distribución que parece en espejo una de otra; claramente estas columnas se corresponden con las dos **barreras absorbentes**. Si nos fijamos en la columna 1, el único 0 corresponde a la probabilidad de pasar de 4 alelos A a 0 alelos A. Esto es trivial, ya que si el alelo A se encuentra fijado, al no existir mutación es imposible que las poblaciones evolucionen a otro estado alélico. Inversamente, en la columna 5 el único 0 se corresponde a la probabilidad

de pasar de 0 a 4 alelos A. Esto también es trivial, ya que no existe posibilidad de pasar del alelo a fijado a otro estado alélico. El resto de las celdas es fácil de explicar, si recordamos las probabilidades de fijación de un alelo dada su frecuencia inicial en la población. Cuando se parte de 1 solo alelo A (por ejemplo, 3 alelos a), la probabilidad de fijación del alelo a es $\frac{3}{4}$; cuando se parte de 2 alelos A, es de $\frac{2}{4} = \frac{1}{2}$, y cuando se parte de 3 alelos A (es decir, solo 1 alelo a), la probabilidad de fijación del alelo a es $\frac{1}{4}$. Lo mismo acontece de forma inversa para las transiciones representadas en la columna 5.

La conclusión de todo lo anterior no es menor: con la matriz de probabilidad de transiciones $\mathbf{T}$ (tabla 3.5) el estado estacionario (cuando $t \to \infty$) implica que **solo los estados de fijación de alguno de los alelos tienen probabilidad mayor a cero**. El punto de partida solo influirá en la distribución entre ellos, pero no en que solo ellos pueden tener probabilidad positiva de fijación. Pongamos un ejemplo: todas las poblaciones inicialmente tienen 1 alelo A de los 4 posibles (recordar que son dos individuos diploides, $2N = 4$), es decir, $\mathbf{I}_0 = (0; 1; 0; 0; 0)$. Si multiplicamos este vector inicial de probabilidades (probabilidades de encontrar poblaciones del **ensemble** en ese estado), tenemos que $\mathbf{I}_0 \times \mathbf{T}^\infty$ (llamamos $\mathbf{T}^\infty$ a la potencia infinita, aunque en este contexto $\mathbf{T}^{50} \sim \mathbf{T}^\infty$), el resultado es $\mathbf{I}_0 \times \mathbf{T}^\infty = (0{,}75; 0; 0; 0; 0{,}25)$, claramente en línea con lo que esperábamos: ninguna población aún segregando (1, 2 o 3 alelos A), 75 % de las poblaciones con el alelo a fijado y 25 % de las poblaciones con el alelo A fijado.

Si en lugar de haber arrancado con todas las poblaciones con 1 solo alelo A, hubiésemos arrancado con 25 % de las poblaciones con 1 alelo A, 25 % con dos alelos y 50 % con 3 alelos A, entonces $\mathbf{I}_0 = (0; 0{,}25; 0{,}25; 0{,}50; 0)$. Por lo tanto $\mathbf{I}_0 \times \mathbf{T}^\infty = (0{,}4375; 0; 0; 0; 0{,}5625)$, es decir que el 43,75 % de las poblaciones de nuestro **ensemble** tendrían fijado el alelo a y el restante 56,25 % tendría fijado el alelo A.

En general, si la deriva genética es la única fuerza evolutiva actuante (sin selección, sin mutación, sin migración), el equilibrio se alcanza cuando más tarde o más temprano todas las poblaciones se encuentran fijadas para uno de los dos alelos.

Un último punto que cabe mencionar acá. Si observamos la forma en la que crece la suma de los estados correspondientes a las **barreras absorbentes**, es posible demostrar que a partir de determinado punto la tasa de aproximación al estado de equilibrio es igual a $(1 - \frac{1}{2N})$. Dicho de otra forma, a partir de determinado momento, a cada generación la proporción de poblaciones aún segregantes se reduce en $\frac{1}{2N_e}$, proporción que va a parar de los estados correspondientes a las **barreras absorbentes**.

3.5 Tamaño efectivo poblacional

Un punto que mencionamos antes pero no abundamos fue la diferencia entre el tamaño censal de una población, es decir el número N de individuos que la conforman y el tamaño efectivo de esta, que llamamos N_e. A veces resulta claro cuando en determinadas estructuras poblacionales particulares no tiene ningún sentido el número N como un indicador del comportamiento genético de la población, por ejemplo, cuando un solo macho es responsable de todas las crías, pero otras veces no resulta tan claro. En todo caso, dada la dependencia de varias de nuestras ecuaciones de algún número que represente el tamaño poblacional, es necesario definir qué significa ese número a la luz de las cosas que hemos asumido en nuestros modelos.

Un ejemplo claro lo representa el supuesto de tamaño poblacional constante que hemos usado en el modelo de Wright-Fisher. Desafortunadamente, este supuesto es difícil de constatar en la realidad ya que las poblaciones fluctúan de tamaño, algunas en forma de tendencia (al crecimiento o decrecimiento) y otras a través de ciclos (por ejemplo, tamaños poblacionales de presa y predador, que están íntimamente relacionados). En este sentido, podríamos pensar en obtener un número poblacional en una población **perfecta** (para nuestro modelo) en la que de alguna manera el efecto genético de la deriva fuese equivalente al de nuestra población real. Si el tamaño poblacional fluctúa en el tiempo, podríamos, por ejemplo, pensar en que la media aritmética (el clásico promedio, que conoces muy bien) del tamaño censal sería un buen resumen del impacto genético de la deriva en la población. Desafortunadamente, el impacto genético de los tamaños poblacionales pequeños es desproporcionadamente alto, como puede verse en la figura 3.13.

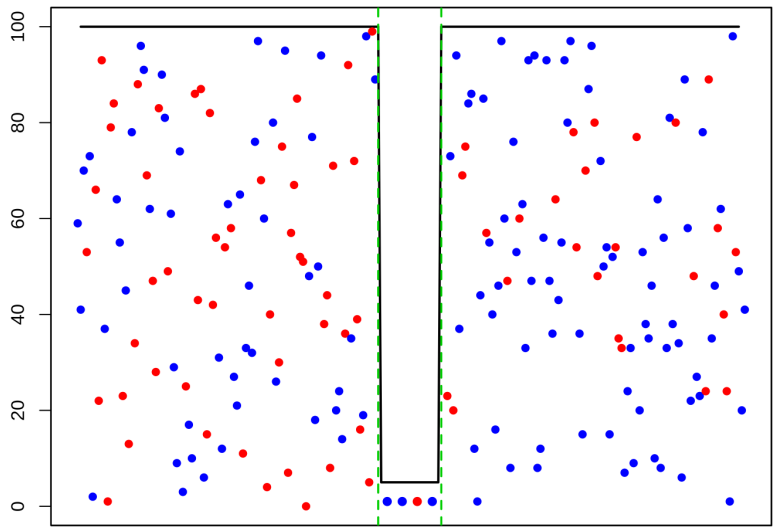


Figura 3.13. Efecto de un cuello de botella poblacional en la frecuencia alélica. La población de 100 alelos, 50 rojos y 50 azules, se ve drásticamente reducida en una generación a 4 alelos (entre las dos líneas verdes), que por muestreo aleatorio son ahora 3 azules y 1 rojo. Luego de ese cuello de botella la población vuelve a expandirse al nivel inicial, pero la frecuencia ya quedó alterada en forma importante (figura propia sobre idea de @Hamilton2009).

En efecto, pese a que la población inicial era de 50 individuos diploides, o sea 100 alelos, 50 rojos y 50 azules, al sufrir un cuello de botella poblacional (sobreviven 2 individuos diploides, marcado por las dos líneas verdes en la figura 3.13) se muestrean 3 alelos azules y 1 rojo (un evento que tiene una probabilidad de $\frac{1}{4} = 25\,\%$). Al expandirse nuevamente la población a 50 individuos, es decir 100 alelos, lo hará desde la frecuencia del cuello de botella. Resulta clara la influencia desproporcionada (respecto a un promedio simple) que tiene el muestreo durante el cuello de botella poblacional.

Veamos entonces una forma más razonable de estimar el tamaño de una población que fluctúa en número. Para hacer nuestro razonamiento supongamos que la población va de la generación 0 hasta la 2. De acuerdo con lo que vimos en la sección «Homocigosidad y heterocigosidad»,

$$1 - F_2 = \left(1 - \frac{1}{2N_1}\right)(1 - F_1) \tag{3.19}$$

$$1 - F_1 = \left(1 - \frac{1}{2N_0}\right)(1 - F_0) \tag{3.20}$$

Sustituyendo (3.20) en (3.19), tenemos

$$1 - F_2 = \left(1 - \frac{1}{2N_1}\right)\left(1 - \frac{1}{2N_0}\right)(1 - F_0) \tag{3.21}$$

Si el tamaño N fuese constante, obtendríamos el N_e generalizando esta expresión para un tiempo t

$$1 - F_t = (1 - \frac{1}{2N_e})^t(1 - F_0) \tag{3.22}$$

Por ejemplo, en el caso en que $t = 2$, la ecuación (3.22) se transforma en

$$1 - F_2 = (1 - \frac{1}{2N_e})^2(1 - F_0) \tag{3.23}$$

Ahora, si igualamos el resultado «real» (ecuación (3.21)) al «ideal» (ecuación (3.23)), tenemos

$$\begin{aligned} 1 - F_2 = (1 - \frac{1}{2N_e})^2(1 - F_0) = (1 - \frac{1}{2N_1})(1 - \frac{1}{2N_0})(1 - F_0) = 1 - F_2 \\ (1 - \frac{1}{2N_e})^2 = (1 - \frac{1}{2N_1})(1 - \frac{1}{2N_0}) \end{aligned} \tag{3.24}$$

Desarrollando la ecuación (3.24), tenemos que

$$1 - 2\frac{1}{2N_e} + \frac{1}{4N_e^2} = 1 - \frac{1}{2N_0} - \frac{1}{2N_1} + \frac{1}{4N_0N_1}\frac{1}{N_e} - \frac{1}{4N_e^2} = \frac{1}{2}\left(\frac{1}{N_0} + \frac{1}{N_1}\right) - \frac{1}{4N_0N_1} \tag{3.25}$$

y considerando que los términos en $\frac{1}{4N_e^2}$ y $\frac{1}{4N_0N_1}$ no afectan casi el resultado, porque su orden es casi similar, llegamos a que

$$\frac{1}{N_e} = \frac{1}{2}(\frac{1}{N_0} + \frac{1}{N_1}) \tag{3.26}$$

es una muy buena aproximación. De hecho, esto significa que N_e es la **media armónica** entre N_0 y N_1. En general, para $t \geq 2$ la media armónica (el recíproco de la media de recíprocos) es un buen indicador del **tamaño efectivo poblacional**

$$\frac{1}{N_e} = \frac{1}{t}\left(\frac{1}{N_0} + \frac{1}{N_1} + \frac{1}{N_2} + ... + \frac{1}{N_{t-1}}\right) \tag{3.27}$$

$$N_e = \frac{t}{\left(\frac{1}{N_0} + \frac{1}{N_1} + \frac{1}{N_2} + ... + \frac{1}{N_{t-1}}\right)} \tag{3.28}$$

Ejemplo 3.3

El venado de campo (*Ozotoceros bezoarticus*) fue una especie típica de nuestro país («extremadamente abundante», en palabras de Charles Darwin, que al pasar por estas tierras no dudó en llevarse ejemplares en el Beagle). Lamentablemente, la población ha sufrido una disminución importante durante el último siglo. Imaginemos que en un momento una población estaba compuesta por 50.000 individuos, y que en la actualidad se compone por 1000 individuos. Si en un esfuerzo descomunal se lograra recuperar el tamaño de la población, ¿cuál sería su tamaño poblacional efectivo, considerando que pasó por dicho cuello de botella?

El tamaño poblacional efectivo (N_e) se calcula a partir de la media armónica de los tamaños poblacionales mencionados:

$$\frac{1}{N_e} = \frac{1}{3}(\frac{1}{50000} + \frac{1}{1000} + \frac{1}{50000}) \Rightarrow$$

$$N_e = \frac{3}{(\frac{1}{50000} + \frac{1}{1000} + \frac{1}{50000})}$$

lo cual da una estimación de

$$N_e \approx 2885$$

Otra fuente importante de diferencias entre el tamaño censal y el tamaño efectivo de una población es el desequilibrio entre el número de machos y el número de hembras, como veremos en detalle en otros capítulos. En general, si una población está constituida por N_m machos y N_h hembras, el número total de animales (individuos) es $N_a = N_m + N_h$, mientras que (como deduciremos en otro capítulo), para una generación determinada

$$N_e = \frac{4N_m N_h}{N_m + N_h} \tag{3.29}$$

Cuando el número de hembras es igual al número de machos ($N_h = N_m = \frac{N}{2}$), algo bastante raro en las poblaciones comerciales (al menos de mamíferos), tenemos que para la misma generación la ecuación (3.29) se transforma en

$$N_e = \frac{4N_m N_h}{N_m + N_h} = \frac{4\frac{N}{2}\frac{N}{2}}{\frac{N}{2} + \frac{N}{2}} = \frac{N^2}{N} = N$$

Es decir, $N_e = N$, el tamaño efectivo poblacional es igual al tamaño censal.

Ejemplo 3.4

El león marino (*Otaria flavescens*) se reproduce de forma estacional, siguiendo una estructura de harén: un macho alfa territorial monopoliza la reproducción, apareándose con varias hembras. Grandi et al.(2012) realizaron censos en colonias de *O. flavescens* en el norte de la Patagonia, contando el número de machos alfa y de hembras sexualmente activas. En el año 2012, el censo de 33 colonias arrojó un conteo de 2534 machos alfa y 12.097 hembras sexualmente activas. ¿Cuál es el tamaño efectivo poblacional en este caso?

Si se asume que a lo largo de las estaciones existe un apareamiento aproximadamente al azar entre machos alfa y hembras sexualmente activas, se puede estimar el tamaño poblacional efectivo teniendo en cuenta el desbalance de sexos según

$$N_e = \frac{4N_h N_m}{N_h + N_m}$$

$$N_e = \frac{4 \times 12097 \times 2534}{12097 + 2534} \Rightarrow N_e = 8381$$

Finalmente, otra versión del tamaño efectivo poblacional es la que sale de nuestra definición del modelo ideal y de cuánto esperamos que sea su varianza. Si llamamos $\Delta p = p_t - p_{t-1}$, es decir a la diferencia entre la frecuencia promedio en la generación t y la generación anterior $(t-1)$, entonces la varianza de Δp estará dada por

$$Var(\Delta p) = \frac{p_{t-1}q_{t-1}}{2N} \tag{3.30}$$

cuando el proceso cumple con todos los requerimientos del modelo de Wright-Fisher. Pero esto mismo podemos usarlo para estimar cuál sería el tamaño efectivo poblacional que

generaría la misma varianza empírica que tenemos en Δp, N_e^v. Es decir, despejando N, que ahora llamaremos N_e^v, que es nuestra incógnita asumiendo que proviene de un proceso perfecto Wright-Fisher, tenemos

$$N_e^v = \frac{pq}{2Var(\Delta p)} \tag{3.31}$$

En resumen, en esta última forma de calcular el tamaño efectivo poblacional hemos asumido que la varianza entre generaciones en la media de frecuencias alélicas proviene de un proceso de Wright-Fisher y esto nos permite estimar cuál sería el tamaño de las poblaciones que se corresponden con esa varianza.

PARA RECORDAR

- El tamaño efectivo de una población (N_e) raramente coincide con su tamaño censal (N), siendo las razones para esto diversas (variación del tamaño en el tiempo, diferente número de machos que de hembras, diferencias muy importantes en el número de descendientes entre reproductores candidatos, etcétera).
- El efecto conocido como **cuello de botella** (*bottleneck*) poblacional se refiere a la **reducción drástica** del número de reproductores en algún momento de la vida de una población determinada y, aunque luego esta experimente una nueva expansión, el cuello de botella dejará huellas indelebles a nivel de la variabilidad genómica.
- El tamaño efectivo poblacional cuando varía el tamaño de la población en el tiempo se puede calcular como la media armónica entre los tamaños en cada generación:

$$N_e = \frac{t}{\left(\frac{1}{N_0} + \frac{1}{N_1} + \frac{1}{N_2} + ... + \frac{1}{N_{t-1}}\right)}$$

- Cuando el número de reproductores machos es diferente del número de reproductores hembras, el tamaño efectivo de la población es inferior a la suma de individuos y lo podemos calcular como

$$N_e = \frac{4N_mN_h}{N_m + N_h}$$

- Si asumimos que se trata de un proceso de Wright-Fisher perfecto, podemos obtener un estimado del tamaño efectivo poblacional que se corresponde con la varianza en la media de frecuencias alélicas entre generaciones a partir de la dicha varianza **observada** mediante

$$N_e^v = \frac{pq}{2Var(\Delta p)}$$

3.6 Aproximación de difusión

Gran parte de los desarrollos de esta sección siguen directamente el libro de Hartl y Clark (2007), que entendemos es de los más claros para el nivel del presente curso, mientras que otros siguen el libro de Walsh y Lynch (2018), que posee un apéndice dedicado al tema. Otras fuentes importantes son el libro de Denny y Gaines (2000) para los procesos de difusión en biología y el libro de Hamilton (2009), que seguimos directamente también (más allá de modificaciones mínimas) y que trata con cierto detalle pero a un nivel muy razonable la aproximación por difusión en genética —que a su vez sigue el de Denny y Gaines (2000)—.
Inevitablemente, esta sección es bastante técnica y usaremos cálculo diferencial e integral en múltiples variables, así como algunos conceptos básicos de ecuaciones diferenciales en derivadas parciales. Si te sientes un poco oxidado en estos temas, en el apéndice B: «Cálculo diferencial e integral» podrás repasar los principales conceptos.
Dada la complejidad de esta sección, no pretendemos que seas capaz de reproducir o aún entender cada paso o derivación (aunque nos pondría muy contentos si así fuese). Si te resulta imposible entender no te desanimes, ya que podrás seguir el resto del curso en forma independiente de las derivaciones de esta sección. La idea es que puedas entender de dónde surgen algunos resultados muy importantes que veremos en este y otros capítulos. Es muy importante para la mente científica conocer y entender el origen de resultados y fórmulas que utilizarás a lo largo de tu carrera, tanto como estudiante como profesional. ***¡Ánimo!***

Como vimos antes en la sección «Cadenas de Markov», la evolución del número de copias de un alelo en poblaciones idénticas es un proceso estocástico discreto y la evolución de los estados puede seguirse a través de la multiplicación de la matriz de probabilidades de transición (cadenas de Markov). Sin embargo, esta aproximación suele ser dificultosa o imposible en la práctica para manejar modelos más complejos y una buena aproximación para comprender el comportamiento a largo plazo viene por el lado de los procesos de difusión. En efecto, si consideramos el tiempo como una variable continua, reescalando el tiempo en generaciones que habíamos utilizado hasta ahora mediante la transformación $\tau = \frac{t}{N}$ y escalando el número de copias del alelo A (i) de tal forma que $x = \frac{i}{2N}$, cuando $N \to \infty$ ambas variables pasan a ser continuas.

Caminatas al azar y procesos de difusión

En la sección «El modelo de Wright-Fisher» vimos el comportamiento de la frecuencia alélica guiada solo por el proceso de deriva genética. Si observamos nuevamente el trazo dejado por cualquiera de las poblaciones, no resulta difícil imaginarse el recorrido de alguien que (posiblemente borracho) a cada instante en el tiempo, decide a través del resultado de tirar una moneda si el siguiente paso es a la derecha o a la izquierda. Más aún, podemos pensar que esta persona se encuentra sobre una línea trazada en el piso y que arranca desde

una determinada posición inicial decidiendo con la moneda hacia dónde dar el siguiente paso. Lo que nos interesa entonces es estudiar el comportamiento de este tipo de procesos y ver si podemos hacer algunas predicciones al respecto. Más divertido aún y más cercano a lo que vamos a describir ahora sería un experimento donde a la salida de un bar (en un lugar medio desierto) reclutamos a todos los borrachos disponibles, los ponemos en «fila india» y les explicamos que al sonido de la campana (que tañeremos en forma regular) deben tirar una moneda y en función del resultado dar un paso a la derecha (si sale cara) o a la izquierda (si sale cruz); o aún quedarse quietos si no sintieron la campana o estaban distraídos. Bueno, también se podría hacer sin borrachos, pero intuimos que será mucho más divertido en el primer caso.

En realidad, existe una motivación física mucho más potente y que usaremos como analogía para luego pasar a ver cómo esa analogía nos puede ayudar a entender los procesos estocásticos en las frecuencias alélicas de nuestras poblaciones. La motivación es muy sencilla: imaginemos que en el instante $t = 0$ dejamos caer una gota de tinta en la posición central de un canal horizontal abierto que contiene agua (el canal se extiende a derecha, signo positivo, e izquierda, signo negativo, de $x = 0$). En ambos extremos del canal tenemos un medio que absorbe las partículas de tinta a medida que van arribando (es decir, son **barreras absorbentes** para las partículas de tinta). El papel de nuestros «borrachos» lo representarán ahora las pequeñas partículas de tinta que formaban la gota y que ahora, a partir del instante $t = 0$, comenzarán a dispersarse en ambas direcciones. Para que nuestra analogía tenga cierto sentido físico y poder estudiar su comportamiento vamos a definir una serie de reglas que se aplican al movimiento individual de cada partícula.

Reglas

1. La partícula comienza en la posición $x = 0$ en el instante inicial, $t = 0$.
2. La partícula se mueve **una distancia fija** δ en cada intervalo de tiempo τ.
3. La partícula se puede mover solo en el eje permitido, con probabilidad $p = \frac{1}{2}$ hacia la derecha (arbitrariamente, puede ser cualquier dirección) y, por lo tanto, con probabilidad $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$ hacia la izquierda (o en sentido opuesto al anterior).
4. El sentido en el que se mueve es independiente del sentido en la movida anterior o, lo que es lo mismo, se trata de un proceso **sin memoria** (recuerda lo visto en la sección «Cadenas de Markov»).
5. Pueden existir tantas partículas como se quiera en la misma posición del eje sin interferir unas con otras (es como si cada una «corriese» por su propio carril, todos los carriles paralelos unos a otros).

En la figura 3.14 podemos ver una representación esquemática de nuestro diseño. En el instante $t = 0$ consideramos que cae la gota de tinta (roja) al agua **al mismo tiempo** que las **partículas** que la constituyen comienzan a difundir hacia izquierda y derecha de acuerdo con las reglas que establecimos antes.

Primero, podríamos analizar qué ocurre con la posición promedio de las partículas a medida que avanza el tiempo, por ejemplo, en una unidad (τ). La posición media está dada por

$$\bar{x} = p(\delta) + q(-\delta) \tag{3.32}$$

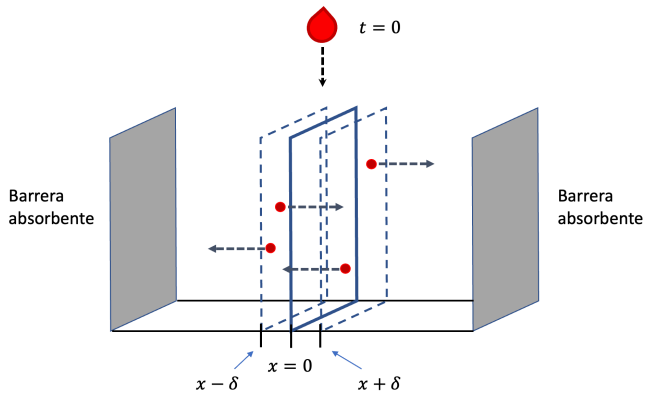


Figura 3.14. Diagrama que representa el escenario para el movimiento de nuestras partículas de tinta. En el instante inicial la gota de tinta cae al mismo tiempo sus partículas constituyentes comienzan a difundir a izquierda y derecha (elaboración propia sobre idea de @Hamilton2009).

Como las partículas se mueven a derecha e izquierda con la misma probabilidad, δ y $-\delta$ unidades de posición, entonces tenemos que luego de una unidad de tiempo τ la posición media será en nuestro caso

$$\bar{x} = \frac{1}{2}(\delta) - \frac{1}{2}(\delta) = 0 \tag{3.33}$$

Es decir que la posición media de las partículas será la misma que al comienzo. Esto no debería extrañarnos dado que por simetría del problema tantas partículas se moverán hacia un lado como otras en el sentido opuesto, lo que da un balance neto de cero respecto al desplazamiento del punto medio. Entonces, ¿nada ha cambiado? Si recuerdas el comportamiento de las distintas poblaciones en la figura 3.5, posiblemente te habrás imaginado que, si bien la media del proceso no cambia, la varianza sí lo hace. Si además recuerdas la fórmula para la varianza

$$Var(x) = E(x^2) - E^2(x) \tag{3.34}$$

y que además en nuestro caso $E^2(x) = E(x)E(x) = \bar{x}\bar{x} = 0 \times 0$, llegaremos enseguida a la conclusión de que la varianza en posición de las partículas es igual a $E(x^2)$ o, lo que es lo mismo, el promedio de las posiciones al cuadrado. Como además, en cada intervalo arbitrariamente pequeño de tiempo la partícula solo puede moverse δ en uno de los sentidos, la ubicación de la partícula i en el tiempo 1 será igual a

$$x_{i,(t=1)} = x_{i,(t=0)} + \delta \tag{3.35}$$

con δ positivo o negativo.

O sea, para obtener una expresión de la varianza en las posiciones de las partículas, luego de un intervalo arbitrariamente pequeño (de $t = 0$ a $t = 1$) debemos comenzar por elevar al cuadrado esta expresión y luego promediar

$$x^2_{i,(t=1)} = (x_{i,(t=0)} + \delta)^2$$
$$x^2_{i,(t=1)} = x^2_{i,(t=0)} + 2x_{i,(t=0)}\delta + \delta^2$$

Por lo tanto, promediando el resultado anterior para las N partículas obtenemos

$$\sigma^2(x_{i,(t=1)}) = \frac{1}{N}\sum_{i=1}^{N}(x^2_{i,(t=0)} + 2x_{i,(t=0)}\delta + \delta^2) \tag{3.36}$$

La expresión anterior se puede simplificar bastante. En primer lugar, por una razón de simetría, habrá tantas partículas que se muevan a la derecha como a la izquierda (recuerda $\bar{x} = 0$, ecuación (3.33)), por lo que el término del medio es igual a cero. Luego

$$\sigma^2(x_{i,(t=1)}) = \frac{1}{N}\sum_{i=1}^{N}x^2_{i,(t=0)} + \frac{1}{N}N\delta^2$$
$$\sigma^2(x_{i,(t=1)}) = \overline{x^2}_{i,(t=0)} + \delta^2$$

Ahora, suponiendo que a $t = 0$ todas las partículas salen de la posición cero, entonces el promedio del cuadrado de sus distancias será 0 también ($\overline{x^2}_{i,(t=0)} = 0$), por lo que nos quedamos con δ^2, es decir, el cuadrado del cambio en posición en un intervalo de tiempo, como el incremento en la varianza en un avance arbitrariamente pequeño del tiempo (de $t = 0$ a $t = 1$). En un intervalo de tiempo de t unidades, por lo tanto, la varianza de la posición de las partículas será de $t\delta^2$, o dicho de otra manera, la varianza en la posición de las partículas se incrementa en forma lineal con el tiempo. En el primer instante la varianza será $1 \times \delta^2 = \delta^2$, en el segundo instante será $2\delta^2$, luego $3\delta^2$, etcétera. La anterior es una observación más que razonable para describir el fenómeno de cómo una gota de tinta se dispersa en el agua a medida que pasa el tiempo.

Si definimos el **coeficiente de difusión** de las partículas (D) como la velocidad con que la tinta se difunde en el agua, tenemos que

$$D = \frac{1}{2}\frac{d\sigma^2}{dt} \tag{3.37}$$

La ecuación anterior es fácil de entender, la varianza representa la dispersión de las partículas en un momento dado y, por lo tanto, su derivada primera respecto del tiempo es su velocidad. Pero la varianza representa dispersión a ambos lados del punto de inicio, por lo que solo la mitad de esta corresponderá a la magnitud de la velocidad en un sentido (de lo contrario, la velocidad neta siempre sería cero). Como $\sigma^2 = t\delta^2$, entonces

$$D = \frac{1}{2}\frac{d\sigma^2}{dt} = \frac{1}{2}\frac{d(t\delta^2)}{dt} = \frac{\delta^2}{2} \tag{3.38}$$

Es decir, $D = \frac{\delta^2}{2}$ es la velocidad de difusión de las partículas, o lo que es lo mismo, un medio del incremento en la varianza en una unidad de tiempo arbitrariamente pequeño.

Volvamos ahora a la genética. ¿Cómo se relaciona esto del proceso de difusión con nuestro tratamiento de la evolución estocástica de las frecuencias alélicas en las poblaciones? En

principio es muy sencillo: si consideramos que cada población del **ensemble** es como una partícula en nuestro modelo de difusión, los «movimientos» que pueden realizar nuestras poblaciones son en lugar del movimiento físico de las partículas, los cambios en frecuencia alélica en la población. O sea, si por ejemplo, partimos de $p_A = 0{,}5$, es equivalente a partir del punto medio en el modelo de difusión. Ahora las **barreras absorbentes** estarán en $p_A = 0$ y $p_A = 1$. Por otra parte, si recuerdas lo tratado en la sección «El modelo de Wright-Fisher», la **varianza por generación** (o lo que sería un intervalo arbitrariamente pequeño en la difusión) era $\sigma^2(p) = \frac{pq}{2N}$ y la varianza a lo largo del tiempo, en la generación t, por ejemplo, igual a $\sigma_t^2(p) = t\frac{pq}{2N}$ (ecuación (3.2), para simplificar la notación llamaremos σ^2 a $\sigma_t^2(p)$), por lo que podemos hacer

$$\frac{d\sigma^2}{dt} = \frac{d(t\frac{pq}{2N})}{dt} = \frac{pq}{2N} \tag{3.39}$$

Si recordamos de la ecuación (3.37) que $D = \frac{1}{2}\frac{d\sigma^2}{dt}$ y sustituimos en ella el resultado de la ecuación (3.39), tenemos que el coeficiente de difusión D es ahora

$$D = \frac{1}{2}\frac{d\sigma^2}{dt} = \frac{1}{2}\frac{pq}{2N} = \frac{pq}{4N} = \frac{p - p^2}{4N} \tag{3.40}$$

Es decir, el coeficiente de difusión para las frecuencias alélicas de nuestras poblaciones es una función de la frecuencia de los alelos (p y q). El denominador de la ecuación (3.40) no depende de las frecuencias, ya que se trata de un número N positivo, multiplicado por una constante positiva (**4**). En cambio, el numerador depende de la frecuencia del alelo **A** (p): si recuerdas de la sección «H-W: la frecuencia de heterocigotas en función de la frecuencia alélica», el producto $p \times q$ tiene un máximo cuando $p = q = \frac{1}{2}$. Es decir, la velocidad de difusión (cambio de frecuencias alélicas) será máxima cerca de $p = \frac{1}{2}$ y tenderá a cero a medida que p se acerque a las barreras absorbentes de 0 y 1. Dicho de otra manera, la velocidad de cambio en la frecuencia en cada población será máximo en el entorno de $p = \frac{1}{2}$, llevando a las poblaciones hacia las regiones donde el coeficiente de difusión es menor, cerca de las barreras absorbentes, donde al ser menor el coeficiente de difusión permanecerá más tiempo. Por otra parte, el coeficiente de difusión también depende del tamaño poblacional efectivo: a mayor número de individuos en cada una de las poblaciones que constituyen el **ensemble**, menor será el coeficiente de difusión o, lo que es lo mismo, menor la velocidad con que cambian las frecuencias de los alelos en cada población.

Procuraremos ahora calcular la probabilidad de que una población se encuentre en una región en particular en el intervalo de frecuencias, o en nuestra analogía de difusión la probabilidad de encontrar una partícula en una posición en particular del eje en que se mueven. Para eso necesitaremos calcular el flujo de partículas en la región de interés o, dicho de otra forma, el número neto de partículas que atraviesan un área definida por intervalo de tiempo. Si consideramos un área A, el flujo neto de partículas que la atraviesa hacia la derecha (arbitrariamente) es igual a la mitad del número de partículas que están a una distancia δ a la izquierda de la misma (recuerda que en cada posición la mitad de las partículas tenderá a moverse a la izquierda y la otra mitad a la derecha), menos la mitad de las partículas que están a una distancia δ a la derecha del área de referencia (δ, porque es la distancia a la que se mueven en una unidad de tiempo).

Es decir

$$N_{neto}(Der) = \frac{1}{2}N(Izq) - \frac{1}{2}N(Der) = -\frac{1}{2}[N(Der) - N(Izq)] \tag{3.41}$$

Pero el flujo es la cantidad de partículas que atraviesan un área en determinado tiempo, por lo que para obtener una expresión de este en el punto x, que llamaremos J_x, debemos dividir por área y por tiempo

$$J_x = -\frac{1}{2}\frac{[N(Der) - N(Izq)]}{At} \tag{3.42}$$

Si multiplicamos esta última expresión por 1, bajo la forma de $\frac{\delta^2}{\delta^2}$, para luego rearreglar los términos, quedará

$$J_x = -\frac{\delta^2}{2t}\frac{1}{\delta}\frac{[N(Der) - N(Izq)]}{A\delta} = -\frac{\delta^2}{2t}\frac{1}{\delta}\left[\frac{N(Der)}{A\delta} - \frac{N(Izq)}{A\delta}\right] \tag{3.43}$$

Si observamos los dos términos entre paréntesis rectos de la ecuación anterior, en ambos casos ($\frac{N(Der)}{A\delta}$ y $\frac{N(Izq)}{A\delta}$) se trata de un número de partículas dividido entre el producto de un área (A) por una distancia (δ), es decir, dividido entre un volumen (recuerda de la escuela cómo calculábamos los volúmenes). Pero el número de partículas dividido entre el volumen que ocupan es la concentración, por lo que podemos llamar $C(Der)$ y $C(Izq)$ a la concentración de partículas a derecha e izquierda de nuestra área. Además, previamente habíamos visto (en la ecuación (3.38)) que $\frac{\delta^2}{2} = D$, por lo que en una unidad de tiempo ($t = 1$), sustituyendo en (3.43) tenemos

$$J_x = -D\frac{C(Der) - C(Izq)}{\delta} \tag{3.44}$$

Si asumimos que las partículas se encuentran uniformemente distribuidas en el intervalo entre x y $x+\delta$, su posición promedio será $\frac{\delta}{2}$ (lo mismo para el otro lado, pero ahora será $-\frac{\delta}{2}$), que tomaremos como posición característica. Entonces $C(Der) = C(x + \frac{\delta}{2})$ y $C(Izq) = C(x - \frac{\delta}{2})$, por lo que la ecuación (3.44) la podemos escribir como

$$J_x = -D\frac{C(x + \frac{\delta}{2}) - C(x - \frac{\delta}{2})}{\delta} \tag{3.45}$$

Si en esta última expresión hacemos el paso al límite cuando las distancias en x son infinitamente pequeñas), es decir cuando $\delta \to 0$, tenemos que

$$J_x = -D\frac{dC}{dx} \tag{3.46}$$

que se conoce como la **primera ecuación de difusión de Fick**. Nuevamente, esta ecuación parece sencilla de interpretar. El coeficiente de difusión es un número positivo que indica la velocidad de movimiento de las partículas en el medio. Por otro lado, $\frac{dC}{dx}$ representa el gradiente de la concentración de partículas (en nuestro caso de poblaciones) y cuando su

signo es positivo indica la dirección hacia la mayor concentración de partículas. Como acá el producto tiene un signo negativo, la ecuación nos indica que las partículas (poblaciones) se moverán en dirección a donde hay una menor concentración (recordar la analogía con la gota de tinta en agua; la tinta se moverá hacia donde hay menor concentración). Obviamente, en el equilibrio, para cada posición en el eje de movimiento (nuestras frecuencias alélicas) tantas partículas se moverán hacia un lado como hacia el otro. Intentemos ahora comprender cómo esto nos puede aportar a entender el comportamiento de las frecuencias alélicas en nuestro **ensemble**. Para esto, debemos recordar que estamos tratando con una aproximación continua (las frecuencias alélicas ya no las expresamos como el número de copias de un alelo sino como este número dividido entre $2N$) y, por lo tanto, la probabilidad la vamos a sustituir por una **densidad de probabilidad**, que llamaremos $\phi(p,x,t)$ (notemos la dependencia del punto inicial de partida p, o sea la frecuencia inicial del alelo **A**). Esta densidad de probabilidad es equivalente a la concentración de partículas. $\phi(p,x,t)$ representa el balance neto entre las poblaciones que arriban a esa frecuencia alélica x y aquellas que se alejan de la misma; pero esto, en nuestro razonamiento previo es el **flujo neto** en el punto x, es decir J_x. De acuerdo con la **ecuación de continuidad**, ya que el material no se crea o destruye,

$$\frac{d\phi(p,x,t)}{dt} + \frac{dJ_{x,t}}{dx} = 0$$
$$\frac{d\phi(p,x,t)}{dt} = -\frac{dJ_{x,t}}{dx}$$

Hasta ahora nos hemos manejado asumiendo que el movimiento de las frecuencias alélicas en las distintas poblaciones se debe exclusivamente al azar (por eso **random drift**), pero para hacer más general nuestra ecuación debemos entender que las frecuencias alélicas también pueden variar de forma sistemática, por ejemplo, a través de la mutación (fuera del equilibrio mutacional) y selección. La analogía con las partículas de tinta sería a través de suponer que estas tienen determinada carga eléctrica y que entre las dos barreras absorbentes existe una diferencia de potencial que hará que las partículas se muevan sistemáticamente en una dirección (efecto superpuesto al aleatorio). Otra analogía para intentar entender esta lógica de efectos superpuestos se plantea en la figura 3.15, en donde se observa el comportamiento de bolitas de diferente densidad al ingresar a un tubo de viento que cuenta a su vez con un sistema de ventilación interno.

Ahora, si llamamos $M(x)$ a ese efecto sistemático, tenemos entonces

$$J_{x,t} = M(x)\phi(p,x,t) - D\frac{d\phi(p,x,t)}{dx} \tag{3.47}$$

por lo que sustituyendo en la ecuación (3.6) tenemos

$$\frac{d\phi(p,x,t)}{dt} = -\frac{dJ_{x,t}}{dx} = -\frac{d}{dx}\left[M(x)\phi(p,x,t) - D\frac{d\phi(p,x,t)}{dx}\right] \tag{3.48}$$

Pero la derivación en forma más general de estas ecuaciones, directamente en nuestro dominio, así como las soluciones para casos particulares, la veremos con más detalle en las subsecciones que siguen.

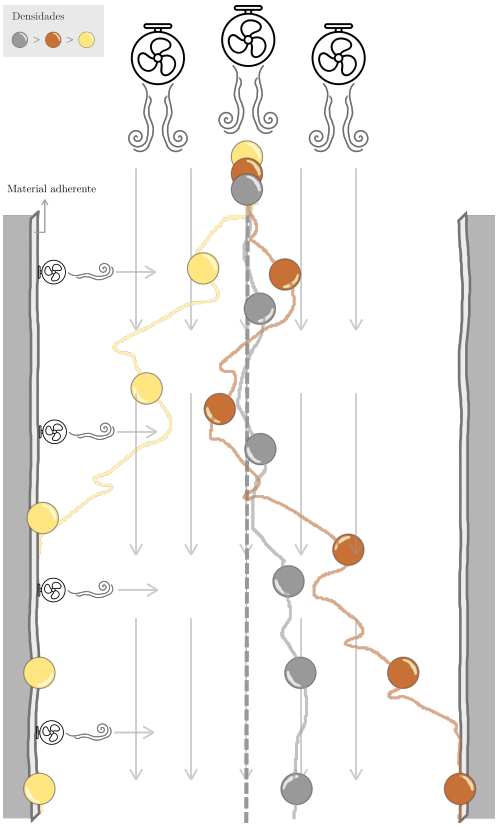


Figura 3.15. Diagrama que representa posibles trayectorias de bolitas de diferente densidad (por ejemplo, una pelotita de ping-pong, una de madera y una de metal) al ingresar a un tubo de viento con un sistema de ventilación interno. El sistema cuenta a su vez con un material adherente a sus costados, por lo que las bolitas que entren en contacto con el borde continuarán su trayectoria sobre él. Dos efectos superpuestos influyen sobre la trayectoria de cada bolita: *a)* el flujo de viento a través del tubo, generado por ventiladores superiores y *b)* una corriente de menor intensidad, de izquierda a derecha, generada por un sistema de ventilación interno. El efecto del flujo de viento superior es aleatorio respecto al plano horizontal, pudiendo cada bolita desplazarse a cada lado con igual probabilidad debido a este estímulo. En cambio, el efecto de la ventilación interna es sistemático, favoreciendo un movimiento de izquierda a derecha en cada bolita. Según su densidad, cada bolita responderá a estos estímulos de manera diferente. Notemos como en el ejemplo la bolita de ping-pong se inclinó tempranamente hacia la izquierda, lo cual favoreció su contacto con dicho borde. Las otras bolitas (de mayor densidad), presentan menores varianzas respecto a su movimiento horizontal.

Kolmogorov Forward Equation

Asumiendo que las poblaciones de nuestro **ensemble** son suficientemente grandes, la frecuencia de los alelos se transforma razonablemente en una variable continua que avanza suavemente en el tiempo, sin grandes saltos. En este sentido, la Kolmogorov Forward Equation (**KFE**) nos permite analizar el comportamiento en el tiempo de la función continua de probabilidad $\phi(p,x,t)$, es decir de la densidad de probabilidad de encontrar la frecuencia alélica en $0 < x < 1$ (notemos que es válida solo entre las **barreras absorbentes**, pero no para ellas, es decir en las poblaciones aún segregantes) en determinado tiempo t, partiendo de una frecuencia inicial p.

Para nuestro desarrollo, análogo a lo que hicimos en «Caminatas al azar y procesos de difusión» vamos a permitir pequeños desplazamientos de nuestras variables x y t, es decir Δx y Δt. Nuevamente, como vimos antes, hay dos fuerzas que pueden hacer que x cambie en un intervalo Δt: una fuerza sistemática que desplace las frecuencias alélicas siempre en una dirección (nuestro análogo era el campo eléctrico para las partículas de tinta) y que llamaremos $M(x)$ y una fuerza aleatoria, la deriva genética (que en nuestra analogía previa era la difusión por movimiento Browniano de las partículas de tinta) que llamaremos $V(x)$. Sin pérdida de generalidad, vamos a asumir que nuestro alelo favorecido es el **A**, cuya densidad de probabilidad (frecuencia) está dada por x (y la inicial del mismo por p) y que la fuerza sistemática es en el sentido creciente de x (o sea, en sentido de la fijación del **A**). En el cuadro que sigue tenemos detallados los posibles movimientos hacia la «posición» x en un intervalo Δt a partir del momento t en que estamos (recordar que en un intervalo Δt x solo puede moverse hacia abajo o hacia arriba, pérdida o fijación, una cantidad Δx), así como las probabilidades asociadas y la causa del movimiento.

Posición (generación t)	Prob. de esa frecuencia	Posibilidad de cambio a x en Δt	Prob. del cambio en Δt	Fuerza
$x - \Delta x$	$\phi(p,x-\Delta x,t)$	$x - \Delta x \to x$	$M(x - \Delta x)$	Sistemática
$x - \Delta x$	$\phi(p,x-\Delta x,t)$	$x - \Delta x \to x$	$\frac{V(x-\Delta x)}{2}$	Deriva
$x + \Delta x$	$\phi(p,x+\Delta x,t)$	$x + \Delta x \to x$	$\frac{V(x+\Delta x)}{2}$	Deriva
x	$\phi(p,x,t)$	xox	$1 - M(x) - V(x)$	Permanece

Para llegar a x en el próximo intervalo $(t + \Delta t)$, como el máximo recorrido posible en x es Δx debo estar a la distancia $\|\Delta x\|$ de x, es decir en $x - \Delta x$ o en $x + \Delta x$, o una tercera alternativa que es estar ya en x y quedarme en esa posición. Las probabilidades de estar en esas posiciones en el tiempo t son las que aparecen en la segunda columna del cuadro anterior. A cada una de ellas le corresponde una probabilidad de movimiento de x determinada, como aparece en la columna **4** del cuadro. Como asumimos que $M(x)$ va hacia la fijación, solo participaría en la probabilidad de movernos de $x - \Delta x$ a x (y no en el sentido contrario); esa es la razón porque $x - \Delta x$ aparece en dos filas del cuadro. La razón de por qué el efecto de la deriva es $\frac{V(x)}{2}$ es que se trata de una fuerza simétrica y, por lo tanto, la posibilidad de aumentar x es igual a la de disminuirla y entonces debemos partir a la mitad su efecto para cada lado. Por último, la probabilidad de que x se quede en donde está es el complemento de las otras probabilidades ($M(x)$ y $V(x)$), ya que al ser probabilidades el conjunto de todos los eventos excluyentes deben sumar 1.

Ahora, para analizar el cambio ocurrido entre $\phi(p,x,t+\Delta t)$ y $\phi(p,x,t)$, si multiplicamos la columna 2 por la **4**, sumamos y reordenamos los términos tenemos

$$\phi(p,x,t+\Delta t) - \phi(p,x,t) = -[M(x)\phi(p,x,t) - M(x-\Delta x)\phi(p,x-\Delta x,t)] +$$
$$\frac{1}{2}[V(x+\Delta x)\phi(p,x+\Delta x,t) - V(x)\phi(p,x,t)] - [V(x)\phi(p,x,t) - V(x-\Delta x)\phi(p,x-\Delta x,t)]$$

Pero el lado izquierdo de la ecuación anterior representa el cambio en ϕ (que podemos representar como $\Delta\phi$) a partir de un cambio en t (por lo tanto Δt). De la misma manera, el primer término del lado derecho de la ecuación representa el cambio en $M\phi$ para un determinado cambio en x (Δx). Finalmente, el segundo término es el cambio del cambio en $V\phi$; notemos que las dos diferencias en ese término (que representamos como $\Delta\Delta V\phi$) para un cambio en dos pasos de x (por lo tanto $\Delta\Delta x$). Poniendo todo junto

$$\frac{\Delta\phi(p,x,t)}{\Delta t} = -\frac{\Delta[M(x)\phi(p,x,t)]}{\Delta x} + \frac{1}{2}\frac{\Delta\{\Delta[V(x)\phi(p,x,t)]\}}{\Delta(\Delta x)} \tag{3.49}$$

Pasando al límite, cuando $\Delta t \to 0$ y $\Delta x \to 0$, la ecuación en diferencias converge a una ecuación diferencial parcial, que se conoce como **Kolmogorov Forward Equation** (o **KFE**)[15]

$$\frac{\partial\phi(p,x,t)}{\partial t} = -\frac{\partial[M(x)\phi(p,x,t)]}{\partial x} + \frac{1}{2}\frac{\partial\{\partial[V(x)\phi(p,x,t)]\}}{\partial(\partial x)} \tag{3.50}$$

$$\frac{\partial\phi(p,x,t)}{\partial t} = -\frac{\partial[M(x)\phi(p,x,t)]}{\partial x} + \frac{1}{2}\frac{\partial^2[V(x)\phi(p,x,t)]}{\partial x^2} \tag{3.51}$$

Retomando la biología del problema, antes mencionamos que $M(x)$ era una fuerza sistemática que llevaba la frecuencia del alelo **A** en una dirección, pero no especificamos mucho más. Esta fuerza podría representar sin problemas la selección, la mutación (o el balance mutacional) y la migración, alguna de ellas las veremos más adelante. Por otro lado, la fuerza aleatoria es igual (en nuestro modelo) a la varianza en las frecuencias alélicas luego de una generación de muestreo al azar en poblaciones de tamaño $2N$ y, por lo tanto, $V(x) = \frac{x(1-x)}{2N_e}$, con N_e tamaño efectivo según la varianza.

Kolmogorov Backward Equation

Una segunda posibilidad que se nos brinda de analizar lo que ocurre con la función de densidad $\phi(p,x,t)$ es **mirar hacia atrás en el tiempo** (de ahí el **Backward**) y ver lo que podría haber ocurrido con p, la frecuencia inicial (en lugar de x) en el primer instante en

[15] Debido al matemático soviético Andrey Nikolaevich Kolmogorov (25 de abril de 1903 - 20 de octubre de 1987). Si bien la ecuación hacia atrás (Kolmogorov Backward Equation) fue un aporte completamente novedoso, la ecuación hacia adelante ya era conocida para los físicos como la ecuación de Fokker-Planck.

el tiempo. Ahora, como todas las poblaciones del **ensemble** parten del mismo p, en ese pequeño incremento de tiempo Δt, los movimientos posibles serían $p - \Delta p$, $p + \Delta p$ o quedarse en p. Las probabilidades correspondientes a esos eventos las vemos en el siguiente cuadro.

Cambio 1.ª gen.	Prob. de ese cambio	Prob. de cambiar a x en $t - \Delta t$	Fuerza
$p \to p + \Delta p$	$M(p)$	$\phi(p + \Delta p,x,t - \Delta t)$	Sistemática
$p \to p + \Delta p$	$\frac{V(p)}{2}$	$\phi(p + \Delta p,x,t - \Delta t)$	Deriva
$p \to p - \Delta p$	$\frac{V(p)}{2}$	$\phi(p - \Delta p,x,t - \Delta t)$	Deriva
$p \to p$	$1 - M(p) - V(p)$	$\phi(p,x,t - \Delta t)$	Permanece

Para la movida $p \to p + \Delta p$ hay dos posibilidades, el efecto de la fuerza sistemática y de la aleatoria, por lo que su efecto combinado será $M(p) + \frac{V(p)}{2}$. Para la movida $p \to p - \Delta p$ solamente $\frac{V(p)}{2}$ y para quedarse en p por lo tanto $1 - M(p) - V(p)$. Ahora, como ya recorrimos un intervalo de tiempo Δt, el tiempo remanente para llegar al estado x en el presente será $t - \Delta t$ y, por lo tanto, las probabilidades asociadas a cada evento serán las que aparecen en la columna 3 del cuadro. Por ejemplo, para $p \to p + \Delta p$, entonces para llegar a x en el tiempo $t - \Delta t$ restante, la probabilidad es de $\phi(p + \Delta p,x,t - \Delta t)$. De la misma manera, para $p \to p - \Delta p$ la probabilidad es de $\phi(p - \Delta p,x,t - \Delta t)$ y finalmente para permanecer en p la probabilidad será de $\phi(p,x,t - \Delta t)$. Ahora, la diferencia entre $\phi(p,x,t)$ y $\phi(p,x,t - \Delta t)$ saldrá de la multiplicación de las columnas 2 y 3 del cuadro, para luego sumarlas, que luego de reordenar los términos nos permite llegar a

$$\phi(p,x,t) - \phi(p,x,t - \Delta t) = M(p)\left[\phi(p + \Delta p,x,t - \Delta t) - \phi(p,x,t - \Delta t)\right] +$$

$$\frac{V(p)}{2}\left\{\left[\phi(p + \Delta p,x,t - \Delta t) - \phi(p,x,t - \Delta t)\right] - \left[\phi(p,x,t - \Delta t) - \phi(p - \Delta p,x,t - \Delta t)\right]\right\}$$

Nuevamente, el lado izquierdo de la ecuación anterior representa el cambio en ϕ (que podemos representar como $\Delta\phi$) a partir de un cambio en t (por lo tanto Δt). Ahora, sin embargo, el primer término del lado derecho de la ecuación es $M(p)$ (que queda afuera del Δ) multiplicado por el cambio en ϕ ($M\Delta\phi$) para un determinado cambio en p (Δp). Notemos que ahora, en lugar de x es en p. Finalmente, el segundo término es ahora $V(p)$ multiplicado el cambio del cambio en ϕ, notemos las dos diferencias en ese término (que representamos como $\Delta\Delta\phi$) para un cambio en dos pasos de p (por lo tanto $\Delta\Delta p$). Poniendo todo junto

$$\frac{\Delta\phi(p,x,t)}{\Delta t} = M(p)\frac{\Delta\phi(p,x,t)}{\Delta p} + \frac{V(p)}{2}\frac{\Delta\{\Delta\phi(p,x,t)\}}{\Delta(\Delta p)} \tag{3.52}$$

Pasando al límite, cuando $\Delta t \to 0$ y $\Delta p \to 0$, la ecuación en diferencias converge a una ecuación diferencial parcial, que se conoce como **Kolmogorov Backward Equation** (o **KBE**)

$$\frac{\partial\phi(p,x,t)}{\partial t} = M(p)\frac{\partial\phi(p,x,t)}{\partial p} + \frac{V(p)}{2}\frac{\partial\{\partial\phi(p,x,t)\}}{\partial(\partial p)} \tag{3.53}$$

$$\frac{\partial\phi(p,x,t)}{\partial t} = M(p)\frac{\partial\phi(p,x,t)}{\partial p} + \frac{V(p)}{2}\frac{\partial^2\phi(p,x,t)}{\partial p^2} \tag{3.54}$$

Solución de las ecuaciones

En el equilibrio, la distribución de probabilidades no cambia con el tiempo, por lo que

$$\frac{\partial \phi(p,x,t)}{\partial t} = 0 \tag{3.55}$$

Cuando una función de este tipo existe la llamamos **distribución estacionaria**, ya que no cambiará con el tiempo, y como esta no depende del punto de partida tampoco (dentro del intervalo abierto de la difusión), la denotamos como $\phi(x)$ solamente. Cuando se cumple la ecuación (3.55), aplicándolo en la ecuación (3.51), tenemos

$$\frac{\Delta \phi(x)}{\partial t} = -\frac{\partial [M(x)\phi(x)]}{\partial x} + \frac{1}{2}\frac{\partial^2 [V(x)\phi(x)]}{\partial x^2} = 0 \ \therefore$$
$$\frac{\partial [M(x)\phi(x)]}{\partial x} = \frac{1}{2}\frac{\partial^2 [V(x)\phi(x)]}{\partial x^2}$$

Si ahora integramos (en x) ambos lados de esta ecuación, tenemos

$$2M(x)\phi(x) = \frac{\partial [V(x)\phi(x)]}{\partial x} \tag{3.56}$$

que se trata de una ecuación diferencial simple, tratable con técnicas de ecuaciones diferenciales ordinarias, que tiene como solución

$$\phi(x) = \frac{C}{V(x)G(x)} \tag{3.57}$$

con C una constante normalizadora tal que $\int_0^1 \phi(p,x,t)dx = 1$, de forma que la ecuación (3.57) pueda considerarse una función de densidad de probabilidad y G la integral indefinida (conocida como **función de escala**)

$$G(x) = exp[-2\int^x \frac{M(y)}{V(y)}dy] \tag{3.58}$$

Ahora, la pregunta que nos hacemos es ¿cómo se relaciona todo esto con nuestros problemas de genética de poblaciones? Veamos, por ejemplo, la determinación de la probabilidad de fijación de un alelo bajo deriva genética. Cuando al menos existe una **barrera absorbente** y esta es accesible, entonces no existe la distribución estacionaria (ya que en algún momento del tiempo todas las poblaciones terminarán en alguna de esas **barreras absorbentes**). Sin embargo, cuando existe más de una, conocer la probabilidad de terminar en una de ellas en particular suele ser importante y, en nuestro caso, conocer la probabilidad de terminar en la **barrera absorbente** $x = 1$, es decir la fijación, dado que partimos de una frecuencia p inicial es algo muy importante. Llamaremos $u(p,t)$ a la probabilidad de fijación en el tiempo t del alelo de interés (por ejemplo, A), cuando partimos de una frecuencia inicial p. Es posible demostrar que $u(p,t)$ es una función que cumple con la **KBE**. En particular, nos interesa conocer esa probabilidad cuando el tiempo $t \to \infty$, es decir cuando todas las poblaciones hayan fijado algún alelo y llamaremos $u(p) = \lim_{t\to\infty} u(p,t)$. En esas condiciones $u(p)$ no

variará con el tiempo y, por lo tanto, su derivada respecto al tiempo será igual a cero, por lo que la ecuación (3.54) quedará

$$\frac{\partial u(p)}{\partial t} = 0 = M(p)\frac{\partial u(p)}{\partial p} + \frac{V(p)}{2}\frac{\partial^2 u(p)}{\partial p^2} \tag{3.59}$$

A veces es posible encontrar esta misma ecuación escrita de la siguiente forma, que reduce un poco el aspecto intimidante de esta última ecuación

$$0 = M(p)\frac{du(p)}{dp} + \frac{V(p)}{2}\frac{d^2 u(p)}{dp^2} \tag{3.60}$$

al pasar la notación desde ∂ a d, lo que es posible ya que a esta altura nuestra $u(p,t)$ ya no depende más del tiempo (no cambiará en el tiempo) y entonces se trata de una función de una sola variable. Tenemos dos formas de ver $u(p)$, primero como la probabilidad última de fijación del alelo A, siendo que partimos desde una frecuencia inicial p; segundo, como la proporción de poblaciones en nuestro **ensemble** en los que finalmente el alelo A será fijado.

Volviendo a la ecuación (3.59), de acuerdo con Kimura (1962), esta tiene la siguiente solución

$$u(p) = \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx} \tag{3.61}$$

con $G(x)$ como fue definido previamente en la ecuación (3.58) y con las condiciones de frontera $u(0) = 0$ y $u(1) = 1$, que en términos llanos quiere decir que un alelo que no existe jamás será fijado y que un alelo fijado ya está fijado.

Finalmente, casi llegamos a donde queríamos. Vamos a determinar la probabilidad de fijación de un alelo que solo experimenta la fuerza de la deriva y, por lo tanto, $M(x) = 0$ y $V(x) = x(1-x)/(2N_e)$. Sustituyendo estos valores en la ecuación (3.58) tenemos que

$$G(x) = exp[-2\int^x \frac{M(y)}{V(y)}dy] = exp[-4N_e\int^x \frac{0}{x(1-x)}dy] = e^{-4N_e 0} = 1 \tag{3.62}$$

es decir que, bajo estas condiciones, $G(x) = 1$. Por lo tanto, sustituyendo este valor en la ecuación (3.61), tenemos ahora

$$u(p) = \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx} = \frac{\int_0^p 1dx}{\int_0^1 1dx} = \frac{p-0}{1-0} = p \tag{3.63}$$

La ecuación (3.63) nos brinda un resultado así de simple como de importante: la probabilidad de fijación de un alelo que solo experimenta deriva genética es igual a su frecuencia inicial en la población (o en el **ensemble** de poblaciones) de referencia, algo que usaremos en la próxima sección. Si lo pensamos un poco, este resultado tiene bastante sentido intuitivo a partir del comportamiento de las gráficas que vimos en la sección «El modelo de Wright-Fisher», en particular la figura 3.5: como se trata de un proceso aleatorio con la misma probabilidad de ir hacia arriba o hacia abajo en la figura, cuanto más cerca arrancamos del borde superior (la fijación del alelo), mayor será la probabilidad de que las poblaciones caigan en la **barrera absorbente** de fijación o, dicho de otra forma, una mayor proporción

de poblaciones irán a terminar en el estado de fijación. ¿Cuánto o cuántas? Bueno, en principio uno podría pensar que es proporcional al punto de arranque, p, lo que afortunadamente se ve confirmado por los cálculos (algo más complejos) precedentes.

PARA RECORDAR

Bajo la condición de que la deriva sea la única fuerza evolutiva actuando, en una población de N individuos diploides, en un locus con dos alelos:

- La **probabilidad de fijación** de un alelo es igual a su frecuencia p
- La ecuación hacia adelante de Kolmogorov (**KFE**) analiza el comportamiento de la función de densidad de probabilidad $\phi(p,x,t)$, es decir la distribución de las frecuencias alélicas x, a medida que avanza el tiempo y tiene la siguiente forma

$$\frac{\partial\phi(p,x,t)}{\partial t} = -\frac{\partial[M(x)\phi(p,x,t)]}{\partial x} + \frac{1}{2}\frac{\partial^2[V(x)\phi(p,x,t)]}{\partial x^2}$$

- La ecuación hacia atrás de Kolmogorov (**KBE**) es generalmente la más útil en genética de poblaciones. Analiza el comportamiento de la función de densidad de probabilidad $\phi(p,x,t)$ respecto a la frecuencia inicial p, a medida que vamos hacia atrás en el tiempo, y tiene la siguiente forma

$$\frac{\partial\phi(p,x,t)}{\partial t} = M(p)\frac{\partial\phi(p,x,t)}{\partial p} + \frac{V(p)}{2}\frac{\partial^2\phi(p,x,t)}{\partial p^2}$$

3.7 Probabilidad de fijación y tiempos de fijación

Como vimos antes, la probabilidad última de fijación de un alelo neutro es igual a la frecuencia inicial de dicho alelo, por ejemplo, p para el alelo A. Inicial, en este contexto, se refiere a cualquier punto en el tiempo que tomemos como referencia, lo que claramente irá variando a medida que avance el tiempo. Más aún, como también vimos antes, cuando la única acción evolutiva es la correspondiente a la deriva genética, el estado de equilibrio se corresponderá a cuando no existan poblaciones segregando, por lo que la probabilidad para cualquier punto entre las dos **barreras absorbentes** (fijación, $p = 1$, pérdida, $p = 0$) será cero. Pese a esto, hay varios resultados interesantes que surgen a partir de la **KBE** y que fueron derivados por Kimura y Ohta (1969). En particular, si bien existían resultados previos para el tiempo de segregación de un alelo (obtenidos a partir de un método diferente), es decir el tiempo hasta que uno de los dos estados absorbentes es alcanzado, no existía una solución para cada uno de los dos tiempos de forma independiente. En particular, asumiendo que el punto de partida es cuando la frecuencia del alelo A es p, el tiempo medio transcurrido hasta la fijación del alelo A ($\bar{t_1}(p)$, notemos la dependencia en p) viene dado por

$$\bar{t_1}(p) = -4N\frac{(1-p)\ln(1-p)}{p} \tag{3.64}$$

mientras que el tiempo medio hasta la pérdida del alelo A ($\bar{t_0}(p)$) es

$$\bar{t_0}(p) = -4N\frac{(p)\ln(p)}{(1-p)} \tag{3.65}$$

Por las dudas, los subíndices 0 y 1 se refieren en esta sección a pérdida ($p = 0$) y fijación ($p = 1$), respectivamente. La derivación de estas ecuaciones es un poco más compleja y excede el alcance de este capítulo (y seguramente tu paciencia), pero para los que estén interesados en cómo se llega hasta eso (a partir de los resultados de la sección previa) una excelente explicación se encuentra en el **Apéndice 1** del libro de Walsh y Lynch (2018), ecuaciones A1.21-A1.28. Volviendo a lo nuestro, los anteriores son los tiempos a la fijación y a la pérdida del alelo. Para calcular el tiempo medio hasta que ocurra alguno de los dos eventos ($\bar{t}(p)$), debemos ponderar cada uno de estos dos tiempos por la probabilidad de que ocurra, la fijación con probabilidad p y la pérdida con probabilidad $1 - p$, por lo que el promedio ponderado de ellos estará dado por $\bar{t} = p\bar{t_1}(p) + (1-p)\bar{t_0}(p)$. Sustituyendo los valores correspondientes a $\bar{t_1}(p)$ (ecuación (3.64)) y $\bar{t_0}(p)$ (ecuación (3.65)), tenemos

$$\bar{t}(p) = p\bar{t_1}(p) + (1-p)\bar{t_0}(p) =$$
$$p\left[-4N\frac{(1-p)\ln(1-p)}{p}\right] + (1-p)\left[-4N\frac{(p)\ln(p)}{(1-p)}\right]$$

$$\bar{t}(p) = -4N[p\ln(p) + (1-p)\ln(1-p)] \tag{3.66}$$

Las curvas correspondientes a los tres tiempos ($\bar{t}$,$\bar{t_0}$, $\bar{t_1}$) se pueden ver en la figura 3.16, como función de la frecuencia inicial del alelo A (p).

Mientras que en $\bar{t_0}$ y $\bar{t_1}$ los máximos tiempos se corresponden con la **proximidad inmediata** a barreras absorbentes (ahora discutiremos cuán cerca), en el caso de $\bar{t}$ el tiempo máximo ocurre cuando la frecuencia del alelo A es exactamente intermedia entre las barreras ($p = \frac{1}{2} = 0{,}5$). En particular, a esta frecuencia $p = \frac{1}{2}$, la ecuación (3.66) se transforma en

$$\bar{t}(p = \frac{1}{2}) = -4N[\frac{1}{2}\ln(\frac{1}{2}) + \frac{1}{2}\ln(\frac{1}{2})] = -4N[2\frac{1}{2}\ln(\frac{1}{2})] = -4N\ln(\frac{1}{2}) = 4N\ln(2)$$
$$\bar{t}_{max} \sim 2{,}77N$$

Es decir, si partimos de una frecuencia del alelo A de $p = 0{,}5$ el tiempo medio que llevará hasta la eventual fijación o pérdida del mismo es de $2{,}77N$ generaciones. Por otro lado, cuando aparece un nuevo alelo en la población su frecuencia es $\frac{1}{(2N)}$ ya que en la población hay N individuos diploides (y, por lo tanto, un total de $2N$ alelos). En este caso, **lo más cerca que podemos estar de su pérdida** antes de que ocurra, el tiempo que le llevará a la fijación será

$$\bar{t_1}(p = \frac{1}{2N}) = -4N\frac{(1-\frac{1}{2N})\ln(1-\frac{1}{2N})}{\frac{1}{2N}} \sim 4N \tag{3.67}$$

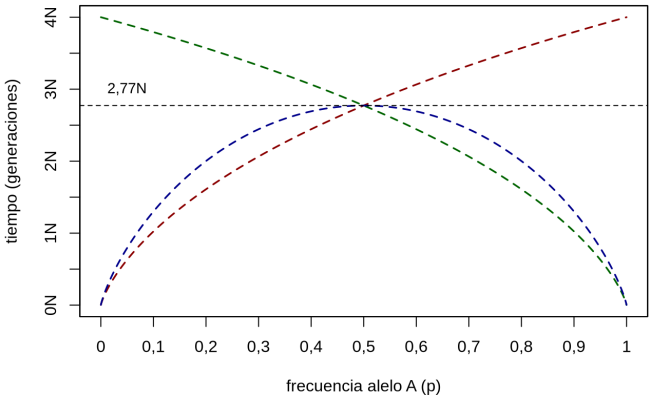


Figura 3.16. Tiempo medio a la fijación (verde), pérdida (roja) y de segregación (azul) en función de la frecuencia del alelo A. Cuando la frecuencia de A es $1/2$ el tiempo medio de segregación es máximo y su valor esperado es de aproximadamente 2,77N generaciones.

La última aproximación basada en que $\frac{\ln(1-x)}{x} \to 1$ cuando $x \to 0$ y que $1-\frac{1}{2N} = \frac{2N-1}{2N} \to 1$ cuando $N \to \infty$, en conjunto una buena aproximación aún para valores tan pequeños como $N = 10$. Es decir, al surgir un nuevo alelo en la población su probabilidad de fijación es $\frac{1}{2N}$ y en la eventualidad de que lo consiga el tiempo requerido será en promedio de $4N$ generaciones.

Finalmente, la probabilidad de perder ese mismo alelo nuevo es $1-\frac{1}{2N}$ y el tiempo promedio a la pérdida será de

$$\bar{t_0}(\frac{1}{2N}) = -4N\frac{(\frac{1}{2N})\ln(\frac{1}{2N})}{(1-\frac{1}{2N})} = -2\frac{\ln(\frac{1}{2N})}{(1-\frac{1}{2N})} \sim 2\ln(2N) \tag{3.68}$$

lo que es claramente menor que el tiempo medio a la eventual fijación ya que para $N > 0$, $2N \gg \ln(2N)$.

Ejemplo 3.5

En la mosca de la fruta (*D. melanogaster*) se han reportado varias mutaciones en el locus vestigial (*vg*) que alteran la forma de las alas de la mosca portadora. Consideremos una población de $N = 60$ individuos homocigotas para el alelo vg_1, el cual se asume produce alas normales. En un momento dado, se mezcla esta población con otra (de tamaño $N = 20$), compuesta por individuos homocigotas para el alelo vg_2, el cual genera una pequeña muesca en las alas de las moscas. Asumiendo que ambos alelos son neutros (es decir, la única fuerza evolutiva en juego es la deriva genética), ¿cuál es la probabilidad de fijación del alelo vg_2 en la población resultante? Si el tiempo generacional de *D. melanogaster* es de 10 días, ¿cuántos años en promedio se espera segregue el alelo en la población? ¿Cuántos años en promedio son necesarios para que se fije en la población?

Una vez se mezclan las poblaciones de moscas, las frecuencias alélicas p y q (para los alelos vg_1 y vg_2, respectivamente) son $p = \frac{120}{160} = \frac{3}{4}$ y $q = \frac{40}{160} = \frac{1}{4}$. Por lo tanto, la probabilidad de fijación del alelo vg_2 es de 0,25. El tiempo medio de segregación para el alelo está dado por

$$\bar{t}_{vg_2} = -4N[q \times ln(q) + (1-q) \times ln(1-q)]$$

$$\bar{t}_{vg_2} = -4 \times 80 \times [0{,}25 \times ln(0{,}25) + (1-0{,}25) \times ln(1-0{,}25)] \approx 180 \text{ generaciones}$$

Dado que cada generación ronda los diez días, se espera que el alelo segregue en promedio por

$$\frac{-4 \times 80 \times [0{,}25 \times ln(0{,}25) + (1-0{,}25) \times ln(1-0{,}25)] \text{ (generaciones)} \times 10 \text{ (días/generación)}}{365 \text{ (días/año)}}$$

$$\approx 4{,}9 \text{ años}$$

El tiempo de fijación medio está dado por

$$\bar{t_1}_{vg_2} = -4 \times N \times \frac{(1-q) \times ln(1-q)}{q}$$

$$\bar{t_1}_{vg_2} = -4 \times 80 \times \frac{(1-0{,}25) \times ln(1-0{,}25)}{0{,}25} \approx 276 \text{ generaciones}$$

$$\frac{-4 \times 80 \times \frac{(1-0{,}25) \times ln(1-0{,}25)}{0{,}25} \text{ (generaciones)} \times 10 \text{ (días/generación)}}{365 \text{ (días/año)}} \approx 7{,}6 \text{ años}$$

Para terminar con este tema, un hecho a primera vista curioso: si ya has tenido un curso de ecología, sin duda te estarás preguntando por la conexión entre la ecuación (3.66) que predice el tiempo medio que un locus con dos alelos se encuentra segregando en un **ensemble** de poblaciones con el índice de diversidad de Shannon-Weaver (a veces conocido como de Shannon o incluso Shannon-Wiener). Si recuerdas este índice, que nos indicaba la diversidad de especies en un sistema ecológico, este era $-\sum_{i=1}^{n} p_i \ln p_i$, para un conjunto de n especies diferentes, donde p_i es la frecuencia de la especie i en el conjunto ($\sum_i p_i = 1$). Si consideramos que $q = 1-p$ y $p+q = 1$, podemos llamar $p_1 = p$ y $p_2 = q$, que cumplen $\sum_i p_i = 1$. Más aún, con este cambio de notación podemos reescribir la ecuación (3.66) como

$$\bar{t}(p) = -4N[p\ln(p) + (1-p)\ln(1-p)] = -4N[p_1 \ln(p_1) + p_1 \ln(p_2)] = -4N \sum_i p_i \ln(p_i)$$

Excepto por el factor de escala $4N$, el índice de Shannon-Wiener para $n = 2$ especies es idéntico al tiempo medio de segregación, lo que llama inmediatamente a pensar, como paralelismo, en las distintas poblaciones segregando como la probabilidad de mantener la diversidad genética dentro de las mismas en el tiempo.

PARA RECORDAR

Bajo la condición de que la deriva sea la única fuerza evolutiva actuando, en una población de N individuos diploides, en un locus con dos alelos:

- La **probabilidad de fijación** de un alelo es igual a su frecuencia p
- El **tiempo a la eventual fijación** de ese alelo es $\bar{t_1}(p) = -4N\frac{(1-p)\ln(1-p)}{p}$
- La **probabilidad de perder un alelo** es igual a $1-p$
- El **tiempo a la eventual pérdida** de ese alelo es $\bar{t_0}(p) = -4N\frac{(p)\ln(p)}{(1-p)}$
- El **tiempo segregando** en ese locus, es decir el tiempo a eventual fijación o pérdida es $\bar{t}(p) = -4N[p\ln(p) + (1-p)\ln(1-p)]$
- El **tiempo máximo segregando** se alcanza cuando el punto de partida es $p = \frac{1}{2}$ y es igual a $\sim 2{,}77N$ generaciones.

3.8 El modelo coalescente

En general, el modelo de Wright-Fisher es una buena aproximación inicial para modelar la evolución de una población, pero las poblaciones reales se suelen apartar de varios de los supuestos, como ya vimos. Un tema del que nos hemos abstenido de tocar hasta ahora es el del número de alelos. Hasta ahora hemos tratado los problemas como un locus con dos alelos, pero se trata de un modelo poco realista desde el punto de vista de la evolución molecular. Es decir, si secuenciamos una región razonable de un gen en varios individuos de la población nos vamos a encontrar con que en ella existen varias posiciones segregantes (en las que hay variabilidad entre los individuos muestreados). Esto es muy razonable, ya que las mutaciones (que son la fuente primaria de variabilidad) ocurren aproximadamente al azar entre las diferentes posiciones posibles en el gen, por lo que en una muestra de secuencias del gen vamos a encontrar muchos alelos diferentes. La idea ahora es ver si podemos trazar y entender la historia **hacia atrás**, es decir, a partir de una muestra del presente extraer conclusiones sobre la historia de la población de la que extrajimos la muestra. Por ejemplo, nos podemos preguntar si a partir de los datos presentes podemos responder algunas de estas cuestiones: ¿Cuál fue el ancestro común más reciente de las copias genéticas existentes en el presente? ¿Cuál era el tamaño de la población en el momento del evento de coalescencia? ¿Con qué frecuencia se «extinguen» las copias genéticas? ¿Qué régimen migratorio operaba en la población histórica?

Una representación de la situación se puede ver en la figura 3.17, donde la generación presente (generación 0) aparece en la parte de abajo de la figura y a medida que nos retrotraemos en el tiempo, generación a generación, vamos subiendo en el gráfico.

En la generación presente, que es lo que usualmente tenemos para muestrear, tenemos 3 alelos diferentes, marcados por los colores (notemos que hay 3 colores: verde, magenta y rojo). Cada una de las copias presentes proviene de una copia en la generación anterior. Si conociésemos de cuál exactamente, la podríamos unir con una línea (técnicamente, con una **arista** en el **grafo** que representa la evolución de nuestra población de alelos en el tiempo). Notemos que el tamaño de la población de alelos es fija, 20 en nuestra figura. A medida que vamos hacia atrás, las copias de cada generación deben venir de alguna copia de la

generación anterior. Sin embargo, como el muestreo en el modelo de Wright-Fisher es con reposición, como vimos antes, con probabilidad $1/(2N)$ podemos volver a extraer una copia del mismo alelo para la nueva generación. Como el tamaño de la población de alelos es fijo, entonces si alguno de ellos es muestreado de forma repetida, algún otro deberá NO dejar copias, es decir, se perderá para el futuro. Si observas con detenimiento la figura verás que en muchas copias no hay una línea que las una con la generación inmediatamente debajo, lo que significa que esa copia no ha dejado descendientes y, por lo tanto, «muere» en nuestra genealogía de los alelos. Más aún, si vuelves a mirar con cuidado, seguramente notarás que varios linajes han desaparecido en el tiempo, por ejemplo, el naranja y el azul, y como en general no tenemos muestras del pasado no tendríamos capacidad de verlos o conocerlos hoy en día.

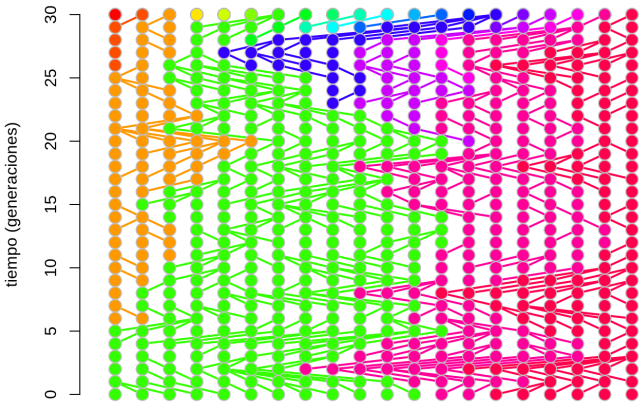


Figura 3.17. Evolución del linaje de los alelos encontrados en la generación presente (generación 0) hacia t generaciones atrás. Figura propia generada con el paquete learnPopGen (Liam Revell) de R (función modificada por nosotros).

Veamos si podemos entender cómo es la distribución de probabilidades asociada con el comportamiento de esta genealogía. De acuerdo con el modelo de Wright-Fisher, si tenemos $2N$ copias de alelos en cada generación, dado que extraemos una cualquiera (que viene de una copia en particular en la generación anterior), la probabilidad de una segunda copia que extraemos de la población provenga de la misma ancestral que la de la primera es $1/(2N)$, por lo que la probabilidad de que vengan de copias ancestrales distintas en la generación previa es de $1-[1/(2N)]$. Ahora, si deseamos saber cuál es la probabilidad de que la tercera copia que sacamos de la generación presente provenga de una copia diferente a las otras dos, esta copia debe ser primero diferente a la primera (con probabilidad $1-[1/(2N)]$), pero también a la segunda, ahora con probabilidad $1-[2/(2N)] = (2N-2)/(2N)$ (el 2 en el numerador de la expresión anterior es porque quedan $2N-2$ posibilidades de copias diferentes); como deben darse los dos eventos, entonces la probabilidad de que la tercera copia sea diferente a las dos anteriores (también diferentes), es decir que las 3 sean diferentes es $\left[\left(1-\frac{1}{2N}\right)\right] \times \left[\left(1-\frac{2}{2N}\right)\right]$. Claramente la misma lógica se puede aplicar para la cuarta copia, para la quinta y así sucesivamente.

Si (siguiendo la notación del libro de Joseph Felsenstein 2004) llamamos G_{kk} a la probabilidad de que k copias vengan de copias diferentes en la generación previa, entonces tenemos que

$$G_{kk} = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{2}{2N}\right)\left(1 - \frac{3}{2N}\right)...\left(1 - \frac{k-1}{2N}\right) = \prod_{i=1}^{k-1}\left(1 - \frac{i}{2N}\right) \quad (3.69)$$

Si operamos un poco, reagrupamos y descartamos los términos del orden $\frac{1}{N^2}$, tenemos entonces

$$G_{kk} = 1 - [1 + 2 + 3 + .. + (k-1)]/(2N) + \text{ términos en } \frac{1}{N^2} \text{ y menores} \quad (3.70)$$

Como $\sum_{i=1}^{k-1} = k(k-1)/2$,[16] entonces,

$$G_{kk} \approx 1 - \frac{k(k-1)}{4N} \quad (3.71)$$

o, visto de otra forma, la probabilidad de coalescer de dos copias en la generación anterior (que es igual a 1 menos la probabilidad de que vengan de distintas copias) es de

$$P(\text{coalescer}) = 1 - G_{kk} \approx \frac{k(k-1)}{4N} \quad (3.72)$$

Otra forma de ver lo anterior es que la probabilidad de coalescer es el producto de la probabilidad de extraer un alelo idéntico ($\frac{1}{2N}$), multiplicada por el número de posibilidades de a dos, que es $k(k-1)/2$ (despreciando la posibilidad de que 3 o más coalescan en la misma generación, un evento extremadamente improbable cuando $k \ll N$),

$$P(\text{coalescer}) = \frac{1}{2N} \times \frac{k(k-1)}{2} = \frac{k(k-1)}{4N} \quad (3.73)$$

Si recuerdas, la distribución geométrica era la que correspondía al número de fracasos hasta obtener el primer éxito. Tiene la forma $\Pr(X = k) = (1-p)^{k-1}p$, con parámetro p igual la probabilidad de éxito, media igual a $1/p$ y varianza $(1-p)/p^2$ (con soporte en $k \in \{1,2,3,...\}$). La distribución exponencial tiene forma $\lambda e^{-\lambda x}$, con soporte en $x \in [0,\infty)$, media $1/p$ y varianza $1/p^2$. Ahora, si llamamos u_k a la variable aleatoria para el tiempo (en generaciones) hasta la primera coalescencia, que me lleva de k a $k-1$ alelos, la misma tendrá una distribución geométrica (muy bien aproximada por una exponencial cuando $k \ll N$) y como $p = \frac{k(k-1)}{4N}$, entonces $\frac{1}{p} = \frac{4N}{k(k-1)}$, por lo tanto con media

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)} \text{ generaciones} \quad (3.74)$$

[16] Recordemos que $\sum_{i=1}^{i=n} i = \frac{n(n+1)}{2}$; si reemplazamos con $n = k-1$ llegamos al resultado mencionado.

y varianza

$$\mathbb{V}(u_k) = \frac{16N^2}{[k(k-1)]^2} \text{ generaciones}^2 \tag{3.75}$$

Es decir, tenemos la distribución del tiempo hasta el primer evento de coalescencia, que me reducirá el número de alelos diferentes en 1. Por esta razón, el siguiente tiempo de coalescencia tendrá también una distribución aproximadamente exponencial, pero ahora su esperanza será

$$\mathbb{E}(u_{k-1}) = \frac{4N}{(k-1)(k-2)} \text{ generaciones} \tag{3.76}$$

Como el denominador es menor que en (3.74), el tiempo esperado para el próximo evento de coalescencia será mayor. De hecho, si dividimos la ecuación (3.76) entre la ecuación (3.74), tendremos la relación entre los tiempos para el primer evento de coalescencia y para el segundo.

$$\frac{\mathbb{E}(u_{k-1})}{\mathbb{E}(u_k)} = \frac{4N}{(k-1)(k-2)} \frac{k(k-2)}{4N} = \frac{k}{k-2} \tag{3.77}$$

El proceso se repite y los tiempos serán cada vez mayores, hasta llegar al último evento de coalescencia que es cuando pasamos de $k = 2$ alelos a $k = 1$ alelo. Para este último evento tendremos entonces que la esperanza es igual a

$$\mathbb{E}(u_2) = \frac{4N}{2(2-1)} = \frac{4N}{2} = 2N \text{ generaciones} \tag{3.78}$$

Como se aprecia en la figura 3.18, en el caso de 5 alelos iniciales, los tiempos de los intervalos hasta el próximo evento de coalescencia van creciendo de acuerdo con la progresión $\frac{4N}{(5\times4)} = \frac{2N}{10} = \frac{N}{5}$, $\frac{4N}{(4\times3)} = \frac{2N}{6} = \frac{N}{3}$, $\frac{4N}{(3\times2)} = \frac{4N}{6} = \frac{2N}{3}$ y $\frac{4N}{(2\times1)} = \frac{4N}{2} = 2N$

Para una muestra conteniendo k alelos diferentes, el tiempo total a la coalescencia, o lo que es lo mismo el tiempo hasta el ancestro más reciente de todos los alelos en la muestra es de

$$t = 4N(1 - \frac{1}{k}) \tag{3.79}$$

con varianza

$$\mathbb{V}(t) = 4N^2 \sum_{i=2}^{k} \frac{1}{\binom{i}{2}^2} \tag{3.80}$$

(según Kingman 1982a; Kingman 1982b; Tajima 1983, ecuaciones 10 y 11).

Un primer resultado de la ecuación (3.79) es que a medida que el número de alelos se acerca a $2N$ (el máximo), como $\frac{1}{(2N)} \to 0$, el tiempo al ancestro común de todos los alelos tenderá a $4N$. Otro resultado más interesante es provisto por Rosenberg y Nordborg (2002), que

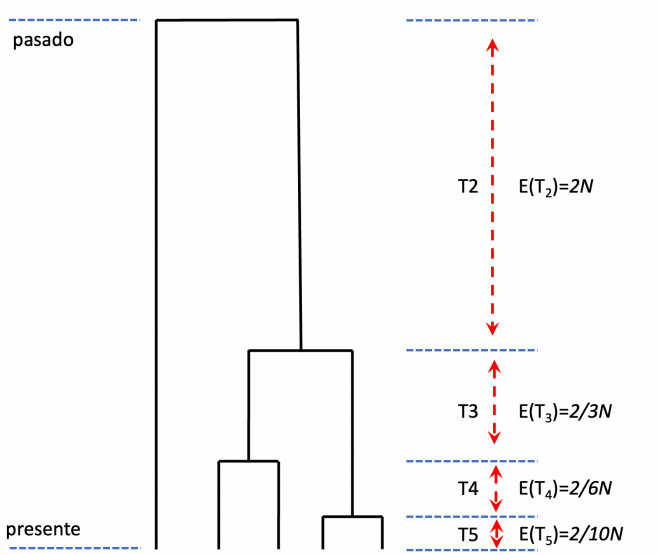


Figura 3.18. Árbol genético representando los tiempos del proceso coalescente, cuya esperanza está dada por $4N/(k(k-1))$ para k alelos. Con 5 alelos inicialmente el tiempo es $4N/20 = 2/10N$ y así sucesivamente. Se aprecia claramente que hacia atrás los tiempos son cada vez mayores. Elaboración propia a partir de idea en Hartl & Clark, 2007.

muestran que la probabilidad de que el ancestro común a los alelos en la muestra sea el mismo que el ancestro común a todos los alelos de la población se guía por la relación $\frac{(k-1)}{(k+2)}$, una relación que crece muy rápido (por ejemplo, para $k = 3$, $\frac{(3-1)}{(3+2)} = \frac{2}{5} = 0{,}40$, pero para $k = 10$ ya es $\frac{(10-1)}{(10+2)} = \frac{9}{12} = \frac{3}{4} = 0{,}75$). Esto, de alguna manera, nos da cierta garantía de que aún con una muestra relativamente pequeña es posible identificar correctamente los patrones evolutivos de la población.

Una de las aplicaciones del modelo coalescente tiene que ver con el supuesto de tamaño poblacional constante. Como la sucesión de intervalos de coalescencia (la esperanza) depende del tamaño poblacional, ecuación (3.74), una reducción del tamaño poblacional, por ejemplo, a la mitad, implicaría también reducir a la mitad el tiempo esperado o, lo que es lo mismo duplicar la velocidad de aparición de estos eventos. Esto tiene dos aristas. Por un lado, si de otra forma conseguimos reconstruir un árbol de alelos, por ejemplo, a partir de las secuencias y comparamos los dos árboles, el del coalescente y el molecular (sumado a otra información, usualmente biológica o histórica) podemos en algunos casos asignar estas discrepancias a fluctuaciones del tamaño poblacional. Por el otro lado, si conocemos alguna función del tamaño poblacional en función del tiempo, $N(t)$, y suponemos, por ejemplo, que el tiempo en una nueva escala ficticia τ pasa proporcional a $\frac{N(t)}{N(0)}$ (con $N(0)$ el tamaño poblacional actual), tenemos que para pequeños intervalo de tiempo ficticio

$$d\tau = \frac{N(0)}{N(t)} dt \tag{3.81}$$

por lo que

$$\tau = \int d\tau = \int \frac{N(0)}{N(t)} dt = N(0) \int \frac{1}{N(t)} dt \tag{3.82}$$

O sea, si conocemos una función que nos relacione el tamaño de la población en el tiempo y su inversa es integrable, entonces tenemos una forma de hacer nuestro coalescente independiente del tamaño poblacional. Veamos un ejemplo. El crecimiento exponencial de una población (por ejemplo, las poblaciones bacterianas al comienzo de la etapa de multiplicación) está dado por una función del tiempo de tipo $N_t = N_0 e^{gt}$, cuando el tiempo va hacia adelante. Ahora, en nuestro caso el tiempo va hacia atrás (en la generación 0, $N(0)$ es la población ya crecida al nivel actual y en tiempo t hacia atrás, $N(t)$ es mucho menor en tamaño), por lo que el exponente debe ser negativo

$$N(t) = N(0) e^{-gt} \tag{3.83}$$

Ahora, si sustituimos esta función del tamaño poblacional en (3.82) e integramos tenemos que

$$\tau = \int_0^t \frac{N(0)}{N(t)} dt = \int_0^t \frac{N(0)}{N(0)e^{-gt}} dt = \int_0^t e^{gt} dt = \frac{1}{g} e^{gt} \Big|_{t=0}^{t} = \frac{1}{g}(e^{gt} - 1) \tag{3.84}$$

Despejando t de la ecuación (3.84), tenemos que

$$t = \frac{\ln(1 + g\tau)}{g} \tag{3.85}$$

Ahora, para terminar alcanza con generar un modelo coalescente con tamaño fijo igual al inicial $N(0)$ y luego reescalar los tiempos de acuerdo la ecuación (3.85).

PARA RECORDAR

- En organismos diploides, en poblaciones de N individuos, la probabilidad de coalescer de dos alelos en la generación anterior es

$$\mathbb{P}\diagdown(\text{coalescer}) = \frac{k(k-1)}{4N}$$

- Como los eventos de coalescencia en cada generación son independientes de la previa, el tiempo a la coalescencia (u_k) tiene distribución geométrica, usualmente aproximada por una distribución exponencial con tasa inversa a la probabilidad de coalescer por generación, es decir

$$\mathbb{E}(u_k) = \frac{4N}{k(k-1)}$$

- Como además en cada evento de coalescencia k se reduce en uno, los tiempos hacia atrás serán cada vez más largos.
- El tamaño poblacional juega un papel muy importante en los tiempos de coalescencia (está en el numerador de la esperanza), por lo que si el mismo varía

a lo largo del tiempo entonces no se cumple uno de los supuestos fundamentales del coalescente. Sin embargo, en los casos en que el tamaño poblacional es una función del tiempo cuya inversa sea integrable, es posible aplicar una transformación del tiempo tal que se puede trabajar en esta escala temporal ficticia con el coalescente sin modificar.

3.9 Conclusión

En este capítulo, exploramos cómo la introducción de una limitación realista al modelo de Hardy-Weinberg (específicamente, el tamaño poblacional finito) lleva a esperar cambios en las frecuencias alélicas a lo largo del tiempo, simplemente por el muestreo azaroso de gametos. Este cambio en las frecuencias alélicas es en efecto un proceso evolutivo, y recibe comúnmente el nombre de ***deriva genética***. Lo que es más, vimos cómo el marco conceptual que nos dio el modelo de Hardy-Weinberg nos permite construir un nuevo modelo (el de Wright-Fisher), el cual nos permite estudiar este proceso evolutivo. También vimos que en un ensemble de poblaciones de tamaño finito, la tendencia observada es una disminución progresiva de la heterocigosidad total respecto a lo esperado bajo equilibrio. Esto ocurre porque se termina rompiendo el régimen de apareamiento aleatorio, un fenómeno en el que ahondaremos en más detalle en capítulos posteriores. A medida que pasa el tiempo en las poblaciones, la deriva lleva al aumento de la representación de alguno de los alelos en cada población, hasta que eventualmente algún alelo se fija en cada subpoblación. El modelo del coalescente nos brinda una herramienta para comprender cómo se da el proceso de aumento progresivo de copias alélicas en una población. Por último es importante resaltar que, si bien procesos netamente aleatorios (y, por lo tanto, ***neutrales*** desde el punto de vista evolutivo) pueden modificar las frecuencias alélicas, es esperable que en las poblaciones naturales también jueguen un papel importante los procesos sistemáticos (o ***selectivos***). Estos serán un próximo objeto de estudio en el siguiente capítulo. Modelos de mayor complejidad matemática, como la aproximación de difusión, nos permiten integrar los procesos estocásticos y los selectivos. Este es un buen ejemplo de cómo la integración de distintas elaboraciones teóricas no solo se puede llevar a cabo a nivel conceptual, sino también matemático, lo cual permite elaborar predicciones que interpelen y enriquezcan las elaboraciones teóricas originales.

3.10 Actividades

3.10.1 Control de lectura

1. ¿Qué es la deriva genética? ¿Cómo afecta a la evolución de las poblaciones?
2. La influencia del azar en la evolución de las frecuencias alélicas de una población se ve estrechamente ligada al tamaño de la población. Utiliza algunas de las ecuaciones presentadas en este capítulo para explicar esta dependencia.
3. ¿Qué es el coeficiente de fijación? Describe cómo este varía según el modelo de Wright-Fisher conforme transcurren las generaciones.
4. En poblaciones de tamaño se espera un aumento de la homocigosidad por efecto del azar, aumentando la frecuencia de algunos alelos a expensas de otros (que a largo plazo van desapareciendo). Según el modelo coalescente, ¿cuánto tiempo es de esperar para que todos los alelos presentes en una población sean réplicas de un mismo alelo ancestral?
5. ¿Qué es el tamaño poblacional efectivo (N_e), y qué utilidad tiene en el marco de la aplicación del modelo de Wright-Fisher? Nombra al menos una situación biológica en la que se posea una ecuación para obtener N_e.

3.10.2 ¿Verdadero o falso?

1. En una población de tamaño finito se espera un decrecimiento lineal de la heterocigosidad a medida que se suceden las generaciones. Esto se debe a que por efecto del muestreo aleatorio es de esperar a largo plazo que alguno de los alelos aumente su número, disminuyendo la frecuencia del resto (y, por lo tanto, la heterocigosidad).
2. Dos poblaciones pueden tener un mismo número de individuos y, sin embargo, el ritmo de la deriva genética puede ser bastante distinto en ambas.
3. Utilizando el modelo de Wright-Fisher podemos predecir de forma exacta la evolución de las frecuencias alélicas en una población.
4. Los tiempos de coalescencia no son iguales durante el proceso que lleva de una muestra de k alelos a su alelo ancestral.
5. Las ecuaciones de Kolmogorov nos permiten modelar cambios tanto hacia atrás (*Kolmogorov Backward Equation*) como hacia adelante (*Kolmogorov Forward Equation*), modelando en conjunto los efectos estocásticos y direccionales en la evolución de una población.

Solución

1. **Falso**. La frase es parcialmente cierta: sí se espera un decrecimiento de la heterocigosidad, y la explicación dada es razonable. No obstante, como se vio en el capítulo cuando se presentó el coeficiente de fijación, la heterocigosidad en la población decrece de forma aproximadamente exponencial (no de forma lineal, como se menciona).
2. **Verdadero**. El punto clave en la afirmación es que el ritmo de la deriva genética dependerá del tamaño poblacional efectivo (N_e), y no del tamaño censal (número de individuos). Un ejemplo claro podría ser el de dos poblaciones de ganado de igual tamaño, una de ellas con igual número de machos y hembras y la otra compuesta casi en su totalidad por hembras. ¿Es de esperar que la deriva genética tenga igual ritmo en ambas poblaciones? No, ya que la diversidad genética aportada generación

a generación en la población que cuenta con un único individuo macho es sustancialmente menor. Esto se verá reflejado en el número poblacional efectivo, el cual podemos utilizar para ajustar el número censal a los modelos teóricos que trabajan el efecto de la deriva genética.

3. **Falso.** Las trayectorias de una población de tamaño finito no pueden ser determinadas con precisión, justamente por la naturaleza estocástica del proceso. No obstante, un conjunto de poblaciones (ensemble) sí tendrá características que son predecibles desde un punto de vista estadístico: no podremos asegurar qué acontecerá en una población particular, pero sí qué tendencias esperamos ver en un conjunto de poblaciones de igual característica.
4. **Verdadero.** Según el modelo del coalescente, se espera que una muestra de k alelos haya provenido $4N$ generaciones de un único alelo ancestral. No obstante, $2N$ generaciones es el tiempo estimado de coalescencia de los últimos dos alelos, lo cual deja claro que el proceso de coalescencia no es homogéneo en el tiempo. Si se lo piensa, esto tiene bastante sentido: si tenemos en cuenta, por ejemplo, cien alelos, es claro que resultará más fácil que suceda el evento de coalescencia que cuando se consideran cincuenta, diez o los últimos dos alelos ancestrales.
5. **Falso.** Sí es cierto que las ecuaciones de Kolmogorov buscan modelar en conjunto los efectos estocásticos y direccionales en la evolución de una población. Sus nombres, sin embargo, no describen la capacidad de hacer predicciones hacia el pasado (como si se tratara de una especie de modelo del coalescente) o el futuro, si no que refieren a cómo se derivan matemáticamente las ecuaciones: en la *Kolmogorov Backward Equation* se recrea el proceso «mirando hacia atrás» en el tiempo y partiendo de la primera generación, intentando razonar y trabajar con las probabilidades de los sucesos que llevan al estado actual del sistema. En la *Kolmogorov Forward Equation* se trabaja partiendo del estado actual del sistema, e intentando razonar los cambios que pueden acontecer si «se avanza» una generación en el tiempo.

3.10.3 Ejercicios

Ejercicio 3.1

En una población con $N = 20$ individuos diploides, el alelo A_1 presenta 6 copias en una generación dada.

a. ¿Cuál es la probabilidad de que el alelo duplique su cantidad de copias en la siguiente generación?

b. El alelo A_2 presenta la mitad de copias en la generación. ¿Tiene la mitad de probabilidad de duplicar su número que la que tiene el alelo A_1?

Solución

a. La probabilidad de pasar de un número i de copias a un número j de copias en la siguiente generación está dada por

$$T_{ij} = \binom{2N}{j} p^j q^{2N-j}$$

(recordemos que $\binom{n}{k} = \frac{n!}{k!(n-k)!}$).

Tenemos $2N = 40$ alelos en la población, Para el alelo A_1 tenemos $p = \frac{6}{40} \Rightarrow q = \frac{34}{40}$, por lo que la probabilidad de pasar de $i = 6$ a $j = 12$ copias alélicas corresponde a

$$T_{6,12} = \frac{40!}{12! \times (40-12)!} \times (\frac{6}{40})^{12} \times (\frac{34}{40})^{28} \approx 0{,}00766$$

b. En el caso del alelo A_2, tenemos por razonamiento análogo una probabilidad de pasar de $i = 3$ copias a $j = 6$ copias en la próxima generación de

$$T_{3,6} = \frac{40!}{6! \times (40-6)!} \times (\frac{3}{40})^{6} \times (\frac{37}{40})^{34} \approx 0{,}0482$$

Como se puede apreciar, el alelo A_2 tiene claramente mayores probabilidades de duplicar su número que el alelo A_1.

Ejercicio 3.2

Una población de ratones tiene fluctuaciones en su tamaño poblacional. La población crece sin mayores problemas hasta alcanzar un tamaño aproximado de $N = 2000$; llegado este punto, una población de gatos se pone manos a la obra, disminuyendo a la población hasta $N = 50$ aproximadamente. Los ratones que logran pasar disimulados son los encargados de restablecer la población de ratones, cerrando el ciclo de la población. Asume que cada ciclo involucra a tres generaciones. ¿Cuál es el coeficiente de fijación para un locus bialélico, luego de que pasen 30 generaciones de ratones?

Solución

Notemos que en cada ciclo el tamaño poblacional efectivo corresponde a

$$\frac{1}{N_e} = \frac{1}{2} \times (\frac{1}{2000} + \frac{1}{50}) \Rightarrow N_e = \frac{2}{(\frac{1}{2000} + \frac{1}{50})}$$

La cantidad de generaciones considerada equivale a 15 de estos ciclos.

Tenemos a su vez que el coeficiente de fijación a tiempo t equivale a

$$F_t = 1 - (1 - \frac{1}{2N_e})^t$$

por lo que al tiempo pedido tenemos $F_{15} = 1 - (1 - \frac{1}{2 \times (\frac{2}{(\frac{1}{2000} + \frac{1}{50})})})^{15} \Rightarrow F_{15} \approx 0{,}0742$

Ejercicio 3.3

Una población de organismos diploides de tamaño N desconocido se estructuró geográficamente en cinco subpoblaciones de igual tamaño. En ellas se segrega un locus con dos alelos (A_1 y A_2, de frecuencias p_i y q_i respectivamente en la subpoblación i). Se estima que el evento de estructuración geográfica ocurrió alrededor de 3000 años atrás. Se genotipan 50 individuos al azar de cada población, observándose el siguiente conteo de individuos:

Genotipo	Subpobl. 1	Subpobl. 2	Subpobl. 3	Subpobl. 4	Subpobl. 5
A_1A_1	2	16	20	26	36
A_1A_2	16	26	22	19	13
A_2A_2	32	8	8	5	1

Estima el tamaño de la población inicial, asumiendo *i)* que las subpoblaciones mantuvieron un tamaño aproximadamente constante a lo largo del tiempo, *ii)* que el tiempo generacional de la especie es de 30 años aproximadamente, *iii)* que las subpoblaciones cumplen todas las suposiciones para asumir que se encuentran en equilibrio de Hardy-Weinberg (salvo, naturalmente, la suposición de tamaño infinito).

Solución

La varianza que se espera observar en el ensemble poblacional para la frecuencia de un alelo (elegimos arbitrariamente el alelo A_1) luego de t generaciones está dada por

$$\sigma_t^2(p) = t \times \frac{pq}{2N_e}$$

Notemos que podemos calcular la varianza observada para la frecuencia alélica p en el ensemble de subpoblaciones a partir de los datos de genotipado. Para ello debemos primero calcular las frecuencias alélicas p_i para cada subpoblación.

Así, por ejemplo, tenemos $p_1 = f_1(A_1A_1) + \frac{1}{2}f_1(A_1A_2) = (\frac{2}{50}) + \frac{1}{2} \times (\frac{16}{50}) = 0{,}2$. Por analogía, se tiene que $p_2 = 0{,}58$; $p_3 = 0{,}62$; $p_4 = 0{,}71$; $p_5 = 0{,}85$. Dado que se asume que las poblaciones tienen aproximadamente el mismo tamaño, el promedio de frecuencias alélicas equivale a $\bar{p} = \frac{1}{n}\sum_{i=1}^{i=n} p_i = \frac{0{,}2+0{,}58+0{,}62+0{,}71+0{,}85}{5} = 0{,}592$. Por lo tanto, la varianza observada es de

$$\sigma_t^2(p) = \frac{1}{n}\sum_{i=1}^{i=n}(p_i - \bar{p})^2$$

$$\sigma_t^2(p) = \frac{1}{5}\Big[(0{,}2 - 0{,}592)^2 + (0{,}58 - 0{,}592)^2 + (0{,}62 - 0{,}592)^2 + (0{,}71 - 0{,}592)^2 + (0{,}85 - 0{,}592)^2\Big]$$

$$\sigma_t^2(p) = 0{,}047016$$

Retomando nuestra primera ecuación, y teniendo en cuenta que pasaron $t = \frac{3000}{30} = 100$ generaciones (dado que se estima un tiempo generacional de alrededor de 30 años), tenemos

$$0{,}047016 = 100 \times \frac{0{,}592 \times (1 - 0{,}592)}{2N_e}$$

en donde se utilizó la estimación de $\bar{p}$ antes realizada.

De aquí podemos obtener un estimado del tamaño poblacional efectivo de las subpoblaciones observadas

$$0{,}047016 = 100 \times \frac{0{,}592 \times (1 - 0{,}592)}{2N_e} \Longrightarrow \frac{1}{N_e} = \frac{2 \times 0{,}047016}{100 \times 0{,}592 \times (1 - 0{,}592)}$$

$$N_e = \frac{100 \times 0{,}592 \times (1 - 0{,}592)}{2 \times 0{,}047016} \Longrightarrow N_e \approx 257$$

Por lo tanto, podemos estimar que el tamaño poblacional efectivo de la población inicial se encontraba en el orden de los 1285 individuos.

Ejercicio 3.4

Una población compuesta por 250 individuos diploides de la especie *Peluca mustensis* se encuentra segregando un locus bialélico (alelos A_1 y A_2 ; frecuencias alélicas denotadas por p y q, respectivamente).

Un genotipado de la población arroja los siguientes conteos:

$$\begin{cases} \text{Genotipo } A_1A_1 : 102 \text{ individuos} \\ \text{Genotipo } A_1A_2 : 96 \text{ individuos} \\ \text{Genotipo } A_2A_2 : 52 \text{ individuos} \end{cases}$$

Determina si la población se encuentra segregando el locus acorde a lo esperado bajo equilibrio de Hardy-Weinberg. De no ser así, estima el valor del coeficiente de fijación (F).

Solución

Se puede estimar las frecuencias alélicas p y q a partir de los conteos genotípicos presentados.

Tenemos que

$$\begin{cases} p = f(A_1A_1) + (\frac{1}{2})f(A_1A_2) = \frac{102}{250} + (\frac{1}{2})(\frac{96}{250}) = 0{,}6 \\ q = 1 - p \Rightarrow q = 0{,}4 \end{cases}$$

Para estas frecuencia alélicas se esperan las siguientes frecuencias genotípicas (y conteos en una población de $N = 250$ individuos):

Genotipo	Frecuencia esperada	Conteo esperado ($N = 250$)
A_1A_1	0,36 ($0{,}6^2$)	90 ($0{,}36 \times 250$)
A_1A_2	0,48 ($2 \times 0{,}6 \times 0{,}4$)	120 ($0{,}48 \times 250$)
A_2A_2	0,16 ($0{,}4^2$)	40 ($0{,}16 \times 250$)

Realizando un test de χ^2 con los conteos obtenidos, podemos establecer si las desviaciones observadas son estadísticamente significativas (en cuyo caso queda refutada la hipótesis nula de segregación en equilibrio de Hardy-Weinberg).

Recordando que $\chi^2 = \sum_{i=1}^{i=n} \frac{(\text{obs.} - \text{esp.})^2}{\text{esp.}}$, tenemos entonces $\chi^2 = \frac{(102-90)^2}{90} + \frac{(96-120)^2}{120} + \frac{(52-40)^2}{40} = 10$. Considerando 1 grado de libertad, tal desviación respecto de los valores esperados se espera observar con una probabilidad $< 0{,}01$ según la distribución de χ^2. Por lo tanto, se rechaza la hipótesis nula, según la cual el locus se encuentra segregando acorde a lo esperado bajo equilibrio de Hardy-Weinberg. Podemos estimar el coeficiente de fijación a partir de la desviación observada en la frecuencia de heterocigotas, considerando que $H_{obs} = 2pq(1 - F) \Rightarrow F = 1 - \frac{H_{obs}}{2pq}$. Reemplazando con los valores correspondientes, tenemos $F = 1 - \frac{96/250}{2 \times 0{,}6 \times 0{,}4} = 0{,}2$, siendo esta nuestra estimación para el coeficiente de fijación.

Σ Ejercicio 3.5

Como ya se ha planteado en este capítulo, existen varias situaciones donde las asunciones del modelo Wright-Fisher no se cumplen. La noción de tamaño poblacional efectivo (N_e) resulta sumamente útil en algunos de estos casos: en lugar de derivar un modelo teórico alternativo, se compara el comportamiento observado en el caso de estudio con el de una población ideal que sí cumple con las asunciones del modelo (y posee este tamaño poblacional).

a. Demuestra que el tamaño poblacional efectivo de una población dioica (con dos sexos), con posible desbalance en la proporción de sexos, equivale a $N_e = \frac{4N_mN_h}{N_m+N_h}$.

[**Pista:** bajo el modelo de Wright-Fisher la probabilidad de que dos alelos sean idénticos por ascendencia no tiene en cuenta la procedencia del alelo (es decir, si el alelo fue aportado a la generación por un macho o una hembra). En una población de organismos diploides dioicos esto debe tenerse en cuenta, ya que todos los organismos de la generación están formados por un gameto masculino y uno femenino.]

b. ¿Cuál es el N_e esperado para un locus asociado al cromosoma X?

c. ¿Cuál es el N_e esperado para un locus mitocondrial? Ten en cuenta que las mitocondrias son heredadas de forma exclusiva por línea materna.

Solución

a. En una población ideal la probabilidad de que dos alelos tomados al azar en una población sean idénticos por ascendencia es de $\frac{1}{2N}$; por lo tanto, en una población de tamaño poblacional efectivo N_e que se comporte acorde al modelo de Wright-Fisher esta probabilidad es de $\frac{1}{2N_e}$. Vayamos ahora al caso de una población diploide dioica, compuesta por N_m machos y N_h hembras (o los sexos análogos según el sistema de definición de sexos de la especie).

¿Cuál es la probabilidad de que dos alelos tomados al azar sean idénticos por ascendencia? Notemos que en este caso los organismos de cada generación no se forman tomando completamente al azar dos gametos de la generación anterior: por definición, cada organismo se conforma muestreando al azar un gameto aportado por un macho y uno aportado por una hembra. Por lo tanto, dos alelos tomados al azar de una generación pueden ser idénticos por ascendencia según dos eventos disjuntos: *i)* se muestrearon dos alelos aportados por un mismo macho de la generación anterior, o *ii*) se muestrearon dos alelos aportados por una misma hembra de la generación anterior. Por cómo están conformados los organismos (un alelo aportado por cada sexo), la probabilidad de muestrear dos alelos aportados por un organismo del mismo sexo en la generación anterior es de $\frac{1}{2} \times \frac{1}{2} = (\frac{1}{4})$. En este punto, la probabilidad de que el alelo muestreado sea el mismo en cada caso se obtiene con un razonamiento análogo al planteado cuando se desarrolló el modelo teórico: la probabilidad de que se trate del mismo alelo es o bien $\frac{1}{2N_m}$ (si los alelos fueron aportados por un macho), o bien $\frac{1}{2N_h}$ (si fueron aportados por una hembra). Estos son eventos disjuntos, por lo que las probabilidades deben ser sumadas para contemplar todos los casos de interés. Dado que buscamos encontrar el tamaño poblacional efectivo N_e tal que lo observado en la población se ajuste a lo esperado idealmente bajo el modelo de Wright-Fisher, tenemos que esta probabilidad y la del modelo teórico deben ser iguales cuando se asigna este tamaño poblacional:

$$\frac{1}{2N_e} = (\frac{1}{4}) \times \frac{1}{2N_m} + (\frac{1}{4}) \times \frac{1}{2N_h}$$

$$\frac{1}{2N_e} = \frac{N_h + N_m}{4 \times 2 \times N_m \times N_h}$$

$$\frac{1}{N_e} = \frac{N_h + N_m}{4N_m N_h} \Rightarrow \therefore N_e = \frac{4N_m N_h}{N_m + N_h}$$

b. Se puede aproximar el problema siguiendo una lógica similar a la planteada para el caso anterior. Debe tenerse en cuenta, no obstante, que en este caso los dos sexos aportan de alelos de forma desigual en la población: $\frac{1}{3}$ de los alelos son aportados por machos y $\frac{2}{3}$ de los alelos son aportados por las hembras. La probabilidad de muestrear en la población dos alelos provenientes de un macho es por lo tanto $(\frac{1}{3}) \times (\frac{1}{3}) = \frac{1}{9}$; la probabilidad de que a su vez se trate del mismo alelo en este caso es de $\frac{1}{N_m}$ (dado que los machos solo aportan N_m alelos). Análogamente, la probabilidad de muestrear dos alelos aportados por una hembra es de $(\frac{2}{3}) \times (\frac{2}{3}) = (\frac{4}{9})$; que se trate del mismo alelo es un evento de probabilidad $\frac{1}{2N_h}$ (dado que las hembras aportan $2N_h$ alelos).

Por lo tanto, tenemos en este caso

$$\frac{1}{2N_e} = (\frac{1}{9}) \times \frac{1}{N_m} + (\frac{4}{9}) \times \frac{1}{2N_h}$$

$$\frac{1}{2 \times N_e} = \frac{2N_h + 4N_m}{9 \times 2 \times N_m \times N_h}$$

$$\frac{1}{N_e} = \frac{2N_h + 4N_m}{9 \times N_m \times N_h} \Rightarrow \therefore N_e = \frac{9N_m N_h}{4N_m + 2N_h}$$

Nótese que la ecuación refleja la disparidad que tiene el número de individuos de cada sexo para los locus ligados al cromosoma X: una población de $N = 10$ con $N_m = 9$ y $N_h = 1$ se comportará acorde a lo esperado para un tamaño poblacional de $N_e = \frac{81}{38} \approx 2$, mientras que una población con $N_m = 1$ y $N_h = 9$ se comportará como una población de tamaño $N_e = \frac{81}{22} \approx 4$, cerca del doble de tamaño.

c. Dado que todo locus mitocondrial es heredado por línea materna, la probabilidad de que dos alelos sean idénticos por ascendencia es de $\frac{1}{N_h}$. Si asumimos paridad en el número de sexos (es decir, $N_h = \frac{N}{2}$), tenemos que

$$\frac{1}{2N_e} = \frac{1}{N_h} = \frac{1}{(N/2)} = \frac{2}{N}$$

$$\frac{1}{N_e} = \frac{4}{N} \Rightarrow \therefore N_e = \frac{N}{4}$$

Por lo tanto, un locus mitocondrial segrega acorde a lo esperado por el modelo de Wright-Fisher para un locus diploide en una población de tamaño cuatro veces menor a la población considerada.

Ejercicio 3.6

(Inspirado en una acotación de *Applied Population Genetics*, de Rodney J. Dyer, disponible en https://dyerlab.github.io/applied_population_genetics/index.html)

Los organismos del orden de insectos *Hymenoptera* (por ejemplo, abejas, avispas u hormigas, entre otros) presentan un sistema de ploidía mixto: las hembras son diploides, mientras que los machos son haploides.

Considera una población de abejas, compuesta por 10.000 machos y una hembra. En un momento dado se observa para un locus segregando en la población una heterocigosidad H_0 de 0,476. ¿Cuál es la heterocigosidad esperada luego de 25 generaciones? Asume, por simplicidad, que todos los machos pueden reproducirse con la abeja reina.

Solución

Notemos que el esquema de ploidías planteado para himenópteros es prácticamente idéntico al planteado para locus ligados al cromosoma X en humanos (uno de los sexos es diploide para el locus considerado, mientras que el otro es haploide). Por lo tanto, se espera que el locus segregue acorde a las expectativas de un locus diploide que se comporte según el modelo de Wright-Fisher con el siguiente tamaño poblacional efectivo (ver **Ejercicio 3.5**, donde se deriva la fórmula para N_e en estos casos):

$$N_e = \frac{9N_mN_h}{4N_m + 2N_h}$$

$$N_e = \frac{9 \times 10000 \times 1}{(4 \times 10000) + (2 \times 1)} \Rightarrow N_e \approx 2{,}250$$

(increíble, ¿no?).

Sabemos a su vez que $H_t \approx H_0 \times e^{\frac{-t}{2N_e}}$, por lo que se espera observar luego de 20 generaciones, en promedio, la siguiente heterocigosidad:

$$H_t \approx 0{,}476 \times e^{\frac{-20}{2 \times 2{,}25}} \approx 0{,}00872$$

Esto equivale a una reducción de la heterocigosidad a menos del 2 % del valor inicial. No olvidemos que estos cálculos solo tienen en cuenta el efecto esperado de la deriva genética sobre la frecuencia esperada de heterocigotas (además de la simplificación sobre el modo de reproducción de la especie). Por lo tanto, el cálculo no necesariamente refleja la variabilidad genética observada en poblaciones reales, donde también es esperable que juegue un rol la acción de otras fuerzas evolutivas.

Ejercicio 3.7

Considerando el coeficiente de fijación F, se tiene que la expectativa para la frecuencia de heterocigotas para un locus bialélico (alelos A_1 y A_2 con frecuencias p y q, respectivamente) en una población de tamaño finito corresponde a $H = 2pq(1 - F)$. Demuestra que las frecuencias esperadas para los genotipos homocigotas son $f(A_1A_1) = p^2(1 - F) + pF$ y $f(A_2A_2) = q^2(1 - F) + qF$.

Solución

Recordemos que las frecuencias alélicas se vinculan a las frecuencias genotípicas según

$$\begin{cases} p = f(A_1A_1) + \frac{1}{2}f(A_1A_2) \\ q = f(A_2A_2) + \frac{1}{2}f(A_2A_2) \end{cases}$$

Por lo tanto, tenemos que $f(A_1A_1) = p - \frac{1}{2}f(A_1A_2)$ y $f(A_2A_2) = q - \frac{1}{2}f(A_1A_2)$. Dado que $f(A_1A_2) = 2pq(1-F)$, tenemos entonces

$$f(A_1A_1) = p - 2(\frac{1}{2})pq(1-F)$$
$$f(A_1A_1) = p - p(1-p)(1-F)$$
$$f(A_1A_1) = p - (p - pF - p^2 + p^2F)$$
$$f(A_1A_1) = p - p + pF + p^2 - p^2F$$
$$f(A_1A_1) = p^2(1-F) + pF$$

Análogamente, tenemos

$$f(A_2A_2) = q - 2\frac{1}{2}pq(1-F)$$
$$f(A_2A_2) = q - (1-q)q(1-F)$$
$$f(A_2A_2) = q - (q - qF - q^2 + q^2F)$$
$$f(A_2A_2) = q - q + qF + q^2 - q^2F$$
$$f(A_2A_2) = q^2(1-F) + qF$$

Vemos entonces que todas las frecuencias genotípicas son corregidas por un factor de corrección de $(1-F)$ respecto a lo esperado bajo equilibrio de Hardy-Weinberg, pero en el caso de los genotipos homocigotas se suma a su vez un factor de pF o qF (según el genotipo), correspondiente al aumento esperado en la frecuencia de homocigotas.

Ejercicio 3.8

Como se mencionó en este capítulo, el tiempo medio transcurrido hasta la fijación o pérdida de un alelo está dado por la ecuación

$$\bar{t} = -4N[p \times ln(p) + (1-p) \times ln(1-p)]$$

donde N es el tamaño poblacional y p la frecuencia del alelo considerado.

La gráfica de esta función sugiere un máximo en $p = \frac{1}{2}$. Demuestra que esto es cierto.

Solución

Notemos que el máximo para la función se da cuando su derivada primera respecto a p equivale a cero. Tenemos entonces

$$\frac{\delta \bar{t}}{\delta p} = 0 \Rightarrow \frac{\delta[-4N[p \times ln(p) + (1-p) \times ln(1-p)]]}{\delta p} = 0$$

Dado que $-4N$ es constante respecto a p, esto equivale a considerar

$$\frac{\delta[p \times ln(p) + (1-p) \times ln(1-p)]}{\delta p} = \frac{\delta[p \times ln(p)]}{\delta p} + \frac{\delta[(1-p) \times ln(1-p)]}{\delta p} = 0$$

El primero de estos términos equivale a $\frac{\delta p}{\delta p} \times ln(p) + p \times \frac{\delta ln(p)}{\delta p}$; recordando que $\frac{\delta ln(f(x))}{\delta x} = \frac{\delta f(x)}{\delta x} \times \frac{1}{f(x)}$ (lo cual utilizaremos también al desarrollar el segundo término), tenemos que

$$\frac{\delta p \times ln(p)}{\delta p} = ln(p) + p \times \frac{1}{p} = ln(p) + 1$$

de forma similar, tenemos

$$\begin{aligned} \frac{\delta[(1-p) \times ln(1-p)]}{\delta p} &= \frac{\delta[(1-p)]}{\delta p} \times ln(1-p) + (1-p) \times \frac{\delta ln(1-p)}{\delta p} \\ &= -1 \times ln(1-p) + (1-p) \times \left(-\frac{1}{1-p}\right) \end{aligned}$$

por lo que

$$\frac{\delta[(1-p) \times ln(1-p)]}{\delta p} = -ln(1-p) - 1$$

Combinando ambos resultados, tenemos

$$\frac{\delta[p \times ln(p) + (1-p) \times ln(1-p)]}{\delta p} = ln(p) + 1 - ln(1-p) - 1 = 0$$

$$ln(p) - ln(1-p) = 0 \Rightarrow ln(p) = ln(1-p)$$

$$p = 1 - p \Rightarrow 2p = 1 \therefore p = \frac{1}{2}$$

con lo que queda demostrado que el máximo valor para $\bar{t}$ ocurre cuando $p = \frac{1}{2}$.

Capítulo 4

Selección natural

«Chi è quel pirla che ha parlato? Come Sarebbe a dire: Proviamo con l'albicocca?».
...
Perché l'è mea pö taant bèll... vèss fiöö del Guglielmo Tell
Perché me de quèla volta suun piò sta quell
E sun cunteent per el mè pà, che l'han fatto eroe nazionale
Ma da allora se vedo una mela comincio a stare male...
ma... le...[1]

Davide Van De Sfroos

En los capítulos anteriores vimos cómo las frecuencias alélicas de las poblaciones evolucionan, simplemente por las «leyes de la herencia», es decir el modelado matemático de los mecanismos moleculares y de los efectos **estocásticos** asociados al muestreo de los gametos en poblaciones de tamaño finito. En otras palabras, hasta ahora no hemos incorporado en nuestros modelos el efecto que la ventaja de algún alelo o genotipo pudiese tener sobre los otros. Como veremos en este capítulo, esto último constituye la base de la selección, el tema de nuestro capítulo actual. En particular, en este capítulo sentaremos las bases de la selección natural, aquella que opera en las poblaciones naturales y dejaremos para otros capítulos en la «Parte III: Genética cuantitativa» el tema de la selección artificial, que, si bien se funda en las mismas bases, su análisis tendrá otros componentes en cuenta que aún no hemos visto.

Bajo cualquier tipo de selección los genotipos ya no tendrán la misma oportunidad de pasar sus gametos a la siguiente generación y como las frecuencias de los alelos están vinculadas a las de los genotipos, por lo tanto, ya no podremos seguir asumiendo que estas frecuencias simplemente siguen en promedio a las de la generación anterior. Central a este capítulo, por tanto, es el concepto de ***fitness***, con el cual nos introduciremos a la temática.

[1] Frase que pertenece a la canción *Il figlio de Guglielmo Tell* (*El hijo de Guillermo Tell*) y que podría traducirse desde el Laghée (variante del Lombardo) como «"¿Quién es el idiota que habló? ¿Cómo que... 'probemos con un damasco'?" ... Porque no es tan lindo, no, ser el hijo de Guillermo Tell. Porque desde aquel día ya no soy el mismo. Y estoy contento por mi papá, que lo hicieron héroe nacional, pero desde entonces, cuando veo una manzana, empiezo a sentirme mal... muy mal». Claramente, la selección natural parece no haber operado aquí para preservar a los genotipos con mayor aversión al riesgo.

OBJETIVOS DEL CAPÍTULO

☐ Presentar y analizar el concepto de selección natural, proceso que genera cambios sistemáticos en las frecuencias alélicas de las poblaciones naturales.

☐ Distinguir entre distintos regímenes de selección natural y sus posibles consecuencias biológicas.

☐ Plantear un modelo matemático explícito que permita plantear de forma precisa la dinámica de las frecuencias alélicas bajo estos distintos regímenes de selección natural.

☐ Analizar y modelar la interacción entre la selección natural y otras fuerzas relevantes en la evolución, tales como la mutación y la deriva genética.

☐ Discutir en qué influye (y en qué no) la magnitud de la intensidad de la selección natural en la evolución de las poblaciones.

4.1 El concepto de *fitness*

En cualquier población natural, los individuos no dejan el mismo número de descendientes que alcanzan la edad reproductiva con éxito y menos aún que son capaces de generar a su vez el mismo número de crías que sus padres. Es decir, existe una variabilidad importante en el número de descendientes que dejan los individuos. Una parte de esta variabilidad es puramente **estocástica**, es decir, producto del azar. Por ejemplo, individuos que mueren antes de la edad reproductiva por eventos dependientes del azar (semillas que caen en terrenos completamente infértiles, animales o plantas con «mala suerte», etcétera). Otra parte de la variabilidad en el éxito reproductivo es en cambio sistemática: si los animales con un genotipo particular en un locus determinado son más exitosos en promedio en el número de descendientes que alcanzan la edad reproductiva, entonces dejarán más descendientes que el resto y en función de que esto haga crecer la frecuencia de su genotipo con relación a los otros entonces tenemos una base genética para la variabilidad. Claramente, la ventaja de unos genotipos sobre otros suele estar asociada al **ambiente**, considerando el término *ambiente* como algo genérico, tiempo y espacio donde y cuando viven nuestros individuos, por lo que usualmente vamos a restringir nuestros desarrollos a lo que consideramos como un **ambiente estable** (a los efectos de evitar que cambien las presiones selectivas durante nuestro análisis).

Por un tema de conveniencia general nosotros vamos a utilizar el término inglés ***fitness***, ya que su traducción más usual al castellano como **aptitud** deja a nuestro entender bastante que desear. Para ser más claros en nuestros conceptos veamos unas definiciones que nos ayudarán a entender la relación entre las diferentes formas del **fitness**, conceptos que desarrollaremos en la siguiente sección. El **fitness absoluto** es la tasa de aumento o crecimiento que predice el número absoluto de individuos de un genotipo determinado en una población a lo largo del tiempo. En otras palabras, se trata del número medio de descendientes

de un **tipo** determinado, por cada progenitor **del mismo tipo** (por ejemplo, el tipo podría ser el genotipo o el alelo, en otros casos).

Por otro lado, para el **fitness relativo** existen dos definiciones diferentes, cada una con sus ventajas, con diferencias en el denominador, es decir en el **relativo a qué**. En las secciones que siguen usaremos preferencialmente la siguiente: es la tasa de crecimiento de los genotipos (cuando consideramos genotipos, por ejemplo) en relación con un genotipo elegido como estándar de comparación (a menudo el genotipo con la mayor aptitud absoluta). Normalmente se simboliza con la letra w en los modelos en los que el tiempo se representa en generaciones discretas (conocida a veces como **selección darwiniana**) y como m en los modelos en los que el tiempo es continuo (a veces llamada **selección malthusiana**, por Thomas Malthus[2]). Otra definición de relativo es la que divide los **fitness absolutos** entre el **fitness medio** ($\bar{w}$), que ocasionalmente usaremos con ventaja sobre la primera definición.

Por último, el **fitness medio** es simplemente el promedio de los **fitness relativos**, es decir la suma de los valores de **fitness relativos** ponderados por la frecuencia de cada genotipo de la población, es decir que para k genotipos, $\bar{w} = \sum_{i=1}^{k} w_i p_i$.

4.2 Selección natural en el modelo de un locus con dos alelos

Las causas del éxito relativo entre los genotipos pueden ser varias. Por ejemplo, los distintos genotipos pueden estar asociados a distintas viabilidades de los cigotos. Luego de esto, durante el crecimiento las viabilidades pueden también ser diferentes, ya sea por la habilidad para crecer más rápido que la de otros genotipos, por ser más resistentes a condiciones ambientales prevalentes, por ser más adversos al riego o por ser menos llamativos como presas, por ejemplo. Ya crecidos los individuos y en plena etapa reproductiva, el éxito puede estar condicionado a varias cosas, de acuerdo con la forma de reproducción, física y etológica de la especie. Por ejemplo, en algunas especies los machos más llamativos (y a menudo más expuestos como presas) son los preferidos por las hembras. En otros casos, los machos disputan su posición privilegiada como reproductores mediante combates, asegurándose el apareamiento con tantas hembras como les sea posible, aunque los machos perdedores suelen encontrar también oportunidades para aparearse. Finalmente, la fertilidad es otro parámetro fundamental para la diferencia en el éxito reproductivo de algunos genotipos, en parte determinado por la acción de genotipos y en otro caso por el éxito o viabilidad de los alelos (por ejemplo, en gametos).

Para estudiar la influencia de la selección en la evolución de las frecuencias de los alelos vamos a asumir que nos encontramos en un locus autosomal con dos alelos, en un organismo diploide y vamos a asumir que no hay diferencias en la viabilidad de los gametos (que en principio está directamente asociado a un **fitness** de alelos). El modelo fundamental que utilizaremos de acá en adelante es el que aparece en la figura 4.1. Los individuos recién nacidos tendrán una **viabilidad** diferencial (probabilidad de supervivencia hasta la edad adulta-reproductiva) que será la base de nuestra **selección**. En efecto, todo el proceso

[2]Thomas Robert Malthus (13 de febrero de 1766 - 29 de diciembre de 1834) fue un clérigo y erudito británico cuyo ensayo *An Essay on the Principle of Population* fue fundamental para el desarrollo de la demografía. En lo que hace a nuestro curso, su aporte fundamental es el modelo que plantea un crecimiento geométrico de las poblaciones, aproximado por un crecimiento exponencial.

en una generación se puede resumir en el cuadro azul de la figura 4.1. Los recién nacidos serán el producto (de la frecuencia) de los gametos en la generación anterior. Luego, durante todo el período hasta llegar a la etapa reproductiva estarán sujetos a la selección, con diferentes viabilidades asociadas a los genotipos en nuestro locus de interés. Esto provocará que los individuos por genotipo que contribuyen al **pool de gametos** no sean en idénticas proporciones a los recién nacidos. Dicho de otra forma, es como si aplicásemos un filtro diferente a cada uno de los genotipos, entre los recién nacidos y los adultos, que cambiarán nuestras frecuencias de genotipos y posiblemente de los alelos (ya que están asociados a los fenotipos).

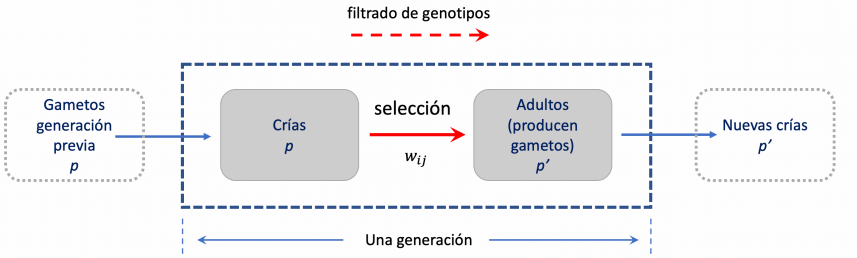


Figura 4.1. Modelo fundamental utilizado para estudiar la selección en un locus con dos alelos. Los gametos de la generación anterior (frecuencia p) producirán las crías de la generación presente, que luego del proceso de selección determinarán que adultos participarán del nuevo pool gamético (con frecuencia p'). Elaboración propia sobre idea de John H. Gillespie (2004).

Podemos llamar a cada uno de los **filtros** anteriores la **viabilidad** de cada genotipo. Obviamente, las nuevas frecuencias de los genotipos antes y después de la selección deberán sumar a uno, ya que se trata de proporciones de un total y no aparecen nuevos genotipos. Si, siguiendo la notación de John H. Gillespie (2004), llamamos w_{ij} a la viabilidad del genotipo con alelos i y j (pudiendo ser $i \neq j$ o $i = j$), entonces la frecuencia del genotipo A_iA_j después de la selección estará dada por el producto de su frecuencia antes de la selección (por ejemplo, p_ip_j si los genotipos se encontraban en equilibrio Hardy-Weinberg), multiplicada por el **filtro** w_{ij} y dividido por el **fitness medio**, de forma que el **fitness relativo** de cada genotipo, ponderado por su frecuencia, sume a uno. Poniendo en un cuadro la situación antes y después de la selección, tenemos

Genotipo	A_1A_1	A_1A_2	A_2A_2
Frecuencia	p^2	$2pq$	q^2
Viabilidad	w_{11}	w_{12}	w_{22}
Frecuencia post-selección	p^2w_{11}	$2pqw_{12}$	q^2w_{22}
Frecuencia relativa post-selección	$\frac{p^2w_{11}}{\bar{w}}$	$\frac{2pqw_{12}}{\bar{w}}$	$\frac{q^2w_{22}}{\bar{w}}$

En el cuadro anterior $\bar{w}$ representa el **fitness promedio** de la población y es igual a

$$\bar{w} = p^2 w_{11} + 2pqw_{12} + q^2 w_{22} \tag{4.1}$$

que no es más que el promedio de los **fitness** de cada genotipo[3]. Además, se cumple $\frac{p^2 w_{11}}{\bar{w}} + \frac{2pqw_{12}}{\bar{w}} + \frac{q^2 w_{22}}{\bar{w}} = \frac{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}}{\bar{w}} = 1$, como era esperado.

Después del proceso de selección. las frecuencias de los gametos serán las correspondientes a los genotipos de la última línea del cuadro anterior y, por lo tanto, la nueva generación tendrá esas frecuencias alélicas, ya que no suponemos ningún proceso de filtrado adicional (por ejemplo, selección cigótica o supervivencia fetal diferencial). Los genotipos con alelos A_1 son A_1A_1 y A_1A_2, de los cuales el primero todos son alelos A_1 y la mitad del segundo también son A_1, por lo que sumando las frecuencias de esta forma obtenemos la frecuencia de A_1 en la siguiente generación

$$p' = \frac{p^2 w_{11}}{\bar{w}} + \frac{1}{2}\frac{2pqw_{12}}{\bar{w}} = \frac{p^2 w_{11} + pqw_{12}}{\bar{w}} \tag{4.2}$$

El cambio ocurrido entre generaciones, que siguiendo la notación de John H. Gillespie (2004) llamaremos $\Delta_s p$ para reforzar la idea de que el cambio se debe a la selección, será la frecuencia actual del alelo A_1, es decir p' (notemos el $'$ para la nueva generación), menos la frecuencia en la generación anterior (o sea, p).

$$\Delta_s p = p' - p = \frac{p^2 w_{11} + pqw_{12}}{\bar{w}} - p = \frac{p^2 w_{11} + pqw_{12} - p\bar{w}}{\bar{w}} \tag{4.3}$$

Si trabajamos un poco el numerador, teniendo en cuenta además que $pq = p(1-p) = p - p^2$ y que $1 - 2p = (1-p) - p = q - p$, tenemos para el numerador de la ecuación (4.3)

$$\begin{aligned} p^2 w_{11} + pqw_{12} - p\bar{w} &= p[pw_{11} + qw_{12} - p^2 w_{11} - 2pqw_{12} - q^2 w_{22}] \\ = p[(p-p^2)w_{11} + q(1-2p)w_{12} - q^2 w_{22}] &= pq[pw_{11} + (q-p)w_{12} - qw_{22}] \\ = p[pqw_{11} + q(1-2p)w_{12} - q^2 w_{22}] &= pq[pw_{11} + (q-p)w_{12} - qw_{22}] \\ = pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})] & \end{aligned} \tag{4.4}$$

Si ahora llevamos este resultado de nuevo a la ecuación (4.3), tenemos una forma simplificada

$$\Delta_s p = \frac{pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}} \tag{4.5}$$

[3] Recuerda que el promedio de una variable es la suma de los valores, w_{11},w_{12},w_{22} multiplicados por sus frecuencias relativas respectivas, p^2, $2pq$ y q^2 en nuestro caso.

Ejemplo 4.1

Una población de *Thrips tabaci* (*trips* de la cebolla), un insecto que constituye una importante plaga agrícola, se encuentra reproduciéndose en un cultivo de cebolla. El locus A codifica para una enzima, la cual es blanco de acción de un plaguicida. En la población se encuentran segregando dos alelos para el locus, el alelo A_1 y el alelo A_2. El primero de estos codifica para una variante sobre la cual el plaguicida es altamente efectivo; el alelo A_2, en cambio, codifica para una variante sobre la cual el plaguicida no tiene efecto. Se asume que la población se encuentra cumpliendo los supuestos del equilibrio de Hardy-Weinberg.

Bajo el régimen del plaguicida, los individuos dejan según su genotipo **5**, **8** o **15** descendientes por cada 10 individuos.

a. ¿Cuál es el **fitness** relativo de cada genotipo?

b. Si en la generación inmediatamente anterior a que se comience a utilizar el plaguicida las frecuencias alélicas son $f(A_1) = 0{,}8$ y $f(A_2) = 0{,}2$, ¿qué frecuencias alélicas se esperan en la siguiente generación?

a. Notemos que el promedio de descendencia es de 0,5, 0,8 y 1,5 para los individuos de genotipo A_1A_1, A_1A_2 y A_2A_2, respectivamente. Estos constituyen los valores de **fitness** absoluto de cada genotipo.

Como se mencionó anteriormente, tenemos al menos dos definiciones de **fitness** relativo, las cuales varían respecto a qué se toma como referencia. La más usual define al **fitness** relativo normalizando respecto al genotipo de mayor **fitness** (en este caso, el genotipo A_2A_2).

Siguiendo esta definición, tenemos los siguientes **fitness** relativos:

$$\begin{cases} w_{A_1A_1} = \frac{0{,}5}{1{,}5} \\ w_{A_1A_1} = \frac{0{,}8}{1{,}5} \\ w_{A_2A_2} = \frac{1{,}5}{1{,}5} = 1 \end{cases}$$

También podemos definir el **fitness** relativo con base en el **fitness** promedio de la población. Este es $\bar{w} = \sum_{i=1}^{i=k} w_i p_i$, donde p_i es la frecuencia de cada genotipo. Bajo equilibrio de Hardy-Weinberg estas frecuencias son
$f(A_1A_1) = 0{,}8^2;\ f(A_1A_2) = 2 \times 0{,}8 \times 0{,}2;\ f(A_2A_2) = 0{,}2^2$. Tenemos entonces
$\bar{w} = 0{,}5 \times 0{,}8^2 + 2 \times 0{,}8 \times 0{,}2 \times 0{,}8 + 0{,}2^2 \times 1{,}5$. Tenemos entonces los **fitness** relativos

$$\begin{cases} w_{A_1A_1} = \frac{0{,}5}{0{,}5 \times 0{,}8^2 + 2 \times 0{,}8 \times 0{,}2 \times 0{,}8 + 0{,}2^2 \times 1{,}5} \\ w_{A_1A_1} = \frac{0{,}8}{0{,}5 \times 0{,}8^2 + 2 \times 0{,}8 \times 0{,}2 \times 0{,}8 + 0{,}2^2 \times 1{,}5} \\ w_{A_2A_2} = \frac{1{,}5}{0{,}5 \times 0{,}8^2 + 2 \times 0{,}8 \times 0{,}2 \times 0{,}8 + 0{,}2^2 \times 1{,}5} \end{cases}$$

b. El cambio en la frecuencia del alelo A_1 está dado por

$$\Delta_s f(A_1) = \frac{f(A_1)f(A_2) \times [f(A_1) \times (w_{A_1A_1} - w_{A_1A_2}) + f(A_2) \times (w_{A_1A_2} - w_{A_2A_2})]}{\bar{w}}$$

En este caso, tenemos

$$\Delta_s f(A_1) = \frac{0{,}8 \times 0{,}2 \times [0{,}2 \times (0{,}5 - 0{,}2) + 0{,}2 \times (0{,}8 - 1{,}5)]}{0{,}5 \times 0{,}8^2 + 2 \times 0{,}8 \times 0{,}2 \times 0{,}8 + 0{,}2^2 \times 1{,}5}$$

Recordando que $\Delta_s f(A_1) = f'(A_1) - f(A_1) \Rightarrow f'(A_1) = f(A_1) + \Delta_s f(A_1)$, por lo que

$$f'(A_1) = 0{,}8 + \frac{0{,}8 \cdot 0{,}2 \cdot \left[0{,}2(0{,}5 - 0{,}2) + 0{,}2(0{,}8 - 1{,}5)\right]}{0{,}5 \cdot 0{,}8^2 + 2 \cdot 0{,}8 \cdot 0{,}2 \cdot 0{,}8 + 0{,}2^2 \cdot 1{,}5} = 0{,}8 - 0{,}02012579 \approx 0{,}77987$$

Esto a su vez implica $f'(A_2) = 1 - f'(A_1)$. Vemos que, como era de esperar, bajo el régimen del plaguicida se da una disminución del alelo A_1 y un consiguiente aumento de la frecuencia del alelo A_2 (que confiere fenotipo de resistencia).

Ya veremos que es posible hacer una transformación de la escala que simplifique aún más la ecuación (4.5), pero el numerador de la misma parece bastante informativo desde ya. El término pq, que es igual a $p(1-p) = p - p^2$ y que aparece multiplicando todo, y ya vimos que alcanza su máximo[4] cuando $p = q = \frac{1}{2}$. También vimos en el capítulo anterior que pq era el numerador de la varianza en frecuencias por deriva. El otro término, entre paréntesis rectos, es el promedio, ponderado por las frecuencias de p y q, respectivamente, de las diferencias en **fitness** entre homocigotos A_1A_1 y heterocigotos y entre estos últimos y los homocigotos A_2A_2.[5]

Veamos si podemos avanzar en la simplificación de todo esto. Como nuestro interés hasta ahora (en este capítulo) es ver cómo el proceso de selección afecta los genotipos y las frecuencias alélicas **dentro** de una población, podemos transformar todos los **fitness** de su escala absoluta a una escala relativa. Por ejemplo, podemos dividir todos los valores entre el mayor valor de **fitness** (asumamos para el ejemplo que se trata dew_{11}). Ahora los **fitness relativos** para los tres genotipos serán $\frac{w_{11}}{w_{11}} = 1$, $\frac{w_{12}}{w_{11}}$ y $\frac{w_{22}}{w_{11}}$. Más aún, si llamamos s al **coeficiente de selección** (su intensidad) contra el genotipo A_2A_2 (sin pérdida de generalidad, ya que la designación de los alelos A_1 y A_2 es completamente arbitraria y podemos elegir a nuestro gusto cuál es uno y cuál el otro) y definimos un parámetro h de tal forma de poder acomodar libremente los valores de **fitness** de los heterocigotos, tenemos que los nuevos fitness relativos estarán dados ahora por

Genotipo	A_1A_1	A_1A_2	A_2A_2
Frecuencia	p^2	$2pq$	q^2
Viabilidad	w_{11}	w_{12}	w_{22}
Viabilidad relativa	$\frac{w_{11}}{w_{11}} = 1$	$\frac{w_{12}}{w_{11}}$	$\frac{w_{22}}{w_{11}}$
Fitness relativo (forma h y s)	1	$1 - hs$	$1 - s$

En nuestro caso, vamos a permitir que $0 \leqslant s \leqslant 1$, donde $s = 0$ corresponde al caso en que no hay ninguna fuerza detrimental contra los genotipos que no son A_1A_1 y $s = 1$ corresponde al caso de **inviabilidad total** del genotipo A_2A_2. El parámetro h, llamado usualmente **efecto del heterocigoto**, es un parámetro que permite ubicar libremente a los heterocigotos entre (y aún por fuera) de los valores de **fitness relativo** de los homocigotos. En efecto, jugando con el parámetro h podemos entrar en los modos de acción génica más relevantes, como se aprecia en el cuadro siguiente

[4] Recuerda que, si hacíamos la derivada y la igualábamos a cero, teníamos los máximos/mínimos de la función. En este caso, $\frac{d(p-p^2)}{dp} = 1 - 2p = 0 \therefore p = \frac{1}{2}$).

[5] Recuerda que w_{11}, w_{12} y w_{22} están todos divididos entre $\bar{w}$, lo que los hace «relativos»).

Modo de acción	Valores de h
A_1 dominante, A_2 recesivo	$h = 0$
A_2 dominante, A_1 recesivo	$h = 1$
Dominancia incompleta	$0 < h < 1$
Sobredominancia	$h < 0$
Subdominancia	$h > 1$

Veamos cómo funciona lo anterior. Si $h = 0$, entonces los **fitness relativos** serán 1 para el genotipo A_1A_1, $1 - hs = 1 - 0 \times s = 1$ para el genotipo A_1A_2 y $1 - s$ para el genotipo A_2A_2, por lo que el genotipo heterocigoto tendrá el mismo valor de **fitness** que el genotipo A_1A_1. Esto es equivalente a un modo de acción con el alelo A_1 dominante sobre el alelo A_2. De la misma manera, si $h = 1$, entonces los **fitness relativos** serán 1 para el genotipo A_1A_1, $1 - hs = 1 - 1 \times s = 1 - s$ para el genotipo A_1A_2 y $1 - s$ para el genotipo A_2A_2, por lo que el genotipo heterocigoto tendrá el mismo valor de **fitness** que el genotipo A_2A_2. Esto es equivalente a su vez a un modo de acción con A_2 dominante sobre A_1. Un caso menos extremo, pero paradigmático también, es el caso en que $h = \frac{1}{2}$; en ese caso, los **fitness relativos** serán 1, $1 - \frac{1}{2}s$ y $1 - s$, con lo que el genotipo heterocigoto estará exactamente en el punto medio de los genotipos homocigotos ($\frac{1+(1-s)}{2} = 1 - \frac{1}{2}s$). En este caso el modo de acción es completamente aditivo (algo sobre lo que abundaremos en el capítulo «El modelo genético básico»). Una representación gráfica de los modos de acción génica descritos anteriormente se puede ver en la figura 4.2, donde las alturas de las barras representan los **fitness relativos** de cada genotipo en las 5 configuraciones del cuadro anterior.

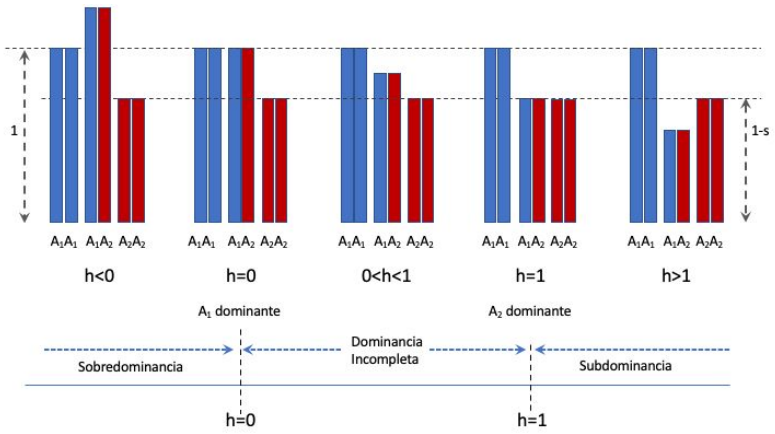


Figura 4.2. Representación gráfica de los modos de acción génica (o alélica) de acuerdo con el parámetro h (asumiendo que s es positivo). En cada caso se representan los 3 genotipos (A_1A_1, A_1A_2 y A_2A_2) y las alturas de las barras representan el **fitness relativo** de cada genotipo. Las barras en azul corresponden al alelo A_1, mientras que las barras rojas corresponden al alelo A_2.

Los casos particulares de $h = 0$ y $h = 1$ merecen una mención aparte. En el caso de $h = 0$, al tipo de selección se lo conoce como **selección contra el recesivo** ya que el heterocigoto contará con un valor igual al del homocigoto con mayor , dejando al otro homocigoto con

el alelo desfavorecido. Esto llevará gradualmente a la disminución de su frecuencia, al principio en forma rápida si el alelo desfavorable es frecuente, disminuyendo luego la velocidad. Por otro lado, al caso en que $h = 1$ se lo conoce como **selección contra el dominante**, ya que ahora el alelo dominante es el desfavorable, lo que llevará a un incremento del alelo recesivo.

Veamos ahora si logramos llegar al cambio de frecuencia esperada en p (es decir, $\Delta_s p$) con la reparametrización de nuestro modelo fundamental. Para ello basta con sustituir en la ecuación (4.5) w_{11}, w_{12} y w_{22} por sus valores relativos (por ejemplo, 1, $1 - hs$ y $1 - s$), lo que nos da

$$\Delta_s p = \frac{pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]}{\bar{w}} = \frac{pq[p(1 - (1 - hs)) + q((1 - hs) - (1 - s)]}{\bar{w}}$$

$$= \frac{pq[p(1 - 1 + hs)) + q(1 - hs - 1 + s]}{\bar{w}} \therefore$$

$$\Delta_s p = \frac{pqs[ph + q(1 - h)]}{\bar{w}} \tag{4.6}$$

A su vez, el **fitness medio** vendrá ahora dado por sustituir los valores w_{11}, w_{12} y w_{22} en la ecuación (4.1) por 1, $1 - hs$ y $1 - s$

$$\bar{w} = p^2 w_{11} + 2pqw_{12} + q^2 w_{22} = p^2 \cdot 1 + 2pq(1 - hs) + q^2(1 - s) = (p^2 + 2pq + q^2) - 2pqhs - q^2 s \therefore$$

$$\bar{w} = 1 - 2pqhs - q^2 s \tag{4.7}$$

Volvamos a la ecuación (4.6) y veamos si podemos comprender de forma intuitiva su significado. El denominador de la ecuación es $\bar{w}$ y es en general positivo, ya que $p^2 + 2pq + q^2 = 1$ y generalmente $2pqhs < 2pq$, lo mismo que $sq^2 < q^2$ (recordar que $0 \leqslant s \leqslant 1$ y usualmente $s \ll 1$). Además, pqs siempre es positivo por definición, ya que $0 \leqslant p \leqslant 1$, $0 \leqslant q \leqslant 1$ y $0 \leqslant s \leqslant 1$, por lo tanto $pqs \geqslant 0$. Es decir, $\frac{pqs}{\bar{w}} \geqslant 0$ y, por lo tanto, el signo de $\Delta_s p$ está dado por el signo de $ph + q(1 - h)$. Dicho de otra forma, la frecuencia de p crecerá ($\Delta_s p$ positivo) cuando $ph + q(1 - h) > 0$ y decrecerá cuando $ph + q(1 - h) < 0$. Una primera conclusión importante de lo anterior es que el sentido del cambio en la frecuencia de p entre generaciones **no depende** de s.

La ilustración de la superficie de $\Delta_s p$ en función de la frecuencia del alelo A_1 (p) y del valor del parámetro h pueden verse en la figura 4.3. Claramente existen regiones donde $\Delta_s p$ toma valores negativos y la forma de esta función de p cambia también en función del valor de h, por lo que estudiaremos los distintos casos en forma separada.

PARA RECORDAR

- El **fitness medio** de una población está dado por $\bar{w} = p^2 w_{11} + 2pqw_{12} + q^2 w_{22}$.
- El cambio en la frecuencia del alelo A_1 debido a la selección en un locus con dos alelos, luego de una generación es igual a $\Delta_s p = \frac{pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]}{\bar{w}}$.

- Otra forma de describir el cambio en la frecuencia del alelo A_1, en una parametrización que depende del efecto de los heterocigotos h y del coeficiente de selección s es $\Delta_s p = \frac{pqs[ph+q(1-h)]}{\bar{w}}$

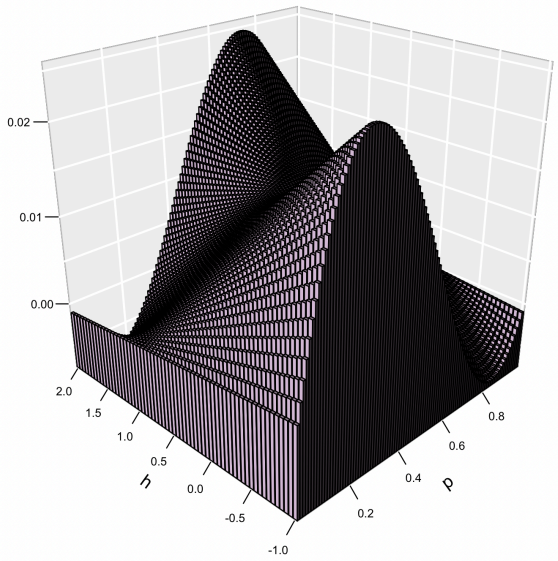


Figura 4.3. Valores de la función $\Delta_s p$ en función de la frecuencia del alelo A_1 (p) y del valor del parámetro h, en el rango completo de h, es decir $-1 < h < 2$. A diferencia del comportamiento en la región restringida $0 < h < 1$, se aprecian regiones del gráfico donde $\Delta_s p$ es negativo, lo que lleva a que la evolución de la frecuencia de p no necesariamente sea en una sola dirección.

4.3 Diferentes formas de selección

La forma en que parametrizamos el cambio en la frecuencia del alelo A_1, o sea $\Delta_s p$, en función de un parámetro h (**efecto del heterocigoto**) y de la frecuencia misma del alelo (p), nos va a permitir estudiar diferentes formas de selección que producirán distintos resultados en la distribución de genotipos en la población, de acuerdo con el modo de acción y la frecuencia de partida. En la figura 4.4 vemos los tres tipos de selección que trataremos, a partir de nuestro modelo fundamental de un locus con dos alelos. Arriba, la selección direccional, en la que la población se desplazará hacia la fijación del alelo favorecido. En el medio, la selección estabilizadora, en la que existe un óptimo en frecuencia intermedia de p y que además es un equilibrio estable. Abajo, la selección disruptiva, en la que el equilibrio de frecuencia intermedia es inestable y cualquier perturbación llevará a la población hacia la fijación o pérdida de uno de los alelos.

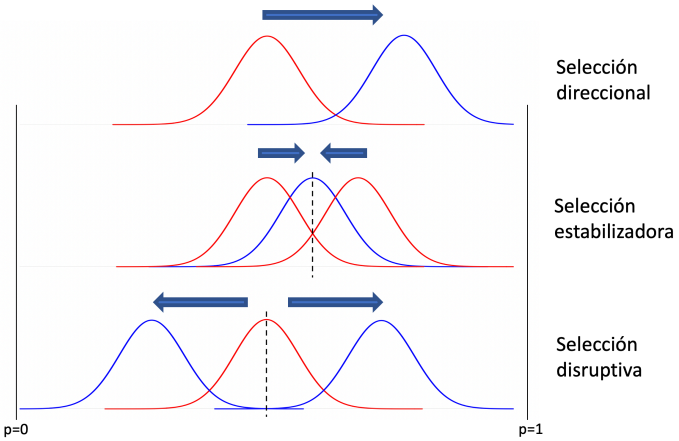


Figura 4.4. Tipos de selección que trataremos, a partir de nuestro modelo fundamental. En rojo las poblaciones antes de la acción de la selección, en azul después de la selección. Arriba, selección direccional, en la que la población se desplaza en una dirección, sin otro equilibrio que la fijación del alelo favorecido. En el medio, selección estabilizadora, en el que existe un punto óptimo en una frecuencia intermedia y el equilibrio es estable. Abajo, selección disruptiva, en la que existe un óptimo inestable en una frecuencia intermedia pero que cualquier pequeña perturbación lleva a que la población se vaya a la fijación o pérdida.

4.3.1 Selección direccional

Cuando existe **dominancia incompleta**, es decir cuando $0 < h < 1$ o, dicho de otra forma el heterocigoto tiene un **fitness** intermedio al de los homocigotas, entonces la expresión $ph + (1-p)(1-h)$ es siempre mayor a cero ya que para que existan dos alelos $0 < p < 1$ (ídem para $q = 1 - p$, notemos que $p = 0$ o $p = 1$ no tienen sentido). Teniendo en cuenta lo que discutimos antes respecto a que el signo de $\Delta_s p$ está determinado y es igual al signo de $ph + (1-p)(1-h)$, como este último en el caso de **dominancia incompleta** es siempre **positivo**, entonces $\Delta_s p$ también será **siempre positivo.** Al ser siempre positivo esto quiere decir que independientemente del punto de partida en la frecuencia del alelo A_1, la misma siempre aumentará. La conclusión lógica de esto es que se irá aproximando a la frecuencia máxima del alelo A_1, a la que nunca alcanzará en tiempo finito (claro que en determinado momento esta cercanía entra dentro del alcance típico de los efectos de la deriva, que podrán terminar de fijar el alelo A_1).

La representación gráfica de la superficie definida por $\Delta_s p$ en función del parámetro h, en la región $0 < h < 1$ y de la frecuencia del alelo A_1 se aprecia en la figura 4.5. Se aprecia con claridad que todos los valores de $\Delta_s p$ son positivos en la región considerada para h, sin depender además del punto en que consideremos la frecuencia de A_1 (es decir, p).

Más aún, para un valor de h determinado, cualquiera en la región $0 < h < 1$, Δp pasa a ser una función exclusivamente en p. Si observamos con cuidado la figura, un valor **fijo** de h, pongamos x es como hacer un corte con un plano $h = x$, paralelo al borde derecho de la figura ($h \to 1$) y en lugar de fijarnos en la superficie vamos a atender a la curva que forma $\Delta_s p$. Veamos que ocurre a distintos valores del parámetro h, siempre dentro del intervalo $0 < h < 1$.

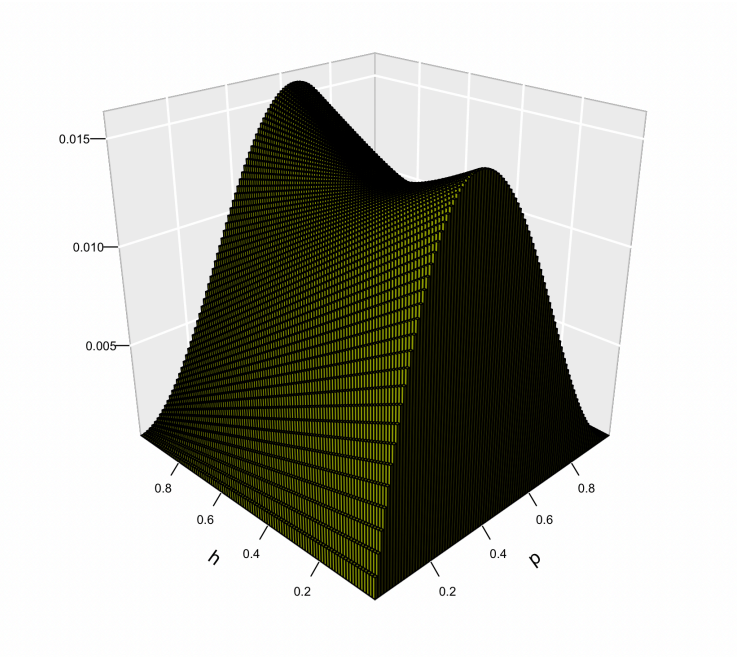


Figura 4.5. Valores de la función $\Delta_s p$ en función del valor del parámetro h y de la frecuencia del alelo A_1 (p). Claramente se aprecia que en el rango $0 < h < 1$ (que es el que está graficado) $\Delta_s p$ es siempre positivo, indicando un crecimiento de p. A su vez, se aprecia que como función de la frecuencia p, $\Delta_s p$ es una parábola con un máximo en $p = q = \frac{1}{2}$, mientras que como función de h tiene un mínimo en $h = \frac{1}{2}$

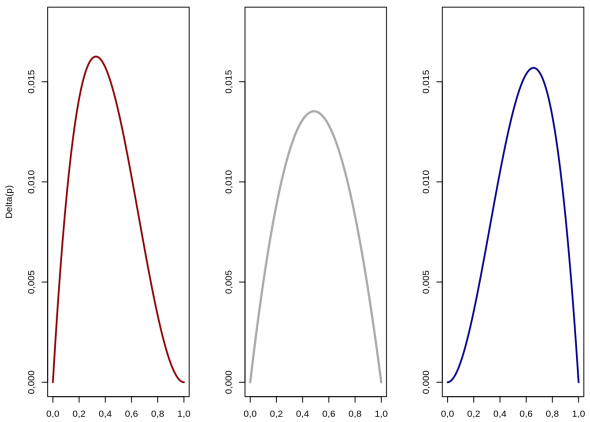


Figura 4.6. Funciones de $\Delta_s p$ para valores fijos del parámetro h. A la izquierda $h = 0{,}001$ (rojo), en el centro $h = \frac{1}{2} = 0{,}5$ (gris) y a la derecha $h = 0{,}999$ (azul). Se aprecia que el máximo de esta función de cambio en la frecuencia del alelo A_1 se desplaza de izquierda a derecha también.

Los ejemplos de selección direccional son abundantes, tanto a nivel de selección natural (que es el foco de este capítulo) como lo que trabajaremos más adelante en los capítulos referidos a la selección artificial. Uno de los ejemplos más usados es el de la **mariposa de los abedules**, *Biston betularia*. Esta mariposa existe en al menos dos coloraciones contrastantes, como se aprecia en la figura 4.7. Antes de la revolución industrial, la mariposa de color claro era la predominante ya que esta coloración le permite un mimetismo mayor cuando se posa sobre la corteza de los abedules (género *Betula*), que además le sirven de alimento. La revolución industrial marcó un hito en la historia de nuestro planeta en lo que hace a contaminación, ya que gran parte de las fuentes de energía usadas (como el carbón) produjeron enormes cantidades de hollín, particularmente en las grandes ciudades y ciudades industriales. La consecuencia de esto fue que gran parte de los árboles donde se posaban las mariposas fueron cambiando de coloración hacia una mucho más oscura, lo que sin duda, en lugar de permitir el mimetismo de la forma clara, las transformaba en objetivo claro de los predadores. En gran parte por esta razón, en poco tiempo el tipo oscuro desplazó al tipo claro en los alrededores de las ciudades industriales, aunque la forma clara siguió predominando en zonas no contaminadas.

Figura 4.7. *Biston betularia* clara (arriba) y oscura (abajo). El tipo claro predominaba antes de la revolución industrial, aún cerca de las grandes ciudades, mientras que el tipo oscuro pasó a predominar durante los primeros tiempos de la revolución industrial en las grandes ciudades. Imágenes de Wikipedia (autor Chiswick Chap, CC BY-SA 3.0).

Otro ejemplo famoso es el de los pinzones de Darwin. Cuando Charles Darwin visitó las islas Galápagos durante el segundo viaje del *Beagle*, recolectó varios pájaros, específicos de cada isla, que creyó en principio eran diferentes especies. Al retornar a Inglaterra, el ornitólogo John Gould logró determinar que se trataba de 18 especies de pinzones estrechamente emparentadas, pese a las importantes diferencias en las morfologías de los picos. En la figura 4.8 se aprecian las diferencias entre 4 de ellas. Más allá de la especulación sobre la selección direccional por el tipo de alimento que persistió durante décadas, los biólogos evolutivos

Peter y Rosemary Grant, que dedicaron alrededor de treinta años al estudio de los pinzones de Darwin, pudieron demostrar la existencia de diversas formas de alimentación de acuerdo con la morfología y tamaño de los picos. Más aún, recientemente se identificaron algunos genes relacionados a la forma y tamaño de los picos, en particular los genes **bone morphogenetic protein 4** (BMP4) (Abzhanov et al. 2004) y la **calmodulina** (CaM) (Abzhanov et al. 2006), lo que aporta una base mecanística a un fenómeno observado a otra escala (la de poblaciones).

Figura 4.8. Pinzones de Darwin: 4 de las 18 especies. En sentido horario, de izquierda arriba, *Geospiza magnirostris*, *Geospiza fortis*, *Certhidea fusca* y *Camarhynchus parvulus*. Imagen de Wikipedia, collage de Kiwi Rex (CC BY-SA 4.0).

En ambos ejemplos podemos observar algo muy importante: la selección natural tiene una relación directa con la evolución de los ambientes. Es decir, en general gran parte de las ventajas o desventajas que proporcionan determinados alelos en determinados locus son apenas temporales o dependientes del ambiente, por ejemplo, de la geografía. En el ejemplo de *Biston betularia* vimos que un cambio temporal en el ambiente en un período de tiempo evolutivo muy corto, del orden de decenas de años (considera que la evolución de la vida en la tierra tiene más de 3 mil millones de años) provocó una ventaja selectiva para los alelos del fenotipo oscuro, solamente en los entornos con fuerte contaminación atmosférica. Es decir, la ventaja es geográfica y también temporal (en el entendido de que es esperable y deseable reducir la contaminación en las grandes ciudades). Esto nos lleva directamente a otro punto relevante que ya hemos discutido: para que la selección pueda actuar es necesario contar con variabilidad genética y si la selección direccional es muy fuerte y rápida como para que aparezcan nuevas variantes relevantes, el éxito adaptativo en un ambiente y tiempo puede transformarse en una catástrofe si las condiciones cambian rápidamente.

En el segundo ejemplo vimos un fenómeno muy importante, del que ya hablamos algo en el capítulo previo: la **especiación**. Las islas con cierta distancia geográfica entre ellas (la escala depende de la capacidad de moverse de cada especie) suelen constituir un típico ejemplo y modelo para la **aparición de nuevas especies**, es decir la especiación. En el capítulo anterior vimos que los eventos estocásticos producto de tamaños poblacionales finitos (y usualmente pequeños), en ausencia de migración (y de otras fuerzas evolutivas) podían

llevar a la fijación de alelos distintos en las poblaciones del **ensemble**. Más aún, vimos que el **efecto fundador** podía ser responsable de fuertes variaciones en los puntos de partida de las distintas poblaciones «La subdivisión poblacional y la evolución de las frecuencias alélicas». Por ejemplo, si la distancia entre islas de un archipiélago es suficiente como para hacer el evento de arribo de una determinada nueva especie algo infrecuente, entonces una vez arribada la primera pareja reproductiva, si las condiciones son favorables, tendrá un espacio de varias generaciones para que se desarrolle la población. Esto provocará que cuando ocurran nuevos arribos la población ya contará con muchos individuos (descendientes de la primera pareja) y hasta posiblemente un nicho saturado. Durante todo este tiempo la población fue además posiblemente evolucionando en forma adaptativa a las condiciones específicas de esta isla, lo que podría llegar a crear barreras reproductivas con los nuevos arribos (tanto físicas cómo genéticas o etológicas). El reforzamiento del aislamiento por las razones anteriores es un mecanismo clásico postulado para la especiación, que se conoce como **especiación alopátrica** (algo más veremos en otro capítulo, en la sección «Mecanismos de especiación»).

Ejemplo 4.2

Determinar el efecto de heterocigotos (h) y el coeficiente de selección (s) para el caso planteado en el ejemplo **4.1**. ¿De qué tipo de acción génica (alélica) se trata? Si la situación continúa invariante, ¿qué se espera respecto a las frecuencias de los alelos A_1 y A_2?

Notemos que el homocigota de referencia es el A_2A_2 ya que es el de mayor valor y, por lo tanto, el **fitness** relativo de los organismos homocigotas A_1A_1 está dado por $w_{A_1A_1} = 1-s = \frac{0,5}{1.5} \Rightarrow s = 1-\frac{0,5}{1,5} = \frac{2}{3}$. A su vez, el **fitness** relativo de los organismos heterocigotas A_1A_2 está dado por $w_{A_1A_2} = 1-hs \Rightarrow 1-hs = \frac{0,8}{1,5}$, por lo que $hs = h \times \frac{2}{3} = 1-\frac{0,8}{1,5} \Rightarrow h = (1-\frac{0,8}{1,5}) \times \frac{3}{2} = 0,7$

Dado que $0 < h < 1$ nos encontramos frente a un caso de dominancia incompleta. De hecho, ya contábamos con esta información, en tanto el **fitness** relativo de los organismos heterocigotas se encontraba en un punto intermedio entre los de los organismos homocigotas. De mantenerse inalteradas las presiones selectivas (por ejemplo, la aplicación del plaguicida), es de esperar un aumento progresivo de la frecuencia del alelo A_2. En un tiempo finito el efecto exclusivo de la selección direccional no será capaz de fijar el alelo. No obstante, la acción de la deriva genética puede llevar a esto, eliminándose de la población el alelo A_2.

PARA RECORDAR

- Bajo **dominancia incompleta**, el **efecto de los heterocigotos** cumple $0 < h < 1$ y para **coeficientes de selección** no nulos, $s > 0$ (es decir, $w_{11} > w_{22}$), entonces $\Delta_s p > 0$ para todo el intervalo $0 < p < 1$, por lo que $p \to 1$.
- La frecuencia de A_1 crecerá hacia la fijación, pero no la alcanzará en tiempo finito si solo depende de la selección. Sin embargo, en la vecindad de $p = 1$, el efecto de la deriva puede terminar fijando el alelo A_1.
- Bajo **dominancia incompleta** no existe un punto de equilibrio intermedio entre $p = 0$ y $p = 1$, y la frecuencia del alelo A_1 tiende a crecer indefinidamente (o sea, en la misma dirección), por lo que a este tipo de selección se le conoce como **selección direccional**.

4.3.2 Selección estabilizadora

El segundo tipo de selección que vamos a analizar ocurre cuando la **sobredominancia** es el modo de acción génica. Si recuerdas de la figura 4.2, sobredominancia era cuando $h < 0$, lo que llevaba a que el **fitness relativo** del heterocigoto fuese mayor que el de ambos homocigotos y por tanto mayor que 1 (el fitness del homocigoto con mayor fitness). En términos matemáticos, como el **fitness relativo** del heterocigoto era $1 - hs$ y $s > 0$, $h < 0$, entonces $-hs > 0$ y $1 - hs > 1$.

Razonemos primero basado en nuestra intuición, antes de pasar al desarrollo matemático. Si el ***fitness*** **relativo** de los heterocigotos es mayor que el de ambos homocigotos, entonces esperamos que haya mayor cantidad de estos, ¿cierto? Pero los heterocigotos por definición tienen un alelo de cada tipo (por ejemplo, un A_1 y un A_2), entonces en principio no esperamos que desaparezca de la población ninguno de los dos. Ahora, ¿cuál será el punto de equilibrio, si es que existe? Por un lado, los heterocigotos aportan cada uno un alelo distinto, pero los homocigotos seguirán apareciendo generación tras generación ya que bajo apareamiento al azar los heterocigotos también producen homocigotos. ¿En qué proporción?

Para estudiar el comportamiento de la frecuencia del alelo A_1 bajo estas condiciones, veamos qué ocurre con la ecuación (4.6) en el equilibrio, es decir cuando ya no haya más cambios en la frecuencia (por ejemplo, $\Delta_s p = 0$). Si ponemos esta restricción en la ecuación (4.6), el numerador de la misma debe ser 0, por lo que (descartando el caso particular de $s = 0$ o cuando $p = 0$ o $q = 0$)

$$\Delta_s p = 0 = \hat{p}\hat{q}s[\hat{p}h + \hat{q}(1-h)] \Leftrightarrow \hat{p}h + \hat{q}(1-h) = 0$$
$$\hat{p}h + (1-\hat{p})(1-h) = 0 \ \therefore \hat{p}h + 1 - \hat{p} - h + \hat{p}h$$
$$= 0 \Leftrightarrow 2\hat{p}h - h + 1$$
$$\hat{p}(2h-1) - (h-1) = 0 \ \therefore$$

$$\hat{p} = \frac{h-1}{2h-1} \tag{4.8}$$

Es decir, la frecuencia $\hat{p}$ de equilibrio para el alelo A_1 (y para el A_2 su complemento a 1), estará dada por la ecuación (4.8). Por esta razón, llamamos a la selección que ocurre a través de la ventaja de los heterocigotas **selección estabilizadora** (porque la frecuencia de los alelos se estabiliza en un valor intermedio manteniendo la variabilidad). De todas formas, es importante mencionar que en realidad la **selección estabilizadora** también se da bajo otros mecanismos como la **selección dependiente de la frecuencia** (que no vamos a analizar en esta sección). Un hecho primero a notar sobre la **selección estabilizadora** es que la frecuencia de equilibrio **NO depende** del **coeficiente de selección** s y que solo depende del **efecto de los heterocigotos** h.

En la figura 4.9 podemos ver (a la izquierda) el comportamiento de la frecuencia de equilibrio $\bar{p}$ en función del valor de $-1 < h < 0$. Claramente, a medida que $h \to 0$ por la izquierda (es decir, desde los valores negativos) la frecuencia de equilibrio $\hat{p} \to 1$ (recordemos que cuando $h = 0$ los heterocigotos valen 1, igual que los homocigotos A_1A_1, es decir

A_1 dominante sobre A_2). En el resto del rango de valores de h considerados la frecuencia de equilibrio estará en algún lugar intermedio, pero siempre con frecuencia $\hat{p} < 1$.

El comportamiento anterior se puede ver claramente en la figura 4.9 (centro), donde graficamos $\Delta_s p$ en función de p. Ahora, a diferencia de cuando tratamos «Selección direccional», tenemos valores positivos y negativos, es decir que harán que la frecuencia de A_1 aumente o descienda respectivamente. Si nos fijamos en la figura central de 4.9, la línea azul vertical representa el valor de equilibrio para $h = -0{,}99$ que será $\hat{p} = (h-1)/(2h-1) \approx 0{,}679$. Si nos ubicamos a la izquierda de ese valor, los $\Delta_s p$ son todos positivos, por lo que la frecuencia p ser irá incrementando, pero a la izquierda del valor de equilibrio son todos negativos, por lo que p tenderá a bajar. La única solución a esto es que $\hat{p}$ es el punto de equilibrio, como habíamos predicho y que en particular se trata de un **punto de equilibrio estable**: si nos apartamos un poquito para cualquiera de los dos lados del punto de equilibrio, las fuerzas actuantes (representadas en $\Delta_s p$) nos volverán a traer al punto $\hat{p}$.

Finalmente, en la figura 4.9 (izquierda) podemos ver la evolución en el tiempo (medido en generaciones) de la frecuencia p bajo **selección estabilizadora**, partiendo desde diferentes valores de p inicial. Claramente, independiente del punto de partida todas las poblaciones convergen hacia la frecuencia $\hat{p} \approx 0{,}679$ (ya que $h = -0{,}99$). Cuanto más apartado el punto inicial de la frecuencia de equilibrio más esperaríamos que tarde en acercarse a este valor, pero si observas con cuidado la figura del centro verás la enorme diferencia en los valores absolutos de $\Delta_s p$ a la derecha e izquierda. Por lo tanto, si bien la diferencia de frecuencias entre el punto de equilibrio $\hat{p}$ y los puntos de partida son mucho mayores para los que arrancan por debajo de $\hat{p}$, la velocidad de avance hacia la convergencia será mucho mayor que para los que arrancan por encima de $\hat{p}$.

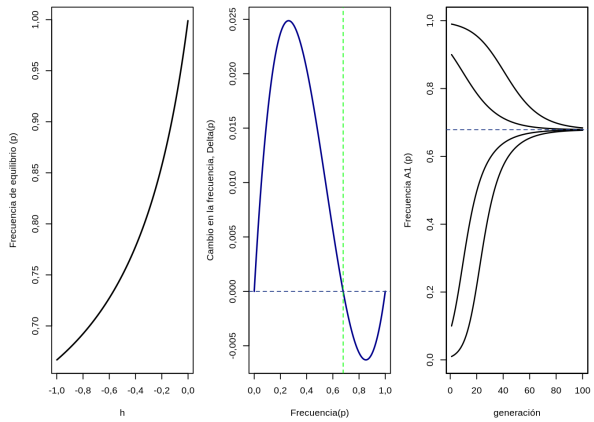


Figura 4.9. Frecuencia de equilibrio para el alelo A_1 ($\hat{p}$) bajo selección estabilizadora (izquierda), cambio entre generaciones de la frecuencia ($\Delta_s p$) dependiendo de la frecuencia p (centro) y evolución de la frecuencia p en sucesivas generaciones a partir de diferentes puntos de partida en la frecuencia inicial p (izquierda). En los gráficos del centro e izquierda $s = 0{,}1$ y $h = -0{,}99$ (por lo que $\hat{p} = (h-1)/(2h-1) \approx 0{,}679$).

Posiblemente, el ejemplo más conocido de **selección estabilizadora** sea el de la **anemia falciforme** o **anemia de células falciformes**. La **anemia de células falciformes** es una enfermedad de base genética que afecta los glóbulos rojos. En efecto, existe un alelo para el gen de la hemoglobina (Hgb) que se conoce como **S** (o como **HgbS**), que produce glóbulos

rojos deformes, con forma de hoz[6] y extremadamente sensibles a la reducción en la tensión de oxígeno. Este alelo es el producto de una sustitución nucleotídica en el gen codificante para globina β, que a su vez lleva a sustituir el aminoácido ácido glutámico que se encuentra en la posición 6 de la cadena polipeptídica por una valina. Las personas que heredan dos alelos ***HgbS***, de madre y padre, padecen la enfermedad que los lleva a una expectativa de vida reducida. En cambio, los individuos que reciben una sola copia de este, los heterocigotos, en principio no se ven afectados o sufren muy poco efecto detrimental (por ejemplo, son portadores de la enfermedad). Hasta acá estamos en una situación conocida, con el alelo normal *A* dominando al alelo *S*. Sin embargo, frente a la presencia de **malaria** (una enfermedad causada por el parásito *Plasmodium falciparum*), aparece un efecto inesperado: los heterocigotos aparecen como relativamente resistentes a la infección y, por lo tanto, en regiones donde la malaria es **endémica** tienen una ventaja considerable frente al homocigoto para el alelo normal. Esto configura un esquema de ventaja para el heterocigoto, o de **sobredominancia**. En la figura 4.10 se puede apreciar un mapa con las regiones de África donde la malaria es **endémica** (a la izquierda), así como un mapa de densidad con las frecuencias del alelo **HgbS**. En ella se aprecia claramente una superposición de las regiones, efecto claro del modo de acción de este alelo bajo la presión selectiva de la **malaria**. Obviamente, sin esta presión selectiva (los lugares donde la malaria ha sido erradicada) la situación vuelve a la relación donde el **fitness relativo** de los heterocigotos es casi idéntico al de los homocigotos para el alelo normal **A**. Esto equivale a decir que $h \approx 0$ y, por lo tanto,la selección se comportará como direccional contra el recesivo. Se tenderá a la eliminación del alelo ***S***, aunque como ya vimos antes en la sección «¿Dónde se 'esconden' los alelos recesivos?», esta selección es altamente ineficiente.

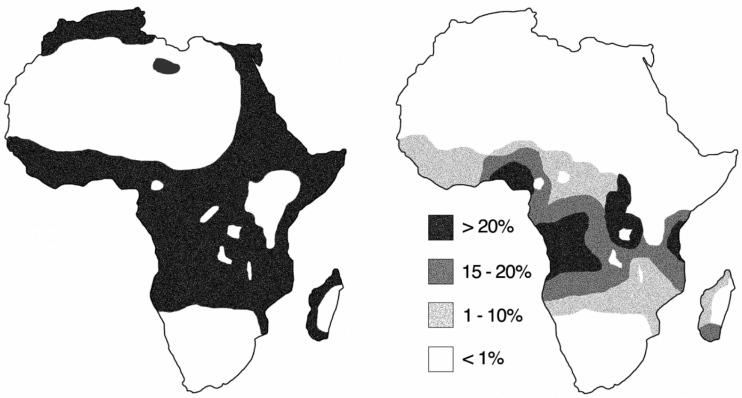


Figura 4.10. Distribución geográfica de la malaria (a la izquierda) y de la frecuencia del alelo *S* (a la derecha). Se observa la importante superposición de rangos geográficos entre la incidencia de malaria y la frecuencia del alelo que incrementa la resistencia a la misma. Figura de dominio público (original work by Anthony Allison, CC0, https://commons.wikimedia.org/w/index.php?curid=8714486).

[6] *Sickle*, en inglés. De ahí el uso de la letra *S* para denotar al alelo.

Ejemplo 4.3

Se ha postulado que la coloración de los pétalos de la flor de *Gymnadenia rhellicani* (una orquídea que crece en los Alpes) se encuentra bajo selección estabilizadora (Kellenberger et al. 2019).

La coloración de los pétalos se debe a la cantidad de pigmento (moléculas de la familia de las antocianinas). Esta se ve controlada por el locus *GrMYB,* codificante para un factor de transcripción que regula la síntesis de estos compuestos. Organismos homocigotas para el alelo GrMYB_1 portan coloración negra, mientras que los homocigotas para GrMYB_2 portan coloración blanca. Los heterocigotas $\mathrm{GrMYB}_1/\mathrm{GrMYB}_2$ presentan coloración roja.

Supongamos que los polinizadores presentan una preferencia según la coloración, siendo los **fitness** relativos $w_{11} = 1$; $w_{12} = 1{,}5$; $w_{22} = 0{,}7$ ¿Cuál es la frecuencia de equilibrio esperada para los alelos GrMYB_1 y GrMYB_2?

El genotipo homocigota GrMYB_2 es el que posee menor **fitness** relativo. Considerando $w_{22} = 1-s = 0{,}7 \Rightarrow s = 1-0{,}7 = 0{,}3$. Por otro lado, tenemos que existe sobredominancia, con el **fitness** de los heterocigotas rojos $w_{12} = 1 - hs = 1{,}5 \Rightarrow hs = 1 - 1{,}5 = -0{,}5$, lo cual implica $h \times 0{,}3 = -0{,}5 \Rightarrow h = \frac{-0{,}5}{0{,}3}$. La frecuencia de equilibrio para el alelo GrMYB_1 se puede calcular entonces considerando

$$f(\hat{\mathrm{GrMYB}_1}) = \frac{h-1}{2h-1} = \frac{(\frac{-0{,}5}{0{,}3}) - 1}{2 \times (\frac{-0{,}5}{0{,}3}) - 1} \approx 0{,}62$$

La frecuencia de equilibrio esperada para el alelo GrMYB_2 es el complementario ($\approx 0{,}38$).

4.3.2 Otra parametrización

En algunos libros se puede encontrar una parametrización diferente para el caso de **sobredominancia** y para la frecuencia de equilibrio. En esta parametrización diferente se asume que el heterocigoto tiene el **fitness** máximo e igual a 1. Por lo tanto, el homocigoto A_1A_1 tendrá un **fitness** de $1-t$ y el homocigoto A_2A_2 tendrá un **fitness** de $1-s$, como se aprecia en la figura 4.11.

Si recordamos de antes, el cambio en frecuencia de p, o sea $\Delta_s p$ lo podíamos escribir como

$$\Delta_s p = \frac{pq[p(w_{11} - w_{12}) + q(w_{12} - w_{22})]}{\bar{w}} \tag{4.9}$$

Sustituyendo en (4.9) los valores de **fitness** definidos antes para esta parametrización, tenemos

$$\Delta_s p = \frac{pq[p((1-t)-1) + q(1-(1-s))]}{\bar{w}} \tag{4.10}$$

Operando en (4.10), tenemos ahora que el cambio es

$$\Delta_s p = \frac{pq[p(-t) + q(s)]}{\bar{w}} = \frac{pq[-pt + qs]}{\bar{w}} \tag{4.11}$$

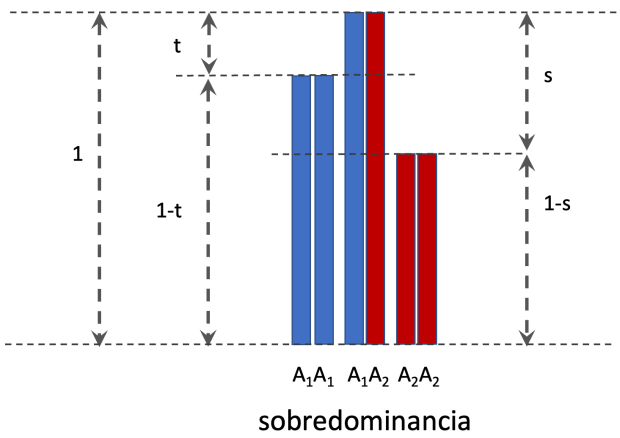


Figura 4.11. Representación gráfica de los **fitness relativos** para los distintos genotipos cuando el modo de acción génica es de **sobredominancia**. Ahora, en lugar de una parametrización en h y s, con el homocigoto A_1A_1 como valor igual a uno, en lugar de eso el heterocigoto será el máximo, con valor de uno y ambos homocigotos con valores de $1-t$ y $1-s$ respectivamente.

y si sustituimos ahora $q = (1-p)$ para quedarnos con un solo parámetro de frecuencia, tenemos ahora

$$\Delta_s p = \frac{pq[-pt + (1-p)s]}{\bar{w}} = \frac{pq[-pt + s - ps]}{\bar{w}} = \frac{pq[s - p(s+t)]}{\bar{w}} \tag{4.12}$$

En el equilibrio, el cambio de frecuencia es cero, es decir $\Delta_s p = 0$. Obviando $p = 0$ y $q = 0$ que quedan fuera de nuestro espacio permitido para **sobredominancia**, tenemos que la única solución a $\Delta_s p = 0$ es

$$\Delta_s \hat{p} = \frac{\hat{p}\hat{q}[s - \hat{p}(s+t)]}{\bar{w}} = 0 \Leftrightarrow s - \hat{p}(s+t) = 0 \Leftrightarrow s = \hat{p}(s+t) \tag{4.13}$$

por lo que

$$\hat{p} = \frac{s}{s+t} \tag{4.14}$$

Por lo tanto, la frecuencia de equilibrio con estos parámetros será $\hat{p} = \frac{s}{s+t}$. Esto es bien fácil de entender de forma intuitiva. Como los heterocigotas tienen un alelo A_1 y un alelo A_2 (y por tanto, no pueden distinguirlos) solo constituyen el fiel de la balanza entre ambos alelos. Donde se juega el partido entonces es en el balance entre los homocigotas. Los parámetros t y s representan el detrimento en los genotipos A_1A_1 y A_2A_2. Entonces, cuanto más importante es s en el total de detrimento $s+t$, mayor será la ventaja de A_1 sobre A_2. Por lo tanto, la proporción de detrimento de A_2 (por ejemplo, $\frac{s}{s+t}$) será a su vez la ventaja del A_1. El equilibrio esperado ocurre cuando la frecuencia de A_1 es igual a la proporción de su ventaja: $\hat{p} = \frac{s}{s+t}$

PARA RECORDAR

- Bajo el efecto de la **sobredominancia**, es decir con el **efecto de los heterocigotos** $h < 0$ y para **coeficientes de selección** no nulos, $s > 0$ (es decir, $w_{11} > w_{22}$), entonces existirá un punto de equilibrio **estable** en el intervalo de frecuencias $0 < p < 1$.
- Este punto de equilibrio solo depende del valor de h y no de s y se encontrará en $\hat{p} = \frac{h-1}{2h-1}$.
- El anterior se trata de un punto de equilibrio estable ya que la selección a ambos lados es hacia ese punto y cualquier perturbación aleatoria de la frecuencia a los lados no hará más que activar la selección nuevamente hacia el punto de equilibrio ($\Delta_s p$ lleva siempre $p \to \hat{p}$).
- Por lo anterior, a la selección bajo el modo de acción génica de **sobredominancia** se le conoce como **selección estabilizadora**.
- En el caso de usar la parametrización alternativa, con $1 - t$ y $1 - s$ como **fitness relativo** para los homocigotos A_1A_1 y A_2A_2 respectivamente, el punto de equilibrio en la frecuencia del alelo A_1, $\hat{p}$, estará dado por $\hat{p} = \frac{s}{s+t}$.

4.3.3 Selección disruptiva

El tercer tipo de selección que vamos a abordar es cuando el modo de acción génica es la **subdominancia** ($h > 1$). En ese caso, siendo que $s > 0$, entonces el heterocigoto se encuentra por debajo de ambos homocigotos, ya que $1 < 1 - s < 1 - hs$ (lo último porque $hs > s$ ya que $h > 1$). Esta situación es la representada más a la derecha en la figura 4.2. Para estudiar el comportamiento de la selección bajo este régimen podemos comenzar por razonar nuevamente como lo hicimos en el caso de la «Selección estabilizadora». Si el heterocigota es ahora el genotipo menos favorecido por la selección, entonces a la inversa del caso anterior los heterocigotos tenderán a desaparecer, aumentando la frecuencia de los homocigotos a sus expensas. Está claro que hay que escapar de la región que maximiza los heterocigotos (que era $p = q = \frac{1}{2}$), pero ahora no existe un sentido o punto hacia donde deba dirigirse $\Delta_s p$ ya que las alternativas son dos: dirigirse hacia donde p es menor (hacia $p = 0$) o dirigirse en el otro sentido (hacia $p = 1$). Esta disyuntiva aparece más clara si graficamos $\Delta_s p$ en función de p para distintos valores de $h > 1$. En la figura 4.12 se puede visualizar lo que ocurre con $\Delta_s p$ para $h = 1{,}1$ (izquierda), $h = 1{,}5$ (centro) y $h = 1{,}9$ (derecha). En principio, de forma inocente, podemos creer que estos gráficos son iguales a los de la figura 4.9 (centro y derecha). En ambas figuras $\Delta_s p$ tiene valores positivos y negativos a ambos lados de los puntos de equilibrio, los que parecen determinados por la ecuación $\hat{p} = \frac{h-1}{2h-1}$. Sin embargo, las gráficas de la figura 4.12 parecen ser en espejo de las de la figura 4.9. ¿Tendrá esto alguna relevancia? Veamos qué ocurre, por ejemplo, con la gráfica del centro de la figura 4.12. De acuerdo con el valor de h el punto de equilibrio estaría en una frecuencia de $\bar{p} = 0{,}25$, lo que aparece marcado por la línea azul vertical. Si partimos a la izquierda del punto de equilibrio, el signo de $\Delta_s p$ es negativo, por lo que a cada generación p irá disminuyendo y a medida que pasen las generaciones $p \to 0$. Si arrancamos a la derecha del punto de equilibrio, el signo de $\Delta_s p$ es positivo, por lo que a cada generación p irá aumentando y a medida que pasen las generaciones $p \to 1$. ¿Entonces por qué decimos que es un punto de equilibrio si en ninguno de los dos casos terminamos en ese valor de $\hat{p}$? Veamos que ocurre si arrancamos exactamente de ese punto $\hat{p}$ en la generación inicial.

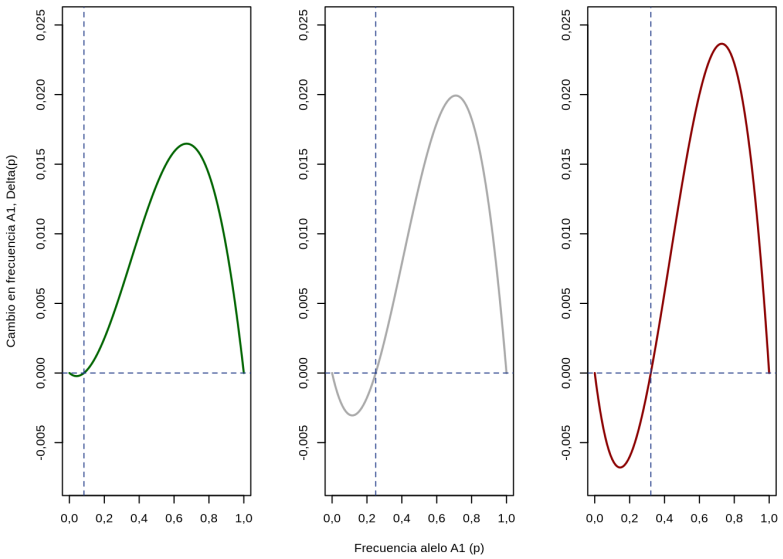


Figura 4.12. Evolución del cambio en frecuencia $\Delta_s p$ en función del punto de partida p y del parámetro h, en el modo de acción génica **subdominancia**. A la izquierda $h = 1{,}1$, en el centro $h = 1{,}5$ y a la derecha $h = 1{,}9$. En todos los casos $s = 0{,}1$.

Por definición, $\hat{p}$ se corresponde con el punto donde $\Delta_s p = 0$, por lo que no cambiará para la siguiente generación, que seguirá estando en $\hat{p}$ y así sucesivamente. Sin embargo, aún la menor perturbación del punto de equilibrio (por deriva, por ejemplo) hará que p se desplace a derecha o izquierda, donde los $\Delta_s p \neq 0$, lo que sellará definitivamente la suerte del alelo A_1: si el desplazamiento inicial es a la izquierda del punto de equilibrio el destino será la desaparición de este alelo, mientras que, si el desplazamiento es hacia la derecha, el destino será la fijación. La evolución a lo largo del tiempo dependiendo del punto de partida se puede ver en la figura 4.13 para 4 puntos de partida diferentes, por arriba y por abajo del punto de equilibrio.

Por esta razón, se dice que el punto $\hat{p}$ en este tipo de acción génica corresponde a un **equilibrio inestable**. Es como tener equilibrado un lápiz en la punta de otro lápiz: el menor movimiento llevará al lápiz en equilibrio a caer por uno de los dos lados.

Desde el punto de vista biológico, la existencia de **subdominancia** se encuentra asociada a la **selección disruptiva**. Imaginemos dos poblaciones de la misma especie que solapan sus rangos geográficos (pero que por distintas razones se reproducen cada una dentro de su grupo) y que arrancan desde la misma frecuencia inicial en el punto de equilibrio inestable. Si por el efecto de la deriva genética (por ejemplo, porque ambas poblaciones son pequeñas en tamaño) una se desplaza hacia $p > \hat{p}$, mientras que la otra en el sentido contrario ($p < \hat{p}$), al cabo de unas cuantas generaciones tendremos dos poblaciones diferenciadas en nuestro *locus* de referencia. Nuevamente, un ejemplo de este tipo de selección parece ser el de los pinzones de Darwin, esta vez los de la isla de Santa Cruz. La diferencia en el tamaño de semillas disponibles para la alimentación de las aves es grande, lo que ha llevado a especializarse en un tipo, desapareciendo las formas intermedias (que tendrían un mayor rango

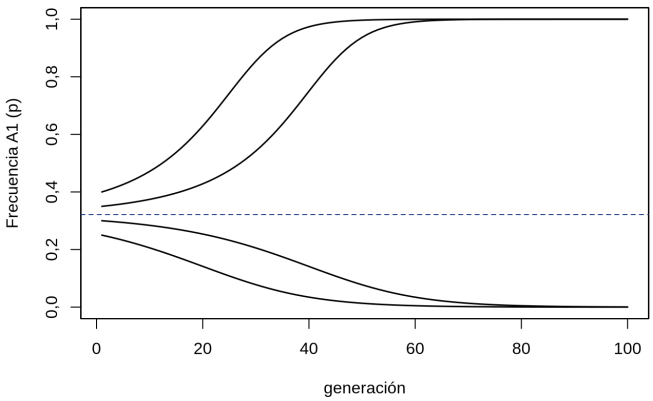


Figura 4.13. Evolución de la frecuencia p en sucesivas generaciones a partir de diferentes puntos de partida en la frecuencia inicial p bajo selección disruptiva. $h = 1{,}9$, $p = 0{,}25; 0{,}30; 0{,}35; 0{,}40$.

de alimentación, pero son menos eficientes en ambos tipos). Otro lugar donde la **selección disruptiva** parece haber jugado un lugar importante es en lo eventos

PARA RECORDAR

- Bajo el efecto de la **subdominancia**, es decir con el **efecto de los heterocigotos** $h > 1$ y para **coeficientes de selección** no nulos, $s > 0$ (es decir, $w_{11} > w_{22}$), entonces existirá un punto de equilibrio **inestable** en el intervalo de frecuencias $0 < p < 1$.
- Este punto de equilibrio solo depende del valor de h y no de s y se encontrará en $\hat{p} = \frac{h-1}{2h-1}$.
- El anterior se trata de un punto de equilibrio **inestable** ya que cualquier perturbación aleatoria de la frecuencia a los lados del equilibrio no hará más que disparar la selección hacia alguno de los extremos, $p \to 1$ si $p > \hat{p}$ (es decir a la fijación) o $p \to 0$ si $p < \hat{p}$ (es decir a la pérdida del alelo).
- Por lo anterior, a la selección bajo el modo de acción génica de **subdominancia** se le conoce como **selección disruptiva.**

4.4 El teorema fundamental de la selección natural

Si sustituimos q por $(1-p)$ en la ecuación (4.7), luego de un poquito de trabajo llegamos a que, expresado solo en función de p, el **fitness medio** (en nuestra escala relativa) es

$$\bar{w} = (1-s) - 2ps(h-1) + p^2 s(2h-1) \tag{4.15}$$

Ahora, si queremos estudiar el comportamiento del **fitness medio** en función de la frecuencia del alelo A_1 (p), debemos hacer la derivada de esta función, que es simplemente (recordemos que aquellos términos que no sean función de p son tratados como constantes)

$$\frac{d\bar{w}}{dp} = -2s(h-1) + 2sp(2h-1) \tag{4.16}$$

En los máximos o mínimos de $\bar{w}$ la derivada respecto a la frecuencia p se hace cero, por lo que

$$\frac{d\bar{w}}{dp} = -2s(h-1) + 2sp(2h-1) = 0 \Leftrightarrow (h-1) = p(2h-1) \Leftrightarrow p = \frac{h-1}{2h-1} \tag{4.17}$$

Notemos que, cuando $2h-1=0 \Leftrightarrow h=\frac{1}{2}$ la derivada primera no existirá. Para saber si se trata de máximos o mínimos debemos estudiar el signo de la derivada segunda (recordar que si el signo de la derivada segunda es positivo en la región del cero de la derivada primera, entonces se trata de un mínimo, mientras que, si el signo es negativo,se trata de un máximo). La derivada segunda de $\bar{w}$ respecto a p es

$$\frac{d^2\bar{w}}{dp^2} = 2s(2h-1) \tag{4.18}$$

y ya que $2s$ es positivo, el signo de la derivada segunda dependerá de $(2h-1)$. Es decir, si $h > \frac{1}{2}$ entonces la derivada segunda será positiva y se tratará de un mínimo; si $h < \frac{1}{2}$ la derivada segunda será negativa y, por lo tanto, se tratará de un máximo. Si recordamos nuestra categorización anterior de los casos de acuerdo con los valores de h, en el rango $0 < h < 1$ no existía un punto de equilibrio. Esto se ve reforzado si estudiamos el rango de valores que tomará $\hat{p}$ en $\hat{p} = \frac{h-1}{2h-1}$, para el rango de $0 < h < 1$ (caso de dominancia incompleta que resulta en selección direccional). Una forma sencilla es graficar esta función de h en ese intervalo. Como se muestra en la figura 4.14 izquierda la función $\hat{p}$ no toma valores admisibles ($0 \leqslant p \leqslant 1$) en el rango $0 < h < 1$. En cambio, fuera de este rango (figura 4.14, derecha) la función $\hat{p}$ si toma valores admisibles, tanto en los casos de **sobredominancia** ($h < 0$), **subdominancia** ($h > 1$) o **dominancia completa** de uno de los dos alelos ($h = 0$ A_1 dominante o $h = 1$ A_2 dominante). Más aún, volviendo a nuestra derivada segunda, para valores de $h < \frac{1}{2}$ era negativa y, por lo tanto, en la región de $h < 0 < \frac{1}{2}$ (la región de **sobredominancia**) también será negativa y los puntos de equilibrio $\hat{p}$ serán

máximos del fitness, que coincide con lo que ya vimos en «Selección estabilizadora». De la misma manera, para valores de $h > \frac{1}{2}$ la derivada segunda era positiva y, por lo tanto, en la región de $h > 1 > \frac{1}{2}$ (la región de **subdominancia**) también será positiva y los puntos de equilibrio $\hat{p}$ serán **mínimos del fitness**, lo que coincide ahora con lo que ya vimos en «Selección disruptiva».

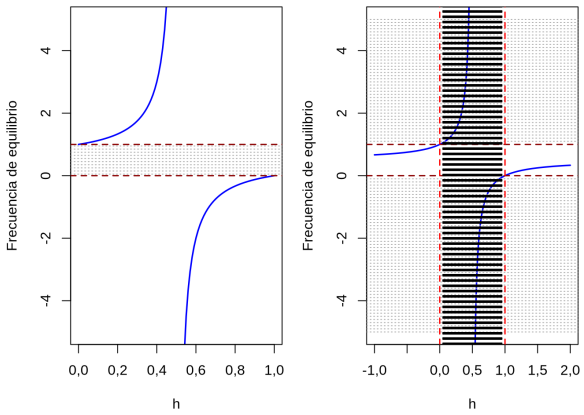


Figura 4.14. Inexistencia de puntos de equilibrio en el intervalo $0 < h < 1$ (gráfica a la izquierda). Para valores de h en el rango de **dominancia incompleta**, la función que representa el punto de equilibrio en la frecuencia del alelo A_1 (por ejemplo, $\hat{p} = \frac{h-1}{2h-1}$) toma valores fuera del rango posible de p, es decir $0 \leqslant p \leqslant 1$ (el rango de valores posibles se representa por las líneas grises entre las dos líneas rojas). Fuera del rango $0 < h < 1$ (gráfica a la derecha) las soluciones de $\hat{p} = \frac{h-1}{2h-1}$ entran dentro de los rangos admisibles para p, **sobredominancia** ($h < 0$), **subdominancia** ($h > 1$) o **dominancia completa** de uno de los dos alelos ($h = 0$ correspondiente a alelo A_1 dominante, o $h = 1$ correspondiente a alelo A_2 dominante).

Siguiendo el planteo de John H. Gillespie (2004), en cierta forma esta división de los modos de selección reclama por una ecuación conceptualmente unificadora. Sewall Wright, en su artículo de 1937 (Wright 1937) muestra que es posible escribir el cambio en frecuencia $\Delta_s p$ en una forma en principio muy diferente

$$\Delta_s p = \frac{pq}{2\bar{w}} \frac{d\bar{w}}{dp} \tag{4.19}$$

Antes de analizar esta formulación de $\Delta_s p$ veamos si podemos arribar a ella. Tomando en cuenta la forma que tiene $\frac{d\bar{w}}{dp}$ en la ecuación (4.16) y sustituyendo en (4.19), tenemos

$$\begin{aligned}
\Delta_s p &= \frac{pq}{2\bar{w}}[-2s(h-1) + 2sp(2h-1)] = \frac{pq}{2} \frac{2s(1-h) + 2sp(2h-1)}{\bar{w}} \\
&= \frac{pqs[(1-h) + p(2h-1)]}{\bar{w}} = \frac{pqs[1-h+2ph-p]}{\bar{w}} = \frac{pqs[1-h+ph-p+ph]}{\bar{w}} \\
&= \frac{pqs[(1-p)(1-h)+ph]}{\bar{w}} = \frac{pqs[ph+q(1-h)]}{\bar{w}} = \Delta_s p
\end{aligned} \tag{4.20}$$

por lo que claramente llegamos a la misma formulación que en nuestra ecuación (4.6).

Lo interesante de la formulación de Wright es el cambio en la frecuencia, $\Delta_s p$ que es ahora mucho más conceptual: se trata del producto de un término que nos da idea de la varianza en frecuencias (el término $\frac{pq}{2\hat{w}}$) y de un término que habla del cambio en el **fitness medio** en función de la frecuencia de p. El primer término es siempre positivo y, por lo tanto, el sentido del cambio estará dado por el signo de $\frac{d\hat{w}}{dp}$. El aspecto interesante de esta relación es que si la derivada tiene signo positivo, también $\Delta_s p$ y, por lo tanto, aumentará la frecuencia de p. Como $\frac{d\hat{w}}{dp} > 0$, entonces el fitness $\hat{w}$ también lo aumentará. Por otro lado, si la derivada tiene signo negativo, entonces también $\Delta_s p$ y, por lo tanto, disminuirá la frecuencia de p. Como $\frac{d\hat{w}}{dp} < 0$, entonces (al ir p hacia menores valores) el fitness $\hat{w}$ también aumentará. Es decir, en nuestro modelo de un locus con dos alelos, con el ***fitness*** **independiente de la frecuencia**, este siempre aumentará en la medida que haya varianza genética.

Relacionado con este concepto de evolución de la población hacia un máximo en el **fitness** se encuentra otra aproximación que motiva el nombre de esta sección. En su libro de 1930 *The Genetical Theory of Natural Selection* (Fisher 1930), Ronald Fisher dedica un capítulo entero a lo que (algo pomposamente) llama **Teorema Fundamental de la Selección Natural**. Su formulación original (traducción nuestra) luce:

> La tasa de aumento de la aptitud de cualquier organismo en cualquier momento es igual a su varianza genética de aptitud en ese momento.

En lo anterior, usamos **aptitud** en lugar de **fitness**. Este concepto fue reformulado con el paso del tiempo y una versión más moderna del mismo, con términos más usuales hoy en día, debida a Edwards (1994) (de nuevo, traducción nuestra) sería:

> La tasa de aumento de la aptitud media de cualquier organismo, en cualquier momento, que es atribuible a la selección natural que actúa a través de los cambios en las frecuencias de los genes, es exactamente igual a su varianza genética en la aptitud en ese momento.

El «teorema» ha sido objeto de una profunda controversia durante todo el siglo XX, cuyo origen probablemente esté en la disputa general entre Wright y Fisher, que continuó con abundantes malinterpretaciones del mismo y con dudas importantes acerca de su significancia. De hecho, alcanza con ver la conclusión aportada por John H. Gillespie (2004) («Thus, it is neither fundamental nor a theorem».[7]) para comprender que su alcance es algo limitado y su importancia difícil de decidir. En principio es totalmente cierto (y matemáticamente correcto) para poblaciones «teóricas», en un modelo de un locus con selección simple.

El tratamiento completo de este teorema requiere de los conceptos de la descomposición de la varianza, en particular de la varianza genética y los efectos aditivos, algo que veremos en el capítulo «El modelo genético básico». Sin embargo, una aproximación simplificada y bastante intuitiva se encuentra en el libro de Hamilton (Hamilton 2009), que nosotros seguiremos acá.

Supongamos una población de un organismo haploide con reproducción asexual o clonal[8]. En este caso, el **fitness medio** de la población será la suma del fitness de cada haplotipo

[7] Que podría traducirse: «Por lo tanto, ni es un teorema ni es fundamental».

[8] En realidad, esto no es un requerimiento en la formulación del teorema, pero sí para nuestro tratamiento simplificado.

(recordar que asumimos que el organismo es haploide, entonces el genotipo es igual al haplotipo) multiplicado por la frecuencia del mismo. Es decir, para los k haplotipos en la población,

$$\bar{w} = \sum_{i=1}^{i=k} (p_i w_i) \tag{4.21}$$

Para cada haplotipo la nueva frecuencia en la siguiente generación, luego de la selección natural, estará dada por

$$p_i' = p_i \frac{w_i}{\bar{w}} = \frac{p_i w_i}{\bar{w}} \tag{4.22}$$

y, por lo tanto, el nuevo **fitness medio** será igual a

$$\bar{w}' = \sum_{i=1}^{k} (p_i' w_i) \tag{4.23}$$

Ahora, si sustituimos el valor de p_i' que obtuvimos en la ecuación (4.22) en la ecuación (4.23), tenemos

$$\bar{w}' = \frac{1}{\bar{w}} \sum_{i=1}^{k} (p_i' w_i) w_i = \frac{1}{\bar{w}} \sum_{i=1}^{k} p_i' w_i^2 \tag{4.24}$$

El cambio relativo (al fitness en la generación previa) estará dado por

$$\Delta\bar{w} = \frac{\bar{w}' - \bar{w}}{\bar{w}} \tag{4.25}$$

y sustituyendo el resultado de la ecuación (4.24) en la ecuación (4.25), tenemos ahora que el cambio relativo lo podemos expresar como

$$\Delta\bar{w} = \frac{\frac{1}{\bar{w}} \sum_{i=1}^{k} p_i' w_i^2 - \bar{w}}{\bar{w}} \tag{4.26}$$

Si ahora multiplicamos y dividimos el término $\bar{w}$ por sí mismo, es decir $\bar{w} = \frac{\bar{w}^2}{\bar{w}}$ tenemos que el cambio se puede escribir como

$$\Delta\bar{w} = \frac{\frac{1}{\bar{w}} \sum_{i=1}^{k} p_i' w_i^2 - \frac{\bar{w}^2}{\bar{w}}}{\bar{w}} \tag{4.27}$$

y luego de reordenar esta ecuación (recordemos que dividir entre $\bar{w}$ es lo mismo que multiplicar por su inverso, es decir $\frac{1}{\bar{w}^2}$), llegamos a que

$$\Delta\bar{w} = \frac{1}{\bar{w}^2}\sum_{i=1}^{k} p_i' w_i^2 - \bar{w}^2 \tag{4.28}$$

Pero si miramos con atención esta última ecuación, el término $\sum_{i=1}^{k} p_i' w_i^2 - \bar{w}^2$ es la **varianza en el fitness** (recordemos que $\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}^2(x)$), por lo que $\Delta_s p \propto \mathbb{V}(w)$. Como el escalado del **fitness relativo** es arbitrario, podemos elegirlo de tal forma que $\bar{w} = 1$ y entonces nos queda la definición del **teorema fundamental de la selección natural**

$$\Delta_s p = \mathbb{V}(w) \tag{4.29}$$

PARA RECORDAR

- Otra forma de escribir el **fitness medio**, ahora solo dependiendo de p (y no de q) es

$$\bar{w} = (1-s) - 2ps(h-1) + p^2 s(2h-1) \tag{4.30}$$

- La tasa de cambio del **fitness medio** en función de la frecuencia se encuentra dado por la derivada primera del mismo, es decir
- Esta derivada tendrá puntos extremos en $\hat{p} = \frac{h-1}{2h-1}$, tanto para $h < 0$ o $h > 1$. De acuerdo con la derivada segunda, al ser negativa para $h < \frac{1}{2}$, para $h < 0 < \frac{1}{2}$ (**sobredominancia**), $\hat{p} = \frac{h-1}{2h-1}$ corresponderá a un máximo, mientras que para $h > 1 > \frac{1}{2}$ (**sobredominancia**), $\hat{p} = \frac{h-1}{2h-1}$ corresponderá a un mínimo del **fitness medio.**
- Otra forma de escribir el cambio esperado en las frecuencias del alelo A_1 que brinda mayor información evolutiva es

$$\Delta_s p = \frac{pq}{2\hat{w}}\frac{d\bar{w}}{dp} \tag{4.31}$$

  es decir, $\Delta_s p$ es el producto de un término referido a la varianza genética por un término que es la tasa de cambio del ***fitness*** **medio** relativo.
- El **teorema fundamental de la selección natural**, debido a Fisher (1930), establece que «la tasa de aumento de la aptitud media de cualquier organismo, en cualquier momento, que es atribuible a la selección natural que actúa a través de los cambios en las frecuencias de los genes, es exactamente igual a su varianza genética en la aptitud en ese momento» (Edwards 1994).

4.5 Equilibrio selección-mutación

Supongamos que ahora nuestro modelo no solo incluye el efecto de la selección favoreciendo al alelo A_1, sino que además tenemos el efecto recurrente de la reintroducción del alelo A_2. Este efecto, si es suficientemente significativo, impedirá la fijación del alelo A_1 y, por lo tanto, ahora existirá un equilibrio entre los dos alelos, cosa que no ocurría en el caso de **dominancia incompleta** si solo consideramos la selección, como vimos en «Selección direccional». Supongamos que el alelo A_1 muta hacia el alelo A_2 con una tasa u, es decir $A_1 \xrightarrow{u} A_2$. Suponemos además, por ahora, **dominancia incompleta** $(0 < h < 1)$.

Asumiendo que s es suficientemente fuerte para mantener la frecuencia de A_2 muy baja y que por lo tanto la incidencia de la mutación reversa ($A_2 \xrightarrow{v} A_1$) es despreciable (porque es el producto de una tasa muy pequeña v por una frecuencia también muy pequeña q) entonces la frecuencia de A_1 en la siguiente generación será

$$p' = p(1-u) \tag{4.32}$$

y el cambio debido a la mutación será (observemos la notación $\Delta_u p$ ahora, con el subscrito $_u$ para indicar mutación)

$$\Delta_u p = p' - p = p(1-u) - p = -up = -u(1-q) = -u + uq \tag{4.33}$$

La tasa de mutación u es usualmente extremadamente pequeña, del orden de 1×10^{-10} por base por replicación para muchos genomas de eucariotas multicelulares (aunque no del mismo orden en plantas). En grandes números esta tasa podría ser del orden de 1×10^{-5} expresada en términos de mutaciones por locus por replicación (con una varianza muy importante entre organismos). Por lo tanto, como $q \approx 0$, entonces

$$\Delta_u p = -u + uq \approx -u \tag{4.34}$$

Veamos ahora cómo es el balance entre el cambio de frecuencias que produce la mutación (ecuación (4.34)), decreciendo la frecuencia de A_1 y la selección que favorece a A_1. Si recuerdas de la ecuación (4.6), el cambio esperado en la frecuencia de A_1 por selección era

$$\Delta_s p = \frac{pqs[ph + q(1-h)]}{\bar{w}} \tag{4.35}$$

Pero si $q \approx 0$, entonces $p \approx 1$ y $\bar{w} \approx 1$ (ya que tanto $2pqhs \approx 0$ como $q^2 s \approx 0$). Además $pqs \times ph \approx qs \times h \approx qhs$ y $pqs \times q(1-h) = pq^2 s(1-h) \approx 0$ (porque $q^2 \ll q \approx 0$). Entonces, con esta condición de que la frecuencia del alelo A_2 se mantenga muy baja ($q \approx 0$), la ecuación anterior se transforma en

$$\Delta_s p \approx qhs \tag{4.36}$$

El juego entre la fuerza de cambio hacia A_2 ($\Delta_u p$) debida a la mutación y hacia A_1 ($\Delta_s p$) debida a la selección se puede ver en la figura 4.15. Para que las dos fuerzas se equilibren el

balance debe ser cero ($\Delta_u p + \Delta_s p = 0$), por lo que, usando $\hat{q}$ para la frecuencia del alelo A_2 en el equilibrio, tenemos que

$$-u + \hat{q}hs \approx 0 \therefore \hat{q}hs \approx u \tag{4.37}$$

y finalmente

$$\hat{q} \approx \frac{u}{hs} \tag{4.38}$$

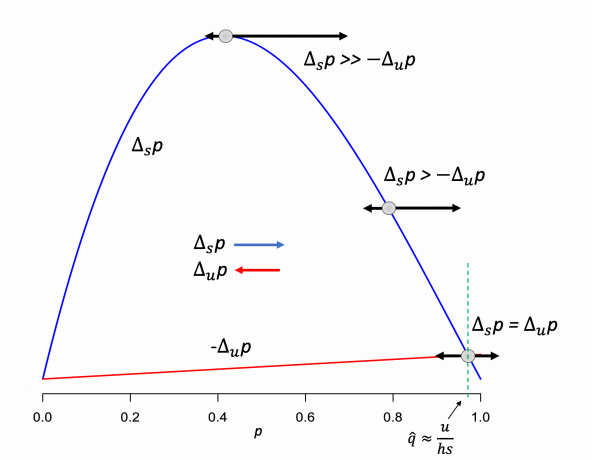


Figura 4.15. Ilustración del equilibrio selección-mutación. La selección tenderá a aumentar la frecuencia del alelo A_1 ($\Delta_s p$, línea azul), mientras que la mutación tenderá a reducirla ($\Delta_u p$, línea roja). Cuando ambas fuerzas son de igual magnitud (abajo a la derecha) se establecerá un equilibrio en una frecuencia $\hat{q} = \frac{u}{hs}$. Los valores usados para generar la figura son $h = \frac{1}{3}$, $s = 0{,}1$ y $u = 1 \times 10^{-3}$. En la figura aparece graficado $-\Delta_u p$ para poder ubicar el punto de corte de ambas fuerzas.

En palabras, la frecuencia de equilibrio para el alelo A_2 es el cociente entre la tasa de mutación con la que es introducido (u) y el producto del **coeficiente de selección** s (en su contra) y **efecto de los heterocigotos** h. Recordando que se trata de una aproximación que depende de que $q \approx 0$ y que además, como vimos en la sección «¿Dónde se 'esconden' los alelos recesivos?», los alelos raros se encuentran en su mayoría en los individuos heterocigotos (porque $2pq \gg q^2$), entonces resulta más claro por qué el denominador de la ecuación (4.38) incluye el efecto de la selección contra los individuos heterocigotos.

Un caso diferente es cuando $h = 0$. Esto sucede cuando A_1 domina completamente a A_2 (es decir, el **alelo perjudicial** es **completamente recesivo**). En ese caso, el cambio de frecuencia debido a la selección es

$$\Delta_s p = \frac{pqs[p \times 0 + q(1-0)]}{1 - 2pqs \times 0 - q^2 s} = \frac{pq^2 s}{1 - q^2 s} \tag{4.39}$$

En el equilibrio, tenemos que el cambio introducido por la mutación más el producido por la selección deben ser cero. Es decir, usando el resultado de la ecuación (4.33), tenemos

$$\Delta_s p + \Delta_u p = 0 = \frac{\hat{p}\hat{q}^2 s}{1-\hat{q}^2 s} - u + u\hat{q} = \frac{\hat{p}\hat{q}^2 s}{1-\hat{q}^2 s} - u(1-\hat{q}) \Leftrightarrow$$

$$u(1-\hat{q}) = \frac{\hat{p}\hat{q}^2 s}{1-\hat{q}^2 s}$$

$$u = \frac{\hat{p}\hat{q}^2 s}{(1-\hat{q}^2 s)(1-\hat{q})}$$

$$u = \frac{\hat{q}^2(1-\hat{q})s}{(1-\hat{q}^2 s)(1-\hat{q})}$$

$$u = \frac{\hat{q}^2 s}{(1-\hat{q}^2 s)} \Leftrightarrow$$

$$\frac{u}{s} = \frac{\hat{q}^2}{(1-\hat{q}^2 s)}$$

En el caso de que nuevamente la frecuencia de equilibrio de A_2 sea muy baja ($\hat{q} \approx 0$) tenemos que $1 - \hat{q}^2 s \approx 1$, lo que nos deja una excelente aproximación en

$$\hat{q} \approx \sqrt{\frac{u}{s}} \tag{4.40}$$

4.5.1 El principio de Haldane-Muller

El principio de Haldane[9]-Muller[10] se relaciona con la reducción en el **fitness medio** de una población como balance entre selección y mutación. En el caso de **selección direccional** e ignorando el efecto de las mutaciones recurrentes, en algún momento el alelo favorecido (A_1) se fijará y entonces $\bar{w} = 1 - 2pqhs - q^2 s = 1$, ya que $q = 0$. Sin embargo, cuando existe mutación recurrente, el alelo desfavorecido volverá a aparecer en forma recurrente y, por lo tanto, la frecuencia de equilibrio será tal que $\hat{q} > 0$. En el caso particular que $h = 0$ (cuando A_1 domina completamente a A_2), entonces el **fitness medio** de la población será $\bar{w} = 1 - q^2 s$ ya que $2pqhs = 0$. Si sustituimos el resultado de la ecuación (4.40) en la expresión anterior, tenemos que

$$\bar{w} = 1 - q^2 s \approx 1 - \left(\sqrt{\frac{u}{s}}\right)^2 s = 1 - u \tag{4.41}$$

El cambio en el **fitness medio** (que llamaremos L[11]) correspondiente a esta **carga mutacional** será la diferencia entre el fitness máximo (igual a 1) y el producto del equilibrio mutación selección, es decir

$$L = w_{max} - \bar{w} = 1 - (1 - u) = u \tag{4.42}$$

es decir, $L = w_{max} - \bar{w}$ es igual a la tasa mutacional u.

En el caso de **dominancia incompleta** ($0 < h < 1$), usando la aproximación de la ecuación (4.38) el **fitness medio** será

$$\bar{w} = 1 - 2\hat{p}\hat{q}hs - \hat{q}^2 s = 1 - 2phs\frac{u}{hs} - \left(\frac{u}{hs}\right)^2 s = 1 - 2pu - \frac{u^2}{h^2 s} \tag{4.43}$$

[9] John Burdon Sanderson Haldane FRS (5 de noviembre de 1892 - 1 de diciembre de 1964), fue un científico británico conocido por sus trabajos en fisiología, genética, biología evolutiva y matemáticas. Junto con Ronald Fisher y Sewall Wright fue uno de los fundadores del neodarwinismo, síntesis de la teoría evolutiva. Haldane profesaba abiertamente sus convicciones socialistas, marxistas y humanistas, siendo un ateo confeso, lo que lo llevó a emigrar a la India en 1956. Entre sus contribuciones notables se encuentran las relacionadas al ligamiento genético, cinética enzimática, el origen de la vida y numerosas contribuciones en genética de poblaciones, entre ellas la propuesta del mecanismo poblacional relacionado al equilibrio alélico en la anemia de células falciformes.

[10] Hermann Joseph Muller (21 de diciembre de 1890 - 5 de abril de 1967) fue un genetista y educador estadounidense, más conocido por sus trabajos sobre los efectos fisiológicos y genéticos de la radiación (mutagénesis). Fue galardonado con el Premio Nobel de Medicina y Fisiología en 1946 por sus descubrimientos acerca del efecto de mutagénesis inducidos por los rayos-X. Durante su vida trabajó inicialmente en EE. UU., se mudó a Alemania en 1932 para realizar una estadía de sabático, pero este fue interrumpido por ascensión del régimen nazi. Se mudó a la Unión Soviética donde fundó un laboratorio muy productivo en el estudio de la «mosca de la fruta» *Drosophila melanogaster*, pero la hegemonía de la visión de Lysenko en la genética y las persecuciones del estalinismo lo llevaron a volver a EE. UU. (con un breve pasaje por Edimburgo), donde culminó su carrera en la Universidad de Indiana.

[11] Esta denominación viene del inglés, ya que el término «carga mutacional» corresponde a *mutational load* en dicho idioma.

Teniendo en cuenta que $p \approx 1$ y despreciando el término $\frac{u^2}{h^2 s}$ por ser extremadamente pequeño para las tasas mutacionales habituales (ver, por ejemplo, Drake et al. 1998), o tener en cuenta que $u^2 \approx 0$), tenemos entonces que en este caso

$$\bar{w} \approx 1 - 2u \tag{4.44}$$

por lo que el cambio de **fitness medio** será ahora

$$L = w_{max} - \bar{w} = 1 - (1 - 2u) = 2u \tag{4.45}$$

En ambos casos la reducción en el **fitness medio** es independiente del efecto del coeficiente de selección. En otras palabras, más allá de lo fuerte o débil que sea el efecto de la selección contra un alelo perjudicial, la **carga mutacional** (por ejemplo, la pérdida en fitness de la población) es solo proporcional a la tasa de mutación. Esta pérdida es equivalente a u en el caso de dominancia completa, y a $2u$ en el caso de dominancia incompleta. Visto de otra manera, cuando el efecto de la mutación es fuerte, la frecuencia de equilibrio de esta será mucho más baja que cuando su efecto es más suave; este balance es la razón de la independencia del **coeficiente de selección**. A este efecto se lo conoce como **principio de Haldane-Muller**.

Ejemplo 4.4

El ejemplo más conocido del uso del principio (a veces conocido como «teorema») de Haldane-Muller es la estimación de la tasa de mutación a partir de genotipos deletéreos. En particular, el ejemplo más conocido es la **enfermedad de Huntington** (a veces conocida como «corea de Huntington») una enfermedad neurodegenerativa llamada así en honor de George Huntington, médico norteamericano que la describió en 1832. Se trata de una enfermedad producto de alelos deletéreos en el gen de la Huntingtina, una proteína vinculada al transporte celular y a factores de transcripción, que interacciona con cerca de otras 100 proteínas. En la secuencia, esta proteína tiene repetidos de la secuencia CAG, que codifica para la Glutamina (por lo que se trata de una región **PoliQ**). El largo «normal» del número de repetidos es menor a 27, de 27 a 35 es una situación intermedia con cierto riesgo en la descendencia, de 36 a 39 se clasifica como de penetrancia incompleta y con más de 40 repetidos se trata de penetrancia total. El modo de acción génica es de **dominancia completa** del alelo perjudicial. Hartl y Clark (2007) dan un ejemplo del cálculo de la tasa mutacional a partir de los datos de afectados por este gen, ejemplo que adaptaremos aquí. El objetivo será entonces tener un rango de estimaciones de la tasa mutacional u.

Si bien se trata de un gen con alelo completamente dominante, la disminución en fitness de los heterocigotos se estima en $w_{12} = 0{,}81$ con relación al normal $w_{11} = 1$ ya que el desarrollo de la enfermedad es en la edad adulta, usualmente después de comenzado la etapa reproductiva. Los homocigotos A_2A_2 no se consideran ya que la frecuencia de ellos es totalmente despreciable (tendrían que tener dos padres con la enfermedad, que ya es rara de por sí). Con estos datos, $1 - hs = 0{,}81 \Leftrightarrow hs = 1 - 0{,}81 = 0{,}19$. En algunas poblaciones europeas, las cifras conocidas indican entre 4 y 15 individuos afectados por cada 100.000 habitantes, es decir una frecuencia de entre $\mathbb{P}(\text{Af}) = 4 \times 10^{-5}$ y $\mathbb{P}(\text{Af}) = 1{,}5 \times 10^{-4}$. Si los afectados son los heterocigotos (las cifras anteriores), es decir $2pq =$ afectados, asumiendo

que $p \approx 1$ (en la cifra mayor de afectados $p = 0{,}99985$), entonces una buena aproximación de la frecuencia de A_2 es

$$2\hat{q} \approx 2pq = \mathbb{P}(\text{Af}) \Rightarrow \hat{q} \approx \frac{\mathbb{P}(\text{Af})}{2} \tag{4.46}$$

Como $h \neq 0$ (porque $hs \neq 0$) la ecuación a utilizar es $\hat{q} = \frac{u}{hs}$, que da lugar a $\hat{u} = \hat{q}hs$ (el $\hat{u}$ lo usamos para indicar que se trata de una «estimación» que hacemos del «verdadero valor»). Sustituyendo los valores correspondientes tenemos que la tasa mutacional varía entre $\hat{u} = qhs = \frac{0{,}19 \times (4 \times 10^{-5})}{2} = 3{,}86 \times 10^{-6}$ y $\hat{u} = qhs = \frac{0{,}19 \times (1{,}5 \times 10^{-4})}{2} = 1{,}425 \times 10^{-5}$.

PARA RECORDAR

- En nuestro modelo de un locus con dos alelos, con **dominancia incompleta**, asumiendo que la fuerza de selección contra el alelo A_2 consigue mantener la frecuencia del mismo muy baja (y, por lo tanto $q \approx 0$), si los alelos A_2 se reintroducen con una tasa u, entonces el balance entre la mutación y la selección está dado aproximadamente por $\hat{q} \approx \frac{u}{hs}$.
- En el caso de que $h = 0$, es decir cuando A_1 es dominante sobre A_2, entonces una aproximación razonable a la frecuencia de equilibrio está dada por $\hat{q} \approx \sqrt{\frac{u}{hs}}$.
- **Principio de Haldane-Muller:** cuando $0 \leq h < 1$, la reducción en el **fitness medio** es independiente del efecto del coeficiente de selección. La **carga mutacional**, es decir la pérdida en fitness de la población es solo proporcional a la tasa de mutación ($L = w_{max} - \bar{w} = u$ en el caso de dominancia completa, $L = w_{max} - \bar{w} = 2u$ en el caso de dominancia incompleta).

4.6 La fuerza de la selección natural

Hasta ahora en el presente capítulo no le habíamos dedicado una atención especial al **coeficiente de selección.** Más aún, algunos resultados importantes de este capítulo son completamente independientes del **coeficiente de selección:** recordemos, por ejemplo, el punto de equilibrio en **sobredominancia** o **subdominancia** ($\hat{p} = \frac{h-1}{2h-1}$). Sin embargo, es este coeficiente de selección el centro de las diferencias de viabilidad entre los genotiposy, por lo tanto, representa aquello que permite que la frecuencia de los alelos vaya cambiando de generación en generación. ¿Es que el valor del **coeficiente de selección** no tiene importancia en las consecuencias de la selección?

En realidad, como es de esperar, este coeficiente sí tiene gran relevancia. Intuitivamente, un coeficiente de selección muy pequeño no hará grandes diferencias entre la viabilidad de los distintos genotiposy, por lo tanto, su impacto será menor en comparación con las otras fuerzas evolutivas como la deriva, la mutación y la migración. Pese a su importancia, hay dos hechos a destacar. En primer lugar, fijado el modelo de acción génica no hay un cambio importante en la forma que adquiere la evolución de las frecuencias alélicas en función del valor de s: como veremos en breve, la forma de esta gráfica permanece relativamente similar. En segundo lugar, pese a su importancia, no es trivial cómo medir el **coeficiente de selección**, y solo a partir de la evolución de las tecnologías de secuenciación y otras *-ómicas*

que ocurrieron a finales del siglo **XX** y principios del **XXI** es posible identificar mejor el efecto de variantes específicas que tienen impacto en el **fitness** de los organismos.

En un estudio reciente, Thurman y Barret (2016) revisaron la literatura en busca de estimaciones publicadas sobre la selección natural que actúa a nivel genético y encontraron más de 3000 estimaciones de coeficientes de selección, la mayoría en plantas (ya que son bastante más fáciles de estudiar), procedentes de 79 estudios distintos. Este tipo de estudio, que combina la información de otros muchos, se conoce como metaanálisis y se trata de estudios importantes para conocer la distribución de valores, así como la coherencia entre los resultados de diferentes estudios. En dicho estudio, los coeficientes de selección se distribuyeron de forma aproximadamente exponencial (negativa), lo que sugiere que el impacto de la selección a nivel genético es generalmente débil, pero en ocasiones puede ser bastante fuerte. Utilizando un modelo mixto lineal generalizado, los autores llegaron a estimar un coeficiente de selección medio de $s = 0{,}095$ (moda *a posteriori*, con un intervalo del 95 % HPD de $0{,}066 - 0{,}124$). Hay que recordar que en la distribución exponencial la media es mayor que la mediana ($\mu = \frac{1}{\lambda}$ y mediana $\frac{\ln 2}{\lambda}$). Por otro lado, estos autores observaron que la selección fue más fuerte cuando se midió en escalas de tiempo más cortas, siendo la magnitud media de s mayor para los estudios que midieron la selección en una sola generación. Finalmente, los autores destacan las dificultades inherentes a esta medición, la diferencia entre los modelos y parametrizaciones usados, y la dificultad de incorporar la estimación del **efecto de los heterocigotos** h.

En la figura 4.16 se puede apreciar la evolución de la frecuencia del alelo A_1 bajo **dominancia incompleta** en diferentes configuraciones. La gráfica de la izquierda corresponde a un **efecto de los heterocigotos** de $h = \frac{1}{4}$, la del medio a $h = \frac{1}{2}$ y la de la derecha a $h = \frac{3}{4}$ representando un conjunto de situaciones intermedias para dicho efecto ($h = \frac{1}{2}$ se corresponde a un modelo completamente aditivo). En cada gráfica, las distintas curvas se corresponden con diferentes **coeficientes de selección**, entre $s = 0{,}05$ (rojo) y $s = 0{,}25$ (negro). Lo primero que observamos es que, como decíamos antes, no parece existir gran diferencia en las formas de las curvas, las que lucen como una S (por ejemplo, aproximadamente **sigmoidales**). Una segunda observación en las gráficas de la figura 4.16 es que, excepto la curva roja (que corresponde al menor s) donde no tenemos una visión completa, el comportamiento es bastante similar: un breve comienzo de evolución lenta que rápidamente se acelera, llevando las curvas a una evolución casi vertical, para finalmente aplanarse y comenzar a acercarse en forma asintótica a la «frecuencia de equilibrio», que en este caso se corresponde con la fijación ($p = 1$) del alelo A_1. Otra vez, excepto para el coeficiente de selección más bajo (curva roja), en el resto de las configuraciones de h y s que mostramos las poblaciones se acercan a la fijación del alelo A_1 en menos de 200 generaciones y la mitad del camino hacia la fijación ($p = 0{,}5$) generalmente en menos de 100 generaciones. Más aún, parece haber una interacción entre coeficientes de selección bajos y $h \to 1$ que potencia el efecto del primero (la curva roja en la gráfica de la derecha se encuentra muy «aplanada»).

Veamos ahora qué ocurre cuando el modo de acción génica es la **sobredominancia**. En la figura 4.17 se muestran diferentes configuraciones del **efecto de los heterocigotos** y del **coeficiente de selección** para dicho modo. Ahora el punto de equilibrio será a una frecuencia intermedia, como vimos en la sección «Selección estabilizadora» ($\hat{p} = \frac{h-1}{2h-1}$). Sin embargo, el comportamiento de las curvas sigue siendo aproximadamente sigmoidal, como en el caso de **dominancia incompleta** (selección direccional). Con **sobredominancia**, el punto de equilibrio se va haciendo menor a medida que h se hace más negativo ($h \to -1$). Eso implica que la «carrera» por alcanzarlo disminuya de izquierda a derecha en las gráficas de la

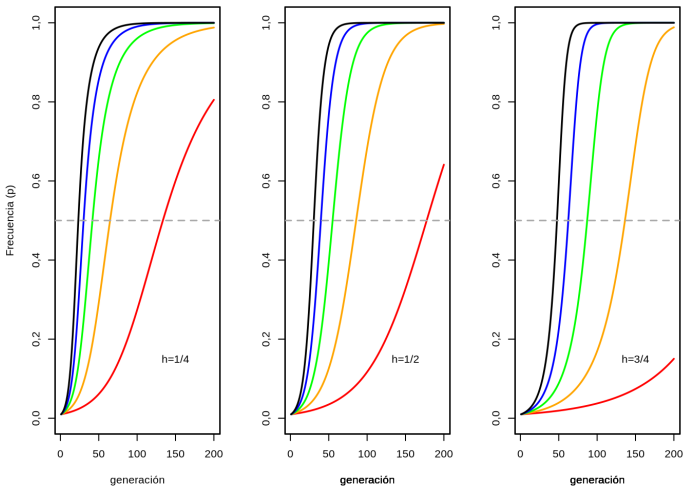


Figura 4.16. Evolución de las frecuencias bajo **dominancia incompleta**, de acuerdo con diferentes niveles del **coeficiente de selección** y del **efecto de los heterocigotos**. Coeficientes de selección $s = 0{,}05$ a $s = 0{,}25$, colores rojo, anaranjado, verde, azul y negro. Izquierda $h = \frac{1}{4}$, centro $h = \frac{1}{2}$ y derecha $h = \frac{3}{4}$.

figura 4.17, lo que hace que no sean directamente comparables. De todas formas, se aprecia claramente que menores coeficientes de selección implican tiempos mayores para alcanzar el punto de equilibrio.

Veamos ahora qué impacto tienen los coeficientes de selección en el tiempo hasta alcanzar la cercanía del punto de equilibrio correspondiente. En la figura 4.18 se ve cuántas generaciones son necesarias para quedar a menos de 1 % del punto de equilibrio (o fijación) de la frecuencia del alelo A_1, dependiendo del coeficiente de selección, para dos modos de acción diferentes.

Claramente, las curvas no son lineales y a medida que vamos reduciendo el coeficiente de **selección** (o sea, si nos movemos por la curva de derecha a izquierda), la misma se vuelve cada vez más empinada. Dicho de otra forma, coeficientes de selección muy pequeños requerirán un número desproporcionadamente alto de generaciones para acercarse a una situación de equilibrio (o de fijación, en el caso de **dominancia incompleta**).

Para terminar con esta sección hagamos una reflexión sobre los tiempos implicados para que la selección natural pueda actuar. En especies de mamíferos de gran tamaño, como los bovinos, por ejemplo, el intervalo generacional es del orden de los 5 años. Con coeficientes de selección débiles (del orden de $s \approx 0{,}01$) y **dominancia incompleta**, como vimos en la figura 4.18, entre 500 y 1000 generaciones son necesarias para llegar al entorno de la fijación de un alelo. Esto significa entre 2500 a 5000 años. Con fuerzas selectivas del orden de $s \approx 0{,}05$ son necesarias unas 175 generaciones, es decir unos 875 años. Coeficientes de selección mayores no son frecuentes para la mayoría de las características, pero con coeficientes de selección del orden de $s \approx 0{,}10$ tardaremos unas 85 generaciones, es decir unos 425 años. Con especies anuales, frecuentes en plantas, los tiempos son claramente inferiores (e iguales al número de generaciones) y en casos de especies que se reproducen más de una vez por año, estos tiempos son aún mucho menores. Si recordamos de la sección «Probabilidad

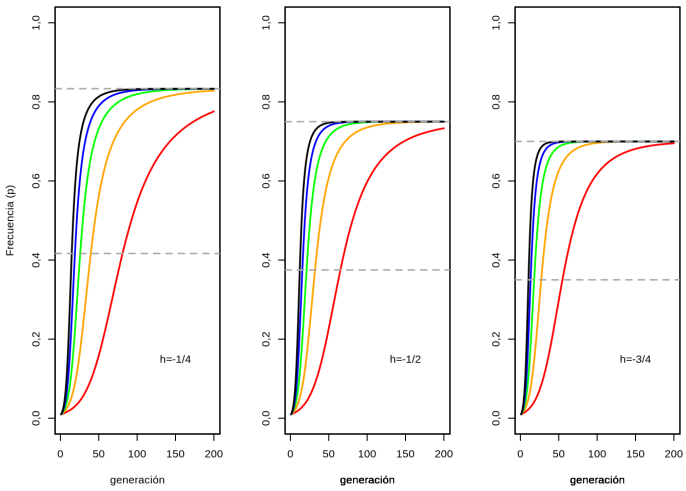


Figura 4.17. Evolución de las frecuencias bajo **sobredominancia**, de acuerdo con diferentes niveles del **coeficiente de selección** y del **efecto de los heterocigotos.** Coeficientes de selección $s = 0{,}05$ a $s = 0{,}25$ (en pasos de 0,05), colores rojo, anaranjado, verde, azul y negro. Izquierda $h = -\frac{1}{4}$, centro $h = -\frac{1}{2}$ y derecha $h = -\frac{3}{4}$.

de fijación y tiempos de fijación» que por efecto de la deriva el tiempo (en generaciones) a la fijación de un alelo que parte de una frecuencia de $p = 0{,}01$ (como en la figura 4.18) es de $\approx 4N$ (ya que $p \approx 0$), aún con coeficientes de selección tan bajos como $s = 0{,}01$ el efecto sistemático de esta prevalecerá sobre la deriva.

Además, como ya vimos en la sección «Equilibrio selección-mutación», su valor es fundamental también a la hora de fijar el punto de equilibrio de las frecuencias alélicas. Un valor alto de s contribuirá a un valor de equilibrio $\hat{q}$ muy cercano a cero, dentro del alcance de los efectos de la deriva y, por lo tanto, de su pérdida.

PARA RECORDAR

- El valor del **coeficiente de selección** es fundamental para conocer el tiempo a la fijación de un alelo bajo **dominancia incompleta** o de alcanzar el equilibrio en **sobredominancia**.
- Los coeficientes de selección son difíciles de determinar bajo selección natural, pero diversos estudios confirman que en su mayoría tienen un valor pequeño, con una distribución del tipo exponencial.
- La forma de la evolución de las frecuencias alélicas bajo **dominancia incompleta** y **sobredominancia** es aproximadamente sigmoidal, con una aceleración en frecuencias más intermedias y una posterior meseta cerca de la fijación o del punto de equilibrio.

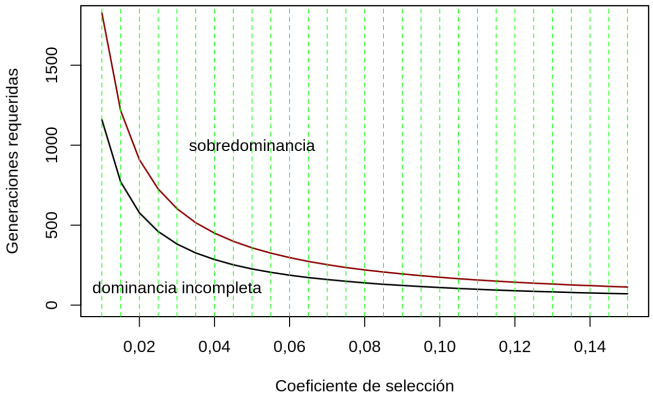


Figura 4.18. Número de generaciones requeridas para quedar a menos de 1 % del punto de equilibrio bajo diferentes **coeficientes de selección** (entre $s = 0{,}01$ y $s = 0{,}15$, de a 0,005) y en dos modos de acción génica (negro **dominancia incompleta** $h = \frac{1}{2}$, rojo **sobredominancia** $h = -\frac{1}{2}$), partiendo de una frecuencia inicial $p = 0{,}01$. Claramente el comportamiento no es lineal y a medida que el coeficiente de selección disminuye el número de generaciones se hace mucho mayor.

4.7 Equilibrio selección-deriva

En el presente capítulo, hasta ahora solo hemos mencionado el efecto de la deriva como usualmente pequeño y, por lo tanto, afectando el resultado de la selección solo en el entorno de los puntos críticos del proceso, es decir en torno a los puntos de equilibrio, fijación o pérdida, de acuerdo con el tipo de selección, pero sin explicitar en forma matemática dicha relación. Es tiempo entonces de presentar esta relación en una forma tal que nos permita evaluar el impacto de cada una de las dos fuerzas en este entorno. La matemática implicada es bastante más compleja que la usada en este capítulo, ya que la matemática de los procesos estocásticos lo es, pero por suerte no deberemos hacer un gran esfuerzo ya que las herramientas las pusimos a punto en el capítulo «Deriva genética», en particular en la sección «Aproximación de difusión» (si antes decidiste no prestarle mucha atención a dicha sección, es probable que sea tiempo de ir de nuevo a ella).

Supongamos que tenemos un locus autosomal con dos alelos en un organismo diploide, con un modo de acción completamente aditivo, es decir $h = \frac{1}{2}$ y con coeficiente de selección s. Como recordarás de antes, el cambio de frecuencia por generación esperado era

$$\Delta_s p = \frac{pqs[ph + q(1-h)]}{\bar{w}} \tag{4.47}$$

que sustituyendo los valores con $h = \frac{1}{2}$ nos da

$$\Delta_s p = \frac{pqs[\frac{1}{2}p + q(1-\frac{1}{2})]}{\bar{w}} = \frac{pqs[\frac{1}{2}(p+q)]}{\bar{w}} \quad \therefore$$
$$\Delta_s p = \frac{pqs}{2}\frac{1}{\bar{w}}$$

Esto era el cambio por generación en la frecuencia de p debido a la selección. Veamos cómo entra este resultado en el equilibrio entre selección y deriva. Para ello vamos a recurrir a nuestra «Aproximación de difusión». Si recuerdas la idea, vamos a pasar de un tiempo discreto en generaciones (y con cambios discretos en frecuencias) a una aproximación suficientemente buena en tiempo y frecuencias continuas. En esta aproximación considerábamos dos fuerzas, una sistemática con dirección determinística, que notamos como $M(x)$ (pero que técnicamente se conoce como «coeficiente de deriva», lo que es algo confuso ya que a veces usamos el término *deriva* para referirnos al azar) y una segunda fuerza aleatoria que notamos como $V(x)$ (y que técnicamente se conoce como «coeficiente de difusión»). En el capítulo anterior, como todavía no habíamos hablado de la selección pusimos la restricción de que $M(x) = 0$, es decir que solo analizamos el efecto del movimiento al azar de las frecuencias. Ahora es tiempo de usar $M(x)$ a nuestra conveniencia, en particular para modelar el efecto de la selección (que como ya vimos, es una fuerza determinística).

Si consideramos que en un intervalo de tiempo infinitesimal el cambio en el **fitness medio** será despreciable, sustituyendo p por x en la ecuación (4.7) para no confundirnos con el punto particular p, nos permite aproximar

$$\lim_{\delta t \to 0} \frac{\Delta_s x}{\delta t} \approx \frac{sx(1-x)}{2} = M(x) \tag{4.48}$$

Por otro lado, recordarás de la sección «Solución de las ecuaciones» que

$$V(x) = \frac{x(1-x)}{2N_e} \tag{4.49}$$

Tenemos entonces todo para estimar la probabilidad de fijación de un alelo A_1 con la ventaja selectiva que definimos antes. Para el cálculo de esta probabilidad, si recuerdas de la sección «Solución de las ecuaciones», tenemos primero que calcular la **función de escala** $G(x)$, como fue definida en la ecuación (3.58), es decir

$$G(x) = \exp[-2\int^x \frac{M(y)}{V(y)}dy] \tag{4.50}$$

Sustituyendo los valores específicos de $M(x)$ y $V(x)$ obtenidos antes (con el cambio de notación de y en lugar de x en el integrando), tenemos que

$$G(x) = \exp\left[-2N_e s \int^x \frac{y(1-y)}{y(1-y)}dy\right] = e^{-2N_e sx} \tag{4.51}$$

Por lo tanto, de acuerdo con la ecuación (3.61), sustituyendo la notación $u(p)$ anterior por la notación $\pi_1(p)$ para la probabilidad de fijación del alelo con frecuencia p (a fin de no confundirnos con la tasa de mutación que también usamos u)

$$\pi_1(p) = \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx} \tag{4.52}$$

sustituyendo los valores de (4.51) en (4.52) tenemos

$$\pi_1(p) = \frac{\int_0^p e^{-2N_e sx}dx}{\int_0^1 e^{-2N_e sx}dx} = \frac{1 - e^{-2N_e sp}}{1 - e^{-2N_e s}} \tag{4.53}$$

Es decir, arribamos a la probabilidad de fijación para este alelo A_1 con ventaja selectiva s del genotipo A_1A_1 respecto al homocigoto A_2A_2, en un locus aditivo que se comporta como completamente aditivo, a partir de una frecuencia determinada p en la que se encuentre dicho alelo. Esto es muy importante: ahora la probabilidad de fijación no será como bajo la acción exclusiva de la deriva, que habíamos visto era $\pi_1(p) = p$, ni tampoco estará determinada por h y s, sino por la combinación de estos efectos.

Veamos qué ocurre con un alelo que recién arriba a la población y que tiene las condiciones definidas antes, es decir $h = \frac{1}{2}$ básicamente. Una nueva mutación que arriba a una población de N individuos diploides tiene frecuencia $p = \frac{1}{2N_e}$ (asumiendo que $N = N_e$). Sustituyendo en la ecuación (4.53), tenemos

$$\pi_1(\frac{1}{2N_e}) = \frac{1 - e^{-2N_e sp}}{1 - e^{-2N_e s}} = \frac{1 - e^{-2N_e s\frac{1}{2N_e}}}{1 - e^{-2N_e s}} \therefore \tag{4.54}$$

$$\pi_1(\frac{1}{2N_e}) = \frac{1 - e^{-s}}{1 - e^{-2N_e s}} \tag{4.55}$$

Si s es muy pequeño (como usualmente lo es, para la mayor parte de las mutaciones que arriban), entonces $e^{-s} \approx 1 - s$ y si además $2N_e s$ es suficientemente grande (es decir N_e es muy grande) para que $e^{-2N_e s} \approx 0$, entonces la ecuación (4.55) se reduce a

$$\pi_1(p) = \frac{1 - e^{-s}}{1 - e^{-2N_e s}} \approx \frac{1 - (1 - s)}{1 - 0} \therefore \tag{4.56}$$

$$\pi_1(p) \approx s \tag{4.57}$$

El resultado anterior es un caso particular de un resultado más general que permite una aproximación a la probabilidad de fijación de un alelo, que sería

$$\pi_1(p) \approx 2(1-h)s \tag{4.58}$$

El sentido intuitivo de la aproximación anterior ya lo hemos discutido en otras instancias. Para los alelos «raros», por ejemplo, las nuevas mutaciones que arriban a la población, la mayoría se van a encontrar en los heterocigotos y por ello no es de extrañar esta dependencia de la probabilidad de fijación de los nuevos alelos en el **fitness relativo** de los heterocigotos. Si sustituimos $h = \frac{1}{2}$ en (4.58) nos da, obviamente, $\pi_1(p) \approx s$.

Nota de precaución

En algunos textos (por ejemplo, Walsh & Lynch, Rice o incluso algunos trabajos de Kimura) se opta por una parametrización distinta de los **fitness** relativos, donde $w_{11} = 1 + 2s;\ w_{12} = 1 + s;\ w_{22} = 1$.
Empleando esta parametrización, utilizando el mismo abordaje a la aproximación por difusión que el planteado en este libro, se llega a la siguiente probabilidad de fijación para un alelo bajo modelo completamente aditivo:

$$\pi_1(p) = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}}$$

Esto puede llevar a una aparente incongruencia. ¿Cómo se establece una equivalencia entre los dos resultados?
Notemos que la parametrización empleada, si se hace con referencia al homocigota A_1A_1 es equivalente a $w_{11} = 1;\ w_{12} = 1 - s;\ w_{22} = 1 - 2s$
donde restamos en cada caso $2s$, la diferencia entre el máximo w_{11} y el mínimo w_{22}. Pero bajo la parametrización empleada en este libro, con $h = \frac{1}{2}$, tenemos $w_{11} = 1;\ w_{12} = 1 - \frac{1}{2} \times s;\ w_{22} = 1 - s$.
Al comparar ambos planteos queda claro que el valor de s empleado en la parametrización de este libro equivale al $2s$ empleado en la otra parametrización. Por lo tanto, el resultado al que arribamos en el desarrollo de este libro es, bajo la otra parametrización

$$\pi_1(p) = \frac{1 - e^{-2N_e(2s)p}}{1 - e^{-2N_e(2s)}} = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}}$$

quedando clara la equivalencia de ambos resultados. Este caso muestra la importancia que posee analizar bajo qué modelos y parámetros se llega a un resultado: claramente, no es lo mismo emplear una ecuación que otra si no se tiene en cuenta qué parametrización se está empleando.

Para poner los resultados anteriores en perspectiva, pongamos algunos números. Si la ventaja selectiva del nuevo alelo es, por ejemplo $s = 0{,}05$, para una población de $N_e = 500$, entonces en el modelo aditivo, utilizando la ecuación (4.55) la probabilidad de fijación es de $\pi_1(\frac{1}{2N_e}) = 0{,}0488 \approx 0{,}05 = s$, que es equivalente a la aproximación por la ecuación (4.57). En palabras, pese a que la ventaja selectiva de esta mutación es relativamente importante, la probabilidad de perderla es del 95 %. Si en cambio la población es ahora de 10 veces el tamaño de la anterior, podríamos intuir que el efecto de la deriva genética será mucho menor y, por lo tanto, la probabilidad de fijación aumentará. Desafortunadamente nuestra intuición es incorrecta y el resultado es casi idéntico. Si bien la deriva genética disminuye con el tamaño poblacional, también lo hace la frecuencia inicial de la nueva mutación, lo que en última instancia se ve reflejado en la aproximación de la ecuación (4.57), donde $\pi_1(\frac{1}{2N_e}) \approx s$ la probabilidad de fijación no depende del tamaño poblacional. Una forma interesante de interpretar este resultado se encuentra en John H. Gillespie (2004): si un nuevo alelo arriba a una población, se encontrará en un heterocigoto; su continuidad en la población va a depender del muestreo en el número de descendientes que tiene ese heterocigoto y no en el tamaño de la población.

PARA RECORDAR

- En nuestro modelo de un locus autosómico en un organismo diploide, **asumiendo que el efecto de los alelos es aditivo**, la probabilidad de fijación de un alelo A_1A_1 que se encuentra en frecuencia p y que en homocigosis tiene ventaja s sobre el genotipo A_2A_2 es

$$\pi_1(p) = \frac{1 - e^{-2N_e sp}}{1 - e^{-2N_e s}}$$

- En el caso particular de que además de todo lo anterior (locus autosómico en un organismo diploide, efecto de los alelos es **aditivo**), la **probabilidad de fijación de un nuevo alelo** A_1A_1 que en homocigosis tiene ventaja s sobre el genotipo A_2A_2 es

$$\pi_1(\frac{1}{2N_e}) = \frac{1 - e^{-s}}{1 - e^{-2N_e s}}$$

Esto es debido a que al tratarse de un alelo nuevo su frecuencia inicial es $p = \frac{1}{2N}$ y asumiendo $N = N_e$, sustituyendo p por ese valor queda la expresión reducida anterior.

- En el caso en que s es muy pequeño, entonces $e^{-s} \approx 1 - s$ y si además $2N_e s$ es suficientemente grande para que $e^{-2N_e s} \approx 0$, entonces la probabilidad de fijación se aproxima razonablemente a $\pi_1(p) \approx s$.
- En general, bajo los supuestos anteriores, la aproximación se generaliza a $\pi_1(p) \approx 2(1 - h)s$.

4.8 Otros tipos de selección y complejidades

En todo este capítulo nos hemos enfocado en entender las principales formas de selección natural, así como de la interacción entre algunas fuerzas evolutivas (mutación, deriva genética y selección). Pese a la aparente complejidad de algunos de los temas tratados, nuestro enfoque ha sido muy restringido para adecuarlo al nivel del presente curso. Esto deja por fuera una gran cantidad de formas de selección o de tópicos directamente relacionados. Pese a que no trataremos estos temas, resulta importante conocer las limitaciones de nuestras aproximaciones y como se distancian de la complejidad biológica, por lo que a continuación vamos a mencionar varios de estos tópicos.

Selección gamética

En ciertas plantas el ciclo de vida implica una fase haploide y otra diploide. En estos organismos las viabilidades asociadas a los alelos suelen ser diferentes en las dos fases, una con genotipos diploides y la otra con haplotipos (es decir, los alelos en estado haploide). A la selección en esta última se le conoce como selección gamética. Esto implica una alternancia en las formas de selección y los fitness de cada una de las fases. Si los **fitness** de A_1 y A_2 en esta fase son v_1 y v_2, entonces (como se muestra en Hartl y Clark (2007)) luego de una fase de selección en los haplotipos las frecuencias serán $p^* = pv_1/\bar{v}$ y $p^* = qv_2/\bar{v}$, con $\bar{v} = pv_1 + qv_2$. Por lo tanto, bajo equilibrio de Hardy-Weinberg estas serán las frecuencias con la que aparecerán los genotipos en la fase diploide, que luego serán seleccionados con viabilidad w_{11}, w_{12} y w_{22}. Por lo tanto, después de esta selección en la fase diploide, la frecuencia de A_1 será

$$p' = \frac{p^2 w_{11} v_1^2 + pq w_{12} v_1 v_2}{p^2 w_{11} v_1^2 + 2pq w_{12} v_1 v_2 + q^2 w_{22} v_2^2} \tag{4.59}$$

y la forma de selección será determinada por la combinación de parámetros v y w.

Selección por fecundidad

Cuando el fitness reproductivo no depende exclusivamente de cada genotipo sino de las combinaciones de parejas reproductivas, entonces la complejidad de los modelos se complica enormemente. Imaginemos, por ejemplo, que machos de un genotipo son atraídos por hembras de otros genotipos con diferente probabilidad, que son diferentes de las probabilidades recíprocas. Ahora, para un locus con dos alelos ya no tenemos 3 **fitness** diferentes sino 9, ya que las combinaciones recíprocas pueden ser diferentes (macho de un genotipo y hembra de otro y el cruzamiento recíproco son diferentes). En general esas combinaciones implican

muchos parámetros y eso complica las determinaciones, más aún que los modelos más realistas no suelen reducirse a sumas o productos de pares de genotipos.

El efecto de la estructura de edades

Salvo en los casos de especies donde no hay solapamiento en la estructura de edades al momento de la reproducción, nuestro modelo es en general una simplificación considerable de lo que ocurre en la vida real. ¿Cómo afecta esto nuestros modelos? En principio es difícil de tratar en forma analítica, ya que al solaparse las generaciones, si la estructura de edades impacta, por ejemplo, en las viabilidades debemos incorporar este efecto en los modelos. Claramente estas complejidades exceden el alcance de nuestro curso.

Selección dependiente de la frecuencia

Hasta ahora hemos mencionado varias veces que nuestras conclusiones se referían exclusivamente a modelos en los que los que el **fitness** es **independiente de la frecuencia de los genotipos**. ¿Por qué es tan importante esa condición? En principio, si bien no vamos a trabajar los aspectos matemáticos del problema, si de alguna manera el **fitness** de los organismos depende de la frecuencia de los genotipos, entonces podríamos imaginarnos que se trata de una carrera donde la línea de meta se va corriendo como una función de la posición de los distintos corredores. Aún con reglas relativamente sencillas, excepto en ejemplos altamente estilizados, el tratamiento matemático suele ser más complejo de lo que pretendemos para este capítulo. Mencionaremos, sin embargo, un caso interesante: con determinada dependencia del fitness en las frecuencias de los genotipos, aún en presencia de **dominancia incompleta** (que usualmente fijaría uno de los dos alelos mediante selección direccional) pueden existir puntos de equilibrio estable intermedios. Si recuerdas lo que vimos en los distintos tipos de selección discutidos, en principio esta **selección estabilizadora** solo se producía con **sobredominancia**. Es decir, ahora entendemos mejor por qué no podemos igualar **sobredominancia** a la causa de la **selección estabilizadora**. Un ejemplo típico de una configuración de este tipo aparece en Hartl y Clark (2007). Si el **fitness** de los genotipos disminuye en forma proporcional a su frecuencia, con una constante c, entonces tenemos $AA : w_{11} = 1 - cp^2$ $Aa : w_{12} = 1 - 2cpq$ $aa : w_2 = 1 - cq^2$.
Haciendo las cuentas, $\Delta p = cpq(q - p)(p^2 - pq + q^2)/\bar{w}$ y, por lo tanto (asumiendo que $c > 0$), los puntos de equilibrio serán las frecuencias de p que hagan igual a cero el numerador. Estos puntos son $p = 0$, $p = 1$ (porque cpq multiplica todo y $p = 1 \Leftrightarrow q = 0$) y $p = \frac{1}{2}$ (cuando $p = q = \frac{1}{2} \Rightarrow p - q = 0$); la expresión $(p^2 - pq + q^2)$ es igual a $1 - 3pq = 3p^2 - 3p + 1 = 0 \Leftrightarrow p(1 - p) = \frac{1}{3}$, pero esto nunca ocurre porque esa expresión es una parábola cuyo máximo es en $p = \frac{1}{2}$ y es $\frac{1}{4} = 0{,}25$, por lo que no tiene ceros. O sea, despreciando los

puntos $p =$ y $p = 1$ que ya hemos discutido en otras oportunidades, en el punto de equilibrio intermedio ($p = \frac{1}{2}$) los nuevos fitness de cada genotipo serán $AA : w_{11} = 1 - \frac{c}{4}$ $Aa : w_{12} = 1 - \frac{c}{2}$ $aa : w_2 = 1 - \frac{c}{4}$y, por lo tanto, el heterocigoto tendrá menos fitness que los homocigotos, es decir que en ese punto aparece como la acción de un efecto de **subdominancia**.

Ambientes heterogéneos y clinas

Como ya vimos en el caso de la mariposa *Biston betularia*, el **fitness** de los distintos genotipos puede variar en el tiempo y en el espacio. Las variaciones en el tiempo desplazarán las frecuencias en una forma relativamente permanente mientras los **fitness** sean estables, aunque les tome cierto tiempo esa aproximación.
En cambio, las diferencias de **fitness** en el espacio, es decir con la geografía, pueden mantener un nivel importante de polimorfismos a pesar de que los heterocigotos no sean el genotipo más ventajoso en ninguno de los ambientes. Un caso especial de variación espacial de las frecuencias son las clinas. Una **clina** es un gradiente en una característica de una especie a lo largo de su rango geográfico. Dicho de otra forma, es la variación gradual de una característica biológica de una especie con la distancia geográfica. En nuestro caso, una clina en las frecuencias alélicas, es decir una variación continua en las frecuencias alélicas con la posición geográfica puede estar relacionada con una presión selectiva determinada, aunque también puede ser producida por otras razones (por ejemplo, dos poblaciones a cierta distancia con frecuencias extremas y migración entre ellas). Más aún, en pequeñas distancias geográficas es teóricamente posible que la clina sea producida por deriva solamente. En este caso, suelen durar poco tiempo y, por lo tanto, se conocen como «clinas transitorias». Un ejemplo claro de clina por selección fue el presentado antes para la malaria y los alelos de la anemia de células falciformes. Otro ejemplo de clina se da en los grupos sanguíneos del locus ABO, por ejemplo, entre Europa (frecuencia alta del A) y Asia (frecuencia alta del B), aunque en este caso parece más producto de la migración que de ventajas selectivas. Por lo tanto, más allá del comportamiento «clinal» de una característica, suele ser dificultoso entender si su base es genética, en particular selectiva y en ese caso cuál es el agente de esa presión selectiva.

Selección diferencial en los sexos y en genes ligados al X

La selección diferencial en los dos sexos implica que en ambos hay diferencias en los **fitness relativos** de los genotipos y, por lo tanto, como usamos dos parámetros para cada uno (h y s), ahora necesitaremos 4 parámetros. Más aún, las frecuencias alélicas en los gametos pueden ser diferentes en ambos sexos y esto duplica los parámetros también. Finalmente, con este tipo de selección puede existir más de

un punto de equilibrio interno, o aún un punto de equilibrio estable polimórfico si la dirección de selección es opuesta en ambos sexos.
En los genes ligados al sexo hay varios problemas similares a la selección diferencial en sexos, aunque ahora para el **fitness** de los genotipos necesitamos tres parámetros: número de genotipos menos uno para cada sexo, pero en machos hay solo dos genotipos posibles porque hay un solo cromosoma X (en mamíferos, en aves es al revés). Las frecuencias alélicas también pueden ser diferentes en los gametos de ambos sexos, aunque como ya vimos antes la frecuencia de gametos en machos será igual a la de hembras en la generación anterior y la de hembras el promedio de machos y hembras en la anterior.

PARA RECORDAR

- Los modelos que hemos tratado *in extenso* en este capítulo son modelos extremadamente simplificados de cómo opera la selección en la realidad.
- En particular, para un tratamiento matemático razonable hemos considerado un locus autosomal de un organismo diploide con solo dos alelos. Normalmente los genes tienen gran cantidad de alelos y las categorías de **fitness** suelen ser difíciles de determinar.
- Uno de los supuestos fundamentales durante todo el capítulo fue que el **fitness relativo** de los genotipos no dependía de la frecuencia de estos. Cuando esto ocurre es necesario introducir esas funciones de dependencia en todas las ecuaciones de cambio de frecuencias. Esto las vuelve más complejas y suelen aparecer otros fenómenos novedosos, como la existencia de puntos de equilibrio interno aún en **dominancia incompleta**.
- La selección diferencial en los sexos, así como los genes ligados al cromosoma homogamético (X en mamíferos, en aves el Z, etcétera) presenta problemas adicionales por el número de parámetros de fitness requeridos, así como la diferencia de frecuencias en los alelos en los dos sexos.
- Los ambientes heterogéneos pueden contribuir a desarrollar clinas en las frecuencias alélicas, aunque también esto puede ser producto de la deriva genética (por corto tiempo) o de procesos fundacionales y migración.
- En especies que tienen dos fases de vida libre alternando, una como diploide y la otra como haploide, se produce el fenómeno de la selección gamética y es necesario adaptar las ecuaciones de cambio de frecuencia para acomodar los parámetros adicionales que corresponden a la fase haploide.
- En la selección por fecundidad es necesario contemplar las combinaciones de fenotipos de los dos sexos, por lo que el número de parámetros a estimar es usualmente mucho mayor que en el modelo fundamental.

4.9 Conclusión

Las poblaciones naturales experimentan cambios en las frecuencias alélicas de los loci que segregan en ellas. Estos cambios no son solo estocásticos: debido al éxito reproductivo diferencial que poseen los diversos genotipos en un ambiente estable (concepto de **fitness**) se producirán cambios sistemáticos en las frecuencias alélicas de la población, de modo tal que el **fitness** promedio de la población aumenta con el tiempo. A este proceso de cambio sistemático se le conoce como selección natural. Las características de este proceso estarán dictaminadas por varios factores, tales como el **fitness** relativo de cada genotipo, el efecto de heterocigotas y, naturalmente, la fuerza del régimen selectivo. En las poblaciones naturales existirá una interacción entre distintas fuerzas que generan cambios en las frecuencias alélicas (deriva genética, selección natural, mutación). El planteo de modelos matemáticos explícitos basados en asunciones claras, como el modelo para locus autosómico con dos alelos planteado en este capítulo o la aproximación por difusión, permite realizar estimaciones que tengan en cuenta estos procesos en simultáneo. Por último, es importante remarcar que la selección natural puede operar a distintos niveles en los sistemas biológicos, acorde a la complejidad de estos.

4.10 Actividades

4.10.1 Control de lectura

1. Describe los diferentes modos de acción génica y los regímenes selectivos asociados a cada caso. Discute qué dinámica se espera para las frecuencias alélicas de una población en cada caso.
2. Considera dos poblaciones que se enfrentan a un mismo régimen selectivo. Ambas se encuentran en igualdad de condiciones, salvo por el hecho de que una posee menor varianza genética que la otra. ¿Cómo afectará esto al cambio en el **fitness** de la población?
3. Describe cómo afecta la magnitud del coeficiente de selección s a las dinámicas de las frecuencias alélicas para los regímenes de selección direccional, estabilizadora y disruptiva.
4. Por efecto de la mutación un alelo perjudicial puede reaparecer recurrentemente en una población, disminuyendo el **fitness** medio de la misma en el equilibrio selección-mutación. Esta disminución, ¿depende de la magnitud de la selección contra el alelo?
5. Las poblaciones naturales tienen tamaño finito, y es de esperar que tanto la selección natural como la deriva genética influyan en los cambios de las frecuencias alélicas. ¿Cómo nos podemos aproximar al estudio de poblaciones teniendo en cuenta el efecto de estas fuerzas evolutivas?

4.10.2 ¿Verdadero o falso?

1. Bajo modo de acción génica de dominancia incompleta se espera un régimen de selección direccional, esperándose a largo plazo la fijación de uno de los alelos de la población.
2. El parámetro h (efecto del heterocigota) es un parámetro fijo, característico de cada locus.
3. Bajo equilibrio de selección-mutación y modo de acción génica de dominancia completa, la carga mutacional L depende únicamente del coeficiente de mutación u y del efecto heterocigota h.
4. La frecuencia de equilibrio esperada para regímenes de selección estabilizadora y disruptiva es la misma, y equivale a $\hat{p} = \frac{h-1}{2h-1}$
5. En una población de tamaño finito el efecto de la deriva genética siempre prevalecerá por sobre el de la selección natural.

Solución

1. **Verdadero.** Bajo el modo de acción génica de dominancia incompleta, los alelos heterocigotos tienen un fenotipo intermedio entre los alelos homocigotos dominantes y recesivos. En este caso, se espera un régimen de selección direccional en el que uno de los alelos tiende a aumentar su frecuencia en la población a largo plazo, hasta eventualmente fijarse.

2. **Falso.** El parámetro h (efecto del heterocigoto) sí será característico para cada combinación de alelos en un locus, pero representará el valor adaptativo de los individuos heterocigotas en una condición dada. Por lo tanto, el parámetro no es estrictamente fijo.

3. **Falso.** En la situaciónplanteada la carga mutacional L depende únicamente del coeficiente de mutación u (tasa de mutación), representando la reducción en la aptitud de la población debido a la acumulación de mutaciones perjudiciales.

4. **Verdadero.** Para selección estabilizadora, la frecuencia de equilibrio es $\hat{p} = \frac{h-1}{2h-1}$, lo cual es independiente del ambiente.

5. **Falso.** El efecto relativo de la deriva genética y la selección natural depende del tamaño de la población. En poblaciones pequeñas, la deriva genética puede tener un efecto significativo sobre el cambio de las frecuencias alélicas, incluso en presencia de selección natural. A medida que el tamaño poblacional considerado aumenta, es de esperar que los efectos de la selección se impongan por sobre el efecto de la deriva genética.

4.10.3 Ejercicios

Ejercicio 4.1

En *D. melanogaster* la hormona juvenil es un regulador clave de procesos como la metamorfosis y la reproducción. La enzima 3-hidroxi-3-metilglutaril-CoA sintasa (HMG-S) es parte de la vía que regula la síntesis de esta hormona. Una población de *D. melanogaster* se encuentra segregando dos alelos codificantes para la enzima: el alelo A_1 codifica para una enzima plenamente funcional, mientras que el alelo A_2 codifica para una enzima menos eficiente. Los individuos homocigotas A_2A_2 ven afectado su desarrollo, dejando en promedio la mitad de descendencia que los individuos que poseen una copia A_1.

a. Modela la situación planteada, estableciendo los **fitness** relativos w_{11}, w_{12} y w_{22} de los genotipos, así como el coeficiente de selección s y el coeficiente de efecto heterocigota h.

b. En un momento dado se genotipan individuos de la población, estableciéndose las siguientes frecuencias genotípicas

$$\begin{cases} f(A_1A_1) = 0{,}40 \\ f(A_1A_2) = 0{,}35 \\ f(A_2A_2) = 0{,}25 \end{cases}$$

Establece las frecuencias genotípicas esperadas en la siguiente generación.

Solución

a. De la letra se deduce que el fenotipo A_1A_1 y el fenotipo A_1A_2 tienen igual **fitness**, lo cual implica que el alelo A_1 es dominante; por lo tanto, el efecto heterocigota $h = 0$. Ambos genotipos son los que presentan mayor ventaja selectiva y, por lo tanto, tenemos **fitness** relativos $w_{11} = w_{12} = 1$. Por último, el hecho de que los organismos A_2A_2 dejen en promedio la mitad de descendencia implica que el **fitness** relativo $w_{22} = \frac{1}{2}w_{11} \Rightarrow w_{22} = 0{,}5$, lo cual implica un coeficiente de selección $s = 0{,}5$.

b. Luego del proceso selectivo se esperan las siguientes frecuencias genotípicas:

$$\begin{cases} f'(A_1A_1) = f(A_1A_1) \times \frac{w_{11}}{\bar{w}} \\ f'(A_1A_2) = f(A_1A_2) \times \frac{w_{12}}{\bar{w}} \\ f'(A_2A_2) = f(A_2A_2) \times \frac{w_{22}}{\bar{w}} \end{cases}$$

donde el **fitness** medio de la población $\bar{w}$ está dado por

$$\bar{w} = f(A_1)^2 w_{11} + 2f(A_1)f(A_2)w_{12} + f(A_2)^2 w_{22}$$

$$\bar{w} = 0{,}4 + 0{,}35 + 0{,}25 \times 0{,}5$$

Tenemos entonces en la siguiente generación

$$\begin{cases} f'(A_1A_1) = 0{,}4 \times \frac{1}{0{,}4+0{,}35+0{,}25\times 0{,}5} \approx 0{,}457 \\ f'(A_1A_2) = 0{,}35 \times \frac{1}{0{,}4+0{,}35+0{,}25\times 0{,}5} \approx 0{,}400 \\ f'(A_2A_2) = 0{,}25 \times \frac{0{,}5}{0{,}4+0{,}35+0{,}25\times 0{,}5} \approx 0{,}143 \end{cases}$$

Puedes corroborar que dichas frecuencias suman 1.

Coherente con lo esperado, se observa un aumento de las frecuencias de los genotipos portadores de A_1 y una disminución de la frecuencia de homocigotas A_2A_2.

Ejercicio 4.2

El hongo *Fusarium oxysporum* es un patógeno del tomate (*Solanum lycopersicum*). Se ha postulado que variantes alélicas del locus XSP10 (a partir de ahora, locus X para simplificar) se encuentran asociadas a la susceptibilidad al patógeno. Considera una población que segrega dos alelos, X_A (no susceptible a *F. oxysporum*) y X_B (susceptible), donde los **fitness** relativos de los genotipos X_AX_A, X_AX_B y X_BX_B son $w_{AA} = 1;\ w_{AB} = 0{,}44;\ w_{BB} = 0{,}3$. En la población se encuentra presente el patógeno antes mencionado.

a. Si la tasa de mutación $X_A \rightarrow X_B$ es $u = 10^{-8}$ y la tasa de mutación $X_B \rightarrow X_A$ es despreciable, ¿qué esperas que suceda con las frecuencias alélicas $f(X_A)$ y $f(X_B)$?

b. En un momento dado se comienza a emplear tiabendazol para disminuir la presencia del hongo en la población. Bajo esta nueva situación los **fitness** relativos de los genotipos son $w_{AA} = 1;\ w_{AB} = 0.8;\ w_{BB} = 0{,}7$. ¿Qué esperas que ocurra con las frecuencias alélicas bajo este nuevo régimen selectivo?

Solución

a. Los **fitness** relativos presentados corresponden a un modo de acción génica de dominancia incompleta. Esto implica un régimen de selección direccional, esperándose a largo plazo un aumento progresivo de la frecuencia del alelo X_A. No obstante, dado que la tasa de mutación $X_A \rightarrow X_B$ no es despreciable, se espera un aumento de esta frecuencia alélica

hasta alcanzarse un equilibrio entre selección y mutación. La frecuencia de equilibrio para el alelo X_B se aproxima en el equilibrio a

$$f(\hat{X_B}) \approx \frac{u}{hs}$$

Dado que $w_{AB} = 0{,}44 = 1 - hs \Rightarrow hs = 1 - 0{,}44 = 0{,}56$, por lo que

$$f(\hat{X_B}) \approx \frac{10^{-8}}{0{,}56}$$

punto en el cual se establece un equilibrio entre mutación y selección bajo el régimen selectivo. La frecuencia del alelo $f(X_A)$ en este punto de equilibrio será por lo tanto de $f(X_A) \approx 1 - (\frac{10^{-8}}{0{,}56})$.

b. El nuevo régimen selectivo continúa siendo de selección direccional, ya que el modo de acción génica es también en este caso la dominancia incompleta. No obstante, observemos que la disminución en la presencia de *F. oxysporum* relaja la presión selectiva sobre los organismos heterocigotas X_AX_B y X_BX_B (notemos que su **fitness** relativo aumentó). Es de esperar por lo tanto que la frecuencia de equilibrio (que llamaremos $f'(\hat{X_B})$). Notemos que en el nuevo régimen selectivo $w_{BB} = 1 - hs = 0{,}8 \rightarrow hs = 1 - 0{,}8 = 0{,}2$, por lo que la frecuencia de equilibrio para el alelo X_B es ahora

$$f'(\hat{X_B}) \approx \frac{10^{-8}}{0{,}2}$$

Notemos que, si bien esta frecuencia de equilibrio es baja, es

$$\frac{f'(\hat{X_B})}{f(\hat{X_B})} = \frac{(10^{-8}/0{,}2)}{(10^{-8}/0{,}56)} = \frac{0{,}56}{0{,}2} = 2{,}8$$

casi tres veces mayor que en el primer escenario planteado.

Ejercicio 4.3

Una población de tamaño $N_e = 200$ se encuentra segregando un locus bialélico (alelos A_1 y A_2). En promedio cada individuo deja aproximadamente un descendiente. En un momento dado las frecuencias genotípicas son

$$\begin{cases} f(A_1A_1) = 0{,}2 \\ f(A_1A_2) = 0{,}3 \\ f(A_2A_2) = 0{,}5 \end{cases}$$

a. ¿Cuál es la probabilidad de fijación de cada alelo?

b. Considera ahora que en el momento planteado las condiciones ambientales cambian, alterando el régimen selectivo. En esta nueva situación los **fitness** relativos de cada genotipo son

$$\begin{cases} w_{11} = 1 \\ w_{12} = 0{,}995 \\ w_{22} = 0{,}99 \end{cases}$$

Estima la probabilidad de fijación para cada alelo bajo el nuevo régimen selectivo.

Solución

a. El escenario planteado corresponde con un escenario selectivamente neutro (no existen diferencias en el **fitness** de los individuos) y con tamaño poblacional constante. En estas condiciones se asume que la única fuerza evolutiva es la deriva genética. Bajo el modelo de Wright-Fisher se puede plantear que la probabilidad de fijación de un alelo está dada por su frecuencia en el momento considerado.

Tenemos entonces para el alelo A_1 una frecuencia alélica de $f(A_1) = f(A_1A_1) + \frac{1}{2} \times f(A_1A_2) = 0{,}2+0{,}5\times0{,}3 = 0{,}35$. Dado que $f(A_1)+f(A_2) = 1$, $f(A_2) = 1-0{,}35 = 0{,}65$. Ambas frecuencias son equivalentes a la probabilidad de fijación de cada alelo.

b. El nuevo régimen selectivo plantea un modo de acción génica de dominancia incompleta, con ventaja selectiva (aparentemente pequeña) para los organismos portadores del alelo A_1. Por lo tanto, la deriva génica no será la única fuerza evolutiva actuando, existiendo también una influencia de la selección natural. Notemos que

$$w_{22} = 1 - s = 0{,}99 \Rightarrow s = 1 - 0{,}99 = 0{,}01$$

A su vez

$$w_12 = 1 - hs = 0{,}995 \Rightarrow hs = 1 - 0{,}995 = 0{,}005$$

$$h \times 0{,}01 = 0{,}005$$

$$h = \frac{0{,}005}{0{,}01} = 0{,}5$$

El modelo de acción génica es por lo tanto completamente aditivo. En este escenario la solución de las ecuaciones de la aproximación por difusión nos plantea que la probabilidad de fijación del alelo A_1, con frecuencia A_1, está dada por

$$\pi_{A_1}(f(A_1)) = \frac{1 - e^{-2N_e s f(A_1)}}{1 - e^{-2N_e s}}$$

donde tenemos $f(A_1) = 0{,}35$, $N_e = 200$, y $s = 0{,}01$, por lo que

$$\pi_{A_1}(0{,}35) = \frac{1 - e^{-2\times 200\times 0{,}01\times 0{,}35}}{1 - e^{-2\times 200\times 0{,}01}}$$

$$\pi_{A_1}(0{,}35) \approx 0{,}767$$

de lo que se deduce que la probabilidad de fijación del alelo A_2 será $1-(\frac{1-e^{-2\times 200\times 0{,}01\times 0{,}35}}{1-e^{-2\times 200\times 0{,}01}}) \approx 0{,}233$.

Es decir, el modelo planteado sugiere que aún una aparentemente pequeña ventaja selectiva para los organismos portadores del alelo A_1 es suficiente para hacer que se revierta la situación planteada cuando se considera solo el efecto de la deriva genética, siendo más probable que se fije este alelo que el alelo A_2.

Ejercicio 4.4

Todas las personas saben que en las remotas Hébridas exteriores surgió un alelo de pigmentación que hace que la lana de las ovejas scottish blackface se vuelva verde bajo la luz crepuscular. Lo que pocas personas saben es otro alelo surgió, que confiere una tonalidad amarillenta. La probabilidad de dejar descendencia no es igual para todos los individuos, debido a la presencia de zorros que detectan y cazan diferencialmente los diferentes tipos de pigmentación. El **fitness** relativo de las ovejas según la presencia de alelos F_V y F_A (causantes de pigmentación verde y amarilla, respectivamente) es

$$\begin{cases} w_{VV} = 1 \\ w_{VA} = 0{,}87 \\ w_{AA} = 0{,}9 \end{cases}$$

Discute qué sucederá con las frecuencias alélicas de la población según se consideren diferentes frecuencias alélicas $f(F_V)$ y $f(F_A)$ de partida.

Solución

Notemos que las ovejas heterocigotas $F_V F_A$ parecen irradiar la luz que más fácilmente detectan los zorros, teniendo el menor **fitness** relativo. Esto implica que la subdominancia es el modo de acción génica, esperándose un régimen de selección disruptiva. Bajo este régimen existe un equilibrio inestable dependiente del efecto heterocigota h, dado por

$$f(\hat{F_V}) = \frac{h-1}{2h-1}$$

Tenemos

$$w_{AA} = 1 - s = 0{,}9 \Rightarrow s = 1 - 0{,}9 = 0{,}1$$

y

$$w_{VA} = 1 - hs = 0{,}87 \Rightarrow h \times 0{,}1 = 0{,}13$$

$$h = \frac{0{,}13}{0{,}1} = 1{,}3$$

Esto es coherente con el modo de acción génica planteado (donde $h > 1$).

La frecuencia de equilibrio se encuentra por lo tanto en

$$f(\hat{F_V}) = \frac{1{,}3 - 1}{2 \times 1{,}3 - 1} = \frac{0{,}3}{1{,}6}$$

$$f(\hat{F_V}) = 0{,}1875 \Rightarrow f(F_A) = 1 - 0{,}1875 = 0{,}8125$$

El valor de $\Delta_s F_V$ dependerá de este punto de equilibrio: si en la población se parte de $f(F_V) > 0{,}1875 \Rightarrow \Delta_s F_V > 0$y, por lo tanto, la frecuencia de este alelo tenderá a aumentar hasta fijarse eventualmente por efecto combinado de la selección natural y la deriva genética. Si en cambio se parte de una frecuencia $f(F_V) < 0{,}1875 \Rightarrow \Delta_s F_V < 0$, por lo que la frecuencia alélica tenderá a disminuir hasta perderse el alelo (fijándose el alelo F_A). Si la población considerada parte de $f(F_V) = 0{,}1875 \Rightarrow \Delta_s F_V = 0$, y las frecuencias alélicas se mantendrán inalteradas.

Ejercicio 4.5

Una población se encuentra segregando los alelos A_1 y A_2 de un locus bialélico. El **fitness** relativo de los organismos homocigotas A_1A_1, A_1A_2 y A_2A_2 está dado por

$$\begin{cases} w_{11} = 1 \\ w_{12} = 1{,}5 \\ w_{22} = 0{,}6 \end{cases}$$

En un momento dado se miden las frecuencias alélicas en la población, estimándose $f(A_1) = 0{,}73$ y $f(A_2) = 0{,}27$.

a. ¿Qué modo de acción génica sugiere el modelo planteado para este locus? Teniendo esto en cuenta, ¿qué esperas que suceda con las frecuencias de los alelos A_1 y A_2?

b. Un tiempo considerable pasa y las condiciones ambientales se modifican, cambiando así el régimen selectivo. Bajo el nuevo régimen selectivo el **fitness** relativo de cada genotipo es

$$\begin{cases} w_{11} = 0{,}55 \\ w_{12} = 0{,}6 \\ w_{22} = 1 \end{cases}$$

¿Qué tipo de régimen selectivo está planteado en esta nueva situación? ¿Qué esperas que suceda ahora con las frecuencias alélicas $f(A_1)$ y $f(A_2)$?

c. Considera la misma situación planteada en **b.**, pero con un coeficiente de selección $s = 0{,}3$. ¿Cómo afecta esto a la situación planteada?

Solución

a. Los organismos con mayor **fitness** relativo son los heterocigotas A_1A_2, lo cual se espera cuando la sobredominancia es el modo de acción génica. El régimen selectivo es de selección estabilizadora, y espera entonces que se alcance una frecuencia de equilibrio $0 < f(\hat{A}_1) < 1$ (ídem para $f(A_2)$). Esta frecuencia de equilibrio es dependiente del efecto de heterocigotas h, siendo

$$f(\hat{A}_1) = \frac{h-1}{2h-1}$$

por lo que es necesario determinar h para estimarla y predecir qué sucederá con las frecuencias alélicas $f(A_1)$ y $f(A_2)$.

Recordemos que $w_{22} = 1 - s$ y que $w_{12} = 1 - hs$. Tenemos entonces

$w_{22} = 0{,}6 = 1 - s \Rightarrow s = 1 - 0{,}6 = 0{,}4$, por lo que

$$w_{12} = 1{,}5 = 1 - 0{,}4 \times h$$

$$0{,}4 \times h = 1 - 1{,}5 = -0{,}5$$

$$h = -\frac{0{,}5}{0{,}4}$$

Esta estimación es coherente con un modo de acción de sobredominancia (donde $h < 0$). La frecuencia de equilibrio es entonces

$$f(\hat{A}_1) = \frac{(-0{,}5/0{,}4) - 1}{2 \times (-0{,}5/0{,}4) - 1}$$

$$f(\hat{A}_1) \approx 0{,}643$$

Por lo tanto, en el largo plazo se espera que la frecuencia alélica $f(A_1)$ disminuya progresivamente, hasta alcanzarse el punto de equilibrio.

b. Notemos que bajo el nuevo régimen selectivo los individuos heterocigotas tienen valores de **fitness** intermedios. Esto corresponde a un modo de acción génica de dominancia incompleta, configurándose un régimen de selección direccional. Siendo el genotipo de mayor **fitness** el homocigota A_2A_2, es de esperar un aumento progresivo de la frecuencia del alelo A_2 hasta alcanzarse su eventual fijación (si consideramos los efectos de la deriva genética aún en poblaciones de tamaño considerable).

c. La fuerza de la selección natural jugará un papel en el tiempo necesario para alcanzarse la fijación del alelo A_2. A igual efecto de heterocigotas h, coeficientes de selección s menores resultarán en un mayor tiempo esperado para la fijación del alelo (lo contrario ocurre para coeficientes mayores).

En la situación planteada en **b.** tenemos

$w_{11} = 0{,}55 = 1 - s \Rightarrow s = 1 - 0{,}55 = 0{,}45$

por lo que si el coeficiente de selección fuera $s = 0{,}3$ se esperaría la misma dinámica, requiriéndose un mayor tiempo generacional para la fijación del alelo A_2.

Capítulo 5

Dinámica de dos loci

> *I think that it is a relatively good approximation to truth — which is much too complicated to allow anything but approximations-that mathematical ideas originate in empirics, although the genealogy is sometimes long and obscure.*[1]
>
> **John von Neumann**

En los capítulos anteriores hemos sentado las bases matemáticas del análisis de la evolución de las frecuencias alélicas en las poblaciones, considerando tanto los fenómenos sistemáticos como los estocásticos. Sin embargo, hasta el momento nuestro modelo fue simplificado en extremo y el tratamiento quedó reducido casi siempre a un locus con un par de alelos. En principio, podríamos pensar que para el tratamiento de más loci alcanzaría con considerar para cada uno de ellos un modelo como el de un locus con dos alelos, como los que hemos trabajado. Eso significaría que nuestros loci son todos independientes unos de otros. Desafortunadamente para el modelado matemático, aunque afortunadamente para nosotros, la meiosis produce un *barajado* de los alelos en diferentes loci, el cual que puede ocurrir en varios puntos del genoma. Obviamente, la probabilidad de que esos eventos discretos que producen el intercambio de cromátidas ocurran entre dos loci en particular es una función de la **distancia de mapeo** entre ellos (cuanto mayor es dicha distancia, mayor la probabilidad de que ese evento ocurra entre los dos loci).

Es momento entonces de comenzar a extender nuestros modelos a la interacción con otro locus cercano y analizar el tipo de dinámicas que pueden surgir de esta interacción. En general, aún los modelos más simplificados de interacción entre dos loci agregan una complejidad importante al análisis, por lo que nos limitaremos a tratar analíticamente los casos más sencillos, mientras que recurriremos a aproximaciones basadas en el modelado computacional en aquellos casos que se complique más allá de nuestras posibilidades. Una nota de advertencia antes de empezar: gran parte del material de este capítulo se encuentra basado en el libro de John H. Gillespie (2004), que es uno de los pocos textos que lo trata como un tema mereciendo un capítulo, la mayoría de ellos apenas considerándolo en alguna sección. Si bien la decisión de qué temas tratar y su profundidad es algo arbitraria, las posibles consecuencias evolutivas de los fenómenos asociados no son despreciables y, por lo tanto, entendimos que merece este tratamiento preferencial.

[1] Works of the Mind, by Robert B. Heywood, editor, vol. 1, no. 1, 180–196, University of Chicago Press, Chicago, 1947. La traducción podría ser algo como «Creo que es una aproximación relativamente buena a la verdad —que es demasiado complicada como para permitir otra cosa que aproximaciones— que las ideas matemáticas se originan en la empiria, aunque la genealogía sea a veces larga y oscura. Sería uno de los orígenes del aforismo atribuido a George Box, de que todos los modelos están mal, pero algunos son útiles. Viene a cuento en este capítulo que recurriremos a modelos bastante simplificados de la realidad, pero aún bastante engorrosos de tratar».

OBJETIVOS DEL CAPÍTULO

☐ Introducir el concepto de desequilibrio de ligamiento entre loci y discutir su relevancia biológica.

☐ Plantear un modelo para la relación entre el desequilibrio de ligamiento y la tasa de recombinación r entre dos loci, así como su evolución en el tiempo.

☐ Presentar medidas para determinar la significancia del desequilibrio de ligamiento entre dos loci, analizando las ventajas y desventajas de estos.

☐ Analizar la relación entre selección y desequilibrio de ligamiento. Se planteará un modelo matemático explícito para un caso sencillo: selección en un locus e impacto en el desequilibrio de ligamiento con un segundo locus neutral.

☐ Discutir diferentes escenarios biológicos, además de la proximidad física, que pueden dar como resultado el desequilibrio de ligamiento entre dos o más loci.

5.1 Desequilibrio de ligamiento y recombinación

En los organismos diploides y con ploidías mayores, durante la meiosis se produce un fenómeno fundamental para la evolución: la **recombinación**. Este proceso se encuentra esquematizado en la figura 5.1. En los organismos diploides todas las células nucleadas, con excepción de los gametos, cuentan con una dotación aproximadamente duplicada, la mitad proveniente de la **hembra** y la otra mitad del **macho**. Como recordarás del curso de **Genética I**, durante la **meiosis I** ocurre ocasionalmente (frecuentemente) la unión física de los cromosomas homólogos en algunos puntos (por ejemplo, en mamíferos usualmente en el orden del número de brazos cromosómicos) llamados **quiasmas** y mediante un proceso complejo se produce un intercambio entre cromátidas no-hermanas. Esto lleva a combinar de una manera diferente los alelos respecto a su arreglo original **hembra-macho**, existiendo ahora fragmentos de uno y de otro en los gametos recombinados.

Este es un proceso aleatorio y los lugares donde ocurren los quiasmas también lo son, aunque no necesariamente con la misma probabilidad para cada punto o región del genoma. La importancia evolutiva de este fenómeno es evidente: la posibilidad de combinar en diferentes formas alelos beneficiosos surgidos en diferentes lugares del mismo cromosoma nos aporta tanto en el efecto aditivo de los mismos como en el efecto agregado de la combinación; este último (estadísticamente el efecto de la **interacción** entre loci), es desde el punto de vista genético la **epistasis**, algo que veremos en el capítulo «El modelo genético básico».

En la figura 5.2 se presenta el modelo simplificado con el que vamos a trabajar en la mayor parte de este capítulo.

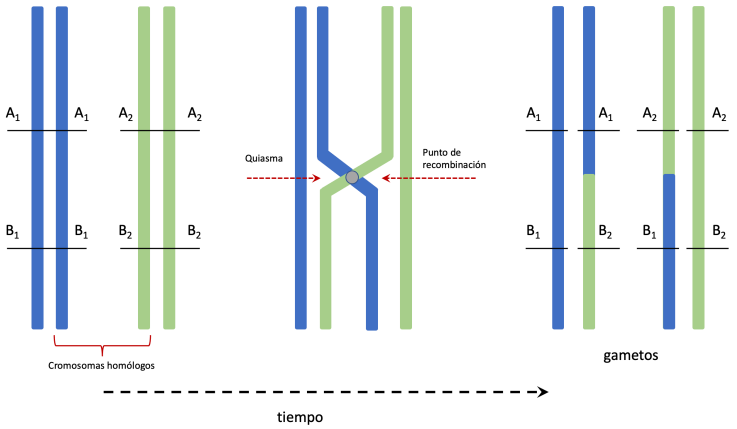


Figura 5.1. Proceso de recombinación durante la meiosis. Durante la meiosis los cromosomas homólogos se aparean y se dan intercambios entre cromátidas no-hermanas, a partir de una unión física de los mismos llamada **quiasma**. En esta representación dos loci, A y B, se encuentran a los lados del punto de recombinación. Los alelos son para cada uno de ellos dos posibles (alelos 1 en azul, alelos 2 en verde). Luego de finalizada la meiosis contamos con 4 gametos haploides, en nuestro caso dos recombinantes y dos idénticos a los parentales.

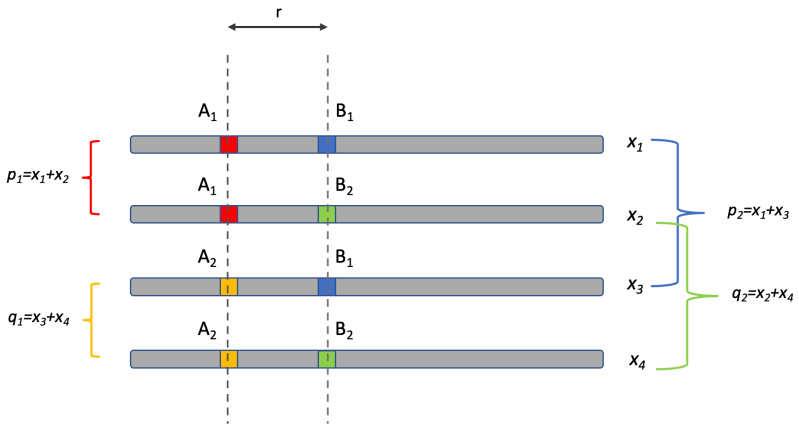


Figura 5.2. Modelo de dos loci con dos alelos cada uno. Las frecuencias de las 4 combinaciones posibles aparecen a la derecha. Los gametos que aparecen en **fase de acoplamiento** son el 1 y el 4, mientras que el 2 y 3 están en **fase de repulsión**. El parámetro r es la probabilidad de producir un gameto recombinante por meiosis. $p_1 = x_1 + x_2$ es la frecuencia del alelo A_1 en el primer locus, mientras que $p_2 = x_1 + x_3$ es la frecuencia del alelo B_1, en el segundo locus.

Tenemos dos loci que asumimos en un principio relativamente cerca en el mismo cromosoma. El primer locus que denotaremos con la letra A tiene dos alelos, A_1 y A_2, con frecuencias p_1 y $q_1 = 1 - p_1$. El segundo locus, que llamaremos B, tiene también dos alelos, B_1 y B_2, con frecuencias p_2 y $q_2 = 1 - p_2$. **Notemos que el subíndice en las frecuencias se refiere al locus,** mientras que usamos las letras p y q para los alelos existentes en cada locus.

A las frecuencias de los $2 \times 2 = 4$ gametos posibles las denominamos x_1 (gameto A_1B_1), x_2 (gameto A_1B_2), x_3 (gameto A_2B_1) y x_4 (gameto A_2B_2), de modo que $x_1+x_2+x_3+x_4 = 1$. A la fracción de gametos recombinantes en la meiosis le llamamos r, con $0 \leqslant r \leqslant \frac{1}{2}$. Si bien en teoría r puede ser mayor a 0,50 (si la recombinación ocurre en los dos pares de cromátidas no-hermanas), se trataría de un fenómeno bastante infrecuente y de ahí la restricción. Más aún, en este capítulo vamos a despreciar la posibilidad de los dobles recombinantes. Si bien es una aproximación, la misma es razonable ya que vamos a trabajar básicamente con loci a relativamente poca distancia. A los gametos que poseen los mismos subíndices en los dos loci los llamamos gametos en **fase de acoplamiento** y a los que poseen distintos subíndices los llamamos en **fase de repulsión**. Tengamos en cuenta que se trata de una distinción arbitraria y simplemente la tomamos para poder describir mejor nuestro modelo, pero en la realidad no existe tal distinción (los alelos no tienen subíndices en la realidad, y a cuál de ellos le asignamos el 1 y a cuál el 2 es completamente arbitrario).

Como vemos en la figura 5.2, la frecuencia de los gametos A_1B_1 está definida por x_1. Si entre los dos loci no existiese ningún tipo de asociación física ni estadística (por ejemplo, **si no estuviesen ligados**) o se encontrasen en **equilibrio de ligamiento**, entonces la probabilidad de obtener un gameto A_1B_1 sería simplemente el producto de tener un alelo A_1 multiplicado por la probabilidad de tener un B_1, es decir $x_1 = p_1p_2$ (recordar que p_1 es la frecuencia de A_1 y p_2 es la frecuencia de B_1). Como ya sabemos, si nuestros loci se encuentran en el mismo cromosoma la probabilidad de que ocurra un entrecruzamiento (***crossover***) entre ellos y, por lo tanto, una recombinación, no es en general despreciable, pero claramente también existe la probabilidad de que ello no ocurra y que por lo tanto vayan a los gametos en la configuración existente en la generación actual.

Veamos ahora cómo podemos expresar la frecuencia del gameto A_1B_1 en la siguiente generación, después de una ronda de apareamiento al azar, en función de la frecuencia en la generación anterior. Existen dos posibilidades mutuamente excluyentes para que en la nueva generación aparezca un gameto A_1B_1: *a)* que se trate de gametos A_1B_1 en la generación previa y que no recombinaron (con probabilidad $(1 - r)x_1$), y *b)* que se trate de un evento de recombinación (con probabilidad r) que arroje A_1B_1 (cuya probabilidad es p_1p_2); como dichos eventos son independientes, su probabilidad conjunta es rp_1p_2. Poniendo en conjunto las consideraciones de los escenarios *a)* y *b)*, tenemos que la probabilidad de A_1B_1 en la generación siguiente (que llamaremos x_1') es

$$x_1' = (1 - r)x_1 + rp_1p_2 \tag{5.1}$$

Claramente, aplicando la misma lógica, las ecuaciones para el resto de las combinaciones de alelos serán

$$x_2' = (1-r)x_2 + rp_1q_2$$
$$x_3' = (1-r)x_3 + rq_1p_2$$
$$x_4' = (1-r)x_4 + rq_1q_2$$

En palabras, la lógica que siguen estas ecuaciones es: a la fracción del mismo gameto que no recombinó le sumamos la proporción de recombinantes del producto de los alelos que forman el gameto.

Veamos ahora qué ocurre con el cambio de frecuencia del gameto A_1B_1. La nueva frecuencia es x_1' y la anterior x_1, por lo que sustituyendo el resultado de (5.1), la diferencia será

$$\Delta_r x_1 = x_1' - x_1 = (1-r)x_1 + rp_1p_2 - x_1 = x_1 - rx_1 + rp_1p_2 - x_1 \Leftrightarrow$$
$$\Delta_r x_1 = -r(x_1 - p_1p_2)$$

Esta formulación nos permite una primera definición del desequilibrio de ligamiento

$$D = x_1 - p_1p_2 \tag{5.2}$$

cuya lógica es sencilla: la diferencia entre la frecuencia real del gameto A_1B_1 y la frecuencia esperada si los dos loci fuesen independientes (que es el producto de la frecuencia de los alelos A_1 y B_1, p_1 y p_2). Sustituyendo en la ecuación (5.1) la definición de D, tenemos ahora

$$\Delta_r x_1 = -rD \tag{5.3}$$

La interpretación de esta ecuación es intuitiva: la frecuencia de los gametos en fase de acoplamiento disminuirá en proporción al producto de la fracción de recombinación y del desequilibrio de ligamiento. Por lo tanto, la diferencia respecto a lo esperado bajo independencia irá disminuyendo en cada generación. Otra forma sencilla de verlo es la siguiente. En el equilibrio, no habrá más cambios en la frecuencia de los gametos de generación a generación, o lo que es lo mismo $\Delta_r x_1 = -rD = 0$. Excepto el caso trivial de $r = 0$ (por ejemplo, en ausencia de recombinación), $\Delta_r x_1 = 0 \Leftrightarrow D = 0$, o lo que es lo mismo

$$D = 0 \Leftrightarrow x_1 - p_1p_2 = 0 \Leftrightarrow x_1 = p_1p_2 \tag{5.4}$$

Es decir, la condición de equilibrio es que la frecuencia del gameto (en este caso A_1B_1, con frecuencia x_1), sea igual al producto de las frecuencias de los alelos que lo componen, en nuestro caso A_1 (frecuencia p_1) y B_1 (frecuencia p_2).

Si trabajamos sobre la ecuación (5.2), podemos ahora definir la frecuencia del gameto A_1B_1 como

$$x_1 = p_1p_2 + D \tag{5.5}$$

Es decir, la frecuencia de cada gameto será lo esperado de las frecuencias de los alelos que lo conforman (en este caso A_1 y B_1), más el aporte del desequilibrio de ligamiento D. Naturalmente este concepto se puede extender a los otros gametos, por lo que las frecuencias serán

$$\begin{aligned} x_1 &= p_1 q_1 + D \\ x_2 &= p_1 q_2 - D \\ x_3 &= q_1 p_2 - D \\ x_4 &= q_1 q_2 + D \end{aligned}$$

El signo negativo para D en los gametos que están en **fase de repulsión** se explica porque el exceso de los gametos que están en **fase de acoplamiento** debe provenir de algún lado.

Para entender mejor la relación entre los gametos en fases de acoplamiento y repulsión con el desequilibrio de ligamiento, veamos si podemos probar la siguiente conjetura

$$D = x_1 x_4 - x_2 x_3 \tag{5.6}$$

En palabras, nuestra conjetura es que el desequilibrio de ligamiento es igual al producto de las frecuencias de los gametos en **fase de acoplamiento** menos el producto de las frecuencias de los gametos en **fase de repulsión.** Para probarlo, podemos igualar la expresión de la ecuación (5.6) con nuestra anterior definición en función de x_1, como aparece en la ecuación (5.2)

$$D = x_1 x_4 - x_2 x_3 \stackrel{?}{=} x_1 - p_1 p_2 = D \tag{5.7}$$

Si recordamos que $p_1 = x_1 + x_2$ y que $p_2 = x_1 + x_3$, reemplazando en la expresión $x_1 - p_1 p_2$, llegamos a que

$$x_1 - p_1 p_2 = x_1 - (x_1 + x_2)(x_1 + x_3) = x_1 - x_1^2 - x_1 x_3 - x_1 x_2 - x_2 x_3 \tag{5.8}$$

Si sacamos factor común x_1 en esta última expresión, tenemos

$$x_1 - x_1^2 - x_1 x_3 - x_1 x_2 - x_2 x_3 = x_1(1 - x_1 - x_2 - x_3) - x_2 x_3 \tag{5.9}$$

Como la suma de frecuencias de los gametos debe ser igual a uno (es decir, $x_1 + x_2 + x_3 + x_4 = 1$), tenemos que $x_4 = 1 - x_1 - x_2 - x_3$. Sustituyendo en la ecuación (5.9) los términos entre paréntesis, tenemos

$$x_1(1 - x_1 - x_2 - x_3) - x_2 x_3 = x_1 x_4 - x_2 x_3 \tag{5.10}$$

Es decir, hemos probado la conjetura de la ecuación (5.7) y, por lo tanto, se confirma la validez de plantear

$$D = x_1x_4 - x_2x_3 \tag{5.11}$$

Veamos si este resultado nos sirve para verificar alguna de las ecuaciones que aparecen en (5.1), la primera de ellas, por ejemplo. Si sustituimos $D = x_1x_4 - x_2x_3$ y utilizamos algunas de las identidades planteadas anteriormente, tenemos

$$x_2 = p_1q_2 - D = p_1q_2 - x_1x_4 + x_2x_3 = (x_1 + x_2)(x_2 + x_4) - x_1x_4 + x_2x_3$$

$$x_2 = x_1x_2 + x_2^2 + x_1x_4 + x_2x_4 - x_1x_4 + x_2x_3 = x_2(x_1 + x_2 + x_3 + x_4) \tag{5.12}$$

Y recordando que $x_1 + x_2 + x_3 + x_4 = 1$, entonces

$$x_2 = x_2 \times (x_1 + x_2 + x_3 + x_4) = x_2 \times (1) = x_2 \tag{5.13}$$

que es lo que queríamos verificar. El procedimiento es el mismo para el resto de las ecuaciones.

Resumiendo, las formas de expresar las frecuencias de los gametos en función de p_i, q_i y D aparecen en el siguiente cuadro que resume las ecuaciones (5.5) y (5.1).

Gametos	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frecuencia (D)	$p_1p_2 + D$	$p_1q_2 - D$	$q_1p_2 - D$	$q_1q_2 + D$

Puesto en conjunto con la ecuación (5.11), los resultados del cuadro anterior nos confirman que cuando D es positivo el desequilibrio de ligamiento produce un exceso de gametos en **fase de acoplamiento** respecto a lo esperado si no existiese ese **ligamiento.** Este exceso debe compensarse con un déficit de gametos en **fase de repulsión.** Cuando D es negativo, se trata de la situación inversa (recordar que la definición de qué pares de gametos consideramos en **fase de acoplamiento** o en **fase de repulsión** es algo completamente arbitrario a nivel molecular).

El otro resultado importante que se desprende del cuadro es el que tiene que ver con los máximos y mínimos para D. Como $x_i \geqslant 0$ para todos los $i \in 1 \times 4$ ya que son las frecuencias de los gametos, entonces

$$\begin{aligned} 0 \leqslant x_1 = p_1p_2 + D \Leftrightarrow D \geqslant -p_1p_2 \\ 0 \leqslant x_2 = p_1q_2 - D \Leftrightarrow D \leqslant p_1q_2 \\ 0 \leqslant x_3 = q_1p_2 - D \Leftrightarrow D \leqslant q_1p_2 \\ 0 \leqslant x_4 = q_1q_2 + D \Leftrightarrow D \geqslant -q_1q_2 \end{aligned} \tag{5.14}$$

Los resultados de (5.14) implican que, agrupándolos por condición,

$$\begin{cases} D_{\text{mín}} = \text{máximo}(-p_1p_2, -q_1q_2) \\ D_{\text{máx}} = \text{mínimo}(p_1q_2, q_1p_2) \end{cases} \tag{5.15}$$

Claramente, $D_{\text{máx}}$ y $D_{\text{mín}}$ dependen ambas de las frecuencias de los alelos, p_1 y p_2, que son específicas a cada situación. Sin embargo, podríamos variar p_1y p_2 a nuestro antojo para ver cuáles son el máximo y el mínimo teóricos de D. La función bivariada (porque depende de p_1 y p_2) para $D_{\text{máx}}$ tiene un máximo cuando $p_1 = \frac{1}{2} = q_1$; $p_2 = \frac{1}{2} = q_2$ y el valor es entonces $D_{\text{máx}} = 0{,}25$, como se aprecia en la figura 5.3. A su vez, para $D_{\text{mín}}$ su valor mínimo teórico es también cuando $p_1 = \frac{1}{2} = q_1$; $p_2 = \frac{1}{2} = q_2$ y equivale a $D_{\text{mín}} = -0{,}25$. En palabras, el valor absoluto del desequilibrio de ligamiento **puede** ser máximo cuando las frecuencias de los dos alelos equivalen a 0,5 en ambos loci.

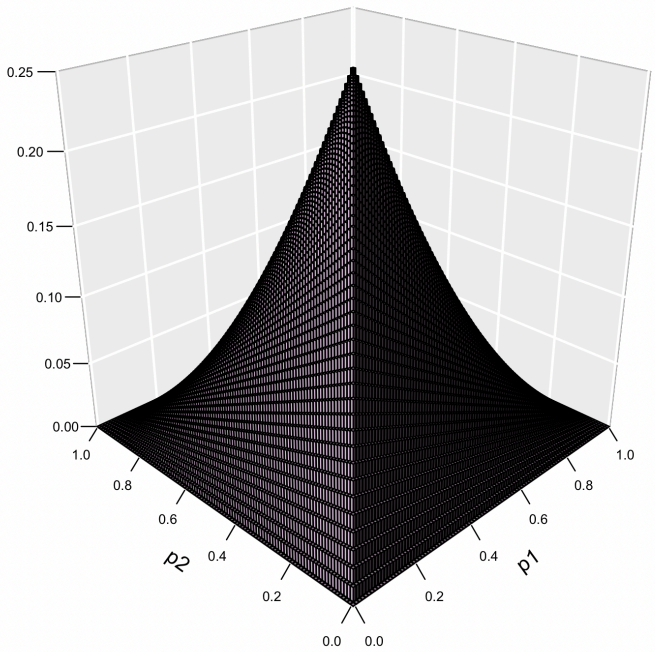


Figura 5.3. Superficie para el valor de $D_{\text{máx}}$ en función de las frecuencias de los alelos A_1 (p_1) y B_1 (p_2). Se aprecia claramente la simetría de la figura, así como el máximo único en $p_1 = q_1 = p_2 = q_2 = \frac{1}{2}$.

Ejemplo 5.1

En un análisis genómico de dos loci (**A** y **B**) físicamente cercanos en el mismo cromosoma autosomal, ambos con dos alelos, de una especie diploide se encontraron los siguientes conteos para los haplotipos posibles

$$\mathbf{A_1B_1} = \mathbf{300}$$
$$\mathbf{A_1B_2} = \mathbf{140}$$
$$\mathbf{A_2B_1} = \mathbf{260}$$
$$\mathbf{A_2B_2} = \mathbf{280}$$

Calcular las frecuencias de los alelos. ¿Es la frecuencia de los haplotipos la esperada con base en estas frecuencias? ¿Hay desequilibrio de ligamiento entre estos loci y si es así cuánto vale? ¿Hay exceso o déficit de los gametos en fase de acoplamiento? ¿Cuánto es el $D_{\text{máx}}$ posible para este par de loci?

Llamemos x_1 a la frecuencia de A_1B_1, x_2 a la de A_1B_2, x_3 a la de A_2B_1 y x_4 a la de A_2B_2, con $x_i = N_i/N$. La frecuencia de A_1 (primer locus), $p_1 = x_1 + x_2$, es decir

$$p_1 = x_1 + x_2 = \frac{N_1}{N} + \frac{N_2}{N} = \frac{300}{980} + \frac{140}{980} \approx 0{,}449 \tag{5.16}$$

Por lo tanto, $q_1 = 1 - p_1 = 1 - 0{,}449 = 0{,}551$. Para el alelo B_1 (segundo locus) será

$$p_2 = x_1 + x_3 = \frac{N_1}{N} + \frac{N_3}{N} = \frac{300}{980} + \frac{260}{980} = 0{,}571 \tag{5.17}$$

y entonces $q_2 = 1 - p_2 = 1 - 0{,}571 = 0{,}429$.

Las frecuencias «reales» de los **4** gametos (haplotipos) posibles son

$$x_1 = \frac{N_1}{N} = \frac{300}{980} = 0{,}306$$
$$x_2 = \frac{N_2}{N} = \frac{140}{980} = 0{,}143$$
$$x_3 = \frac{N_3}{N} = \frac{260}{980} = 0{,}265$$
$$x_4 = \frac{N_4}{N} = \frac{280}{980} = 0{,}286$$

mientras que las esperadas si no existiese desequilibrio de ligamiento serían los productos de las frecuencias de los alelos que los forman, es decir

$$\mathbb{E}_1 = p_1p_2 = 0{,}449 \times 0{,}571 = 0{,}256$$
$$\mathbb{E}_2 = p_1q_2 = 0{,}449 \times 0{,}429 = 0{,}193$$
$$\mathbb{E}_3 = q_1p_2 = 0{,}551 \times 0{,}571 = 0{,}315$$
$$\mathbb{E}_4 = q_1q_2 = 0{,}551 \times 0{,}429 = 0{,}236$$

Un chequeo obvio es que tanto la suma $\sum_i \mathbb{E}_i = 1$ $(0{,}256 + 0{,}193 + 0{,}315 + 0{,}236 = 1)$ así como $\sum_i x_i = 1$ $(0{,}306 + 0{,}143 + 0{,}265 + 0{,}286 = 1)$.

Ahora, podemos armar un cuadro con las frecuencias esperadas y observadas

Gametos	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frecuencia esperada	0,256	0,193	0,315	0,236
Diferencia (D o $-D$)	0,050	−0,050	−0,050	0,050

A partir de este cuadro vemos claramente las diferencias entre las frecuencias esperadas y las observadas, que se corresponden con un valor de $D = 0{,}05$. Los gametos en **fase de acoplamiento**, es decir A_1B_1 y A_2B_2 están en exceso respecto a lo esperado.

De acuerdo con la ecuación (5.15), $D_{\text{máx}} = \text{mínimo}(p_1q_2, q_1p_2)$, por lo que en nuestro caso

$$D_{\text{máx}} = \text{mínimo}(p_1q_2; q_1p_2) = \text{mínimo}(0{,}449 \times 0{,}429; 0{,}551 \times 0{,}571)$$
$$D_{\text{máx}} = \text{mínimo}(0{,}1926; 0{,}3146) = 0{,}1926$$

El conteo de los distintos gametos está influido por el azar y, por lo tanto, lo podemos considerar producto de alguna distribución de probabilidad subyacente. A partir de un conteo de haplotipos (o gametos, por ejemplo) normalmente vamos a derivar no solo las frecuencias de cada uno de ellos, sino que también de los alelos en los distintos loci. En general, al tratarse de un muestreo aleatorio siempre vamos a encontrar valores de D no exactamente iguales a cero, aunque estas diferencias puedan ser mínimas. Eso nos lleva a preguntarnos cómo podemos distinguir cuáles diferencias son estadísticamente significativas, es decir que si muestreamos toda la población (o con $N \to \infty$) detectemos aquellas que «realmente» existen. Para eso debemos antes que nada definir cuál es nuestra hipótesis nula a evaluar o, dicho de otra forma, que esperaríamos nosotros «por defecto».

En el caso de desequilibrio de ligamiento, la hipótesis nula usual es que los loci considerados se encuentran en equilibrio, o lo que es equivalente, que las frecuencias de los alelos en cada locus son independientes de los resultados en otros loci. Por ejemplo, para el caso de dos loci con dos alelos cada uno, usando la notación anterior, la frecuencia esperada para un gameto A_1B_1 sería el producto de la frecuencia de A_1 en el primer locus (p_1) por la frecuencia de B_1 en el segundo locus (p_2) (es decir, bajo la hipótesis nula $\mathbb{E}(x_1) = p_1p_2$). Dado un conteo total de N haplotipos (o gametos), entonces el número esperado de estos que sean A_1B_1 es $\mathbb{E}(n_1) = p_1p_2N$. Claramente, siguiendo esta lógica, el conteo esperado (bajo nuestra hipótesis nula) para los otros gametos será $\mathbb{E}(n_2) = p_1q_2N$, $\mathbb{E}(n_3) = q_1p_2N$ y $\mathbb{E}(n_4) = q_1q_2N$, con $n_1 + n_2 + n_3 + n_4 = N$. Los conteos observados para nuestros haplotipos (gametos) los hemos notado como x_1, x_2, x_3 y x_4 respectivamente, por lo que tenemos un conteo observado y uno esperado. Si recuerdas los cursos de estadística, las anteriores parecen ser las condiciones para la aplicación de un test de independencia. Más aún seguramente recuerdes de esos mismos cursos el test de chi-cuadrado (χ^2), cuya fórmula lucía

$$\chi^2_{gl} = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{5.18}$$

siendo gl los grados de libertad del test, O_i el conteo observado para la «celda» i, y E_i el esperado para esa misma celda. En nuestro caso, podemos arreglar los datos que observamos en una tabla como la siguiente

Gametos	B_1	B_2	**Marginal**
A_2	x_3N	x_4N	$(x_3+x_4)N=q_1N$
Marginal	$(x_1+x_3)N=p_2N$	$(x_2+x_4)N=q_2N$	N

Es de notar que los x_1N,...,x_4N representan los **conteos observados** de los 4 haplotipos posibles (ya que x_1,...,x_4 son las frecuencias relativas de los mismos). Al mismo tiempo, los **conteos esperados** estarán dados por el siguiente cuadro

Gametos	B_1	B_2	**Marginal**
A_2	q_1p_2N	p_2q_2N	q_1N
Marginal	p_2N	q_2N	N

Con esta configuración de las celdas en una tabla podemos volver a formular la ecuación (5.18) para considerar la sumatoria en las i filas y j columnas de la siguiente forma

$$\chi^2_{gl}=\sum_{i,j}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}=\sum_{i,j}\frac{[O_{i,j}-(m_i\times m_j/N)]^2}{(m_i\times m_j/N)} \tag{5.19}$$

donde m_i, $(i=1{,}2)$ y m_j, $(j=1{,}2)$ son los conteos marginales de filas y columnas, respectivamente. En una tabla de 2x2 como la que estamos trabajando, los grados de libertad son $gl=2\times2=4-1=3$, menos 1 para estimar p_1 $(3-1=2)$, menos otro grado de libertad para estimar p_2 $(2-1=1)$, por lo que $gl=1$, con lo que ya tenemos casi todo para nuestro **test de independencia**.[2] Lo único que nos estaría faltando es definir el valor umbral para p que vamos a usar para determinar si la diferencia observada es estadísticamente significativa (a veces conocido como α).

Ejemplo 5.2

Con los datos del ejemplo 5.1 determina si el desequilibrio de ligamiento observado es estadísticamente significativo al valor de $\alpha=1\,\%$.

Ordenando los haplotipos observados en una tabla tenemos

Gametos	B_1	B_2	**Marginal**
A_2	260	280	540
Marginal	560	420	980

[2] En un test de independencia los grados de libertad se calculan como (filas-1) × (columnas-1) ya que, dado el gran total, hacen falta los valores de (filas-1) para saber la que falta y lo mismo (columnas-1) para saber la columna que falta.

El valor esperado para la celda i,j podemos calcularlos de acuerdo con $E_{i,j} = \frac{m_1 \times m_j}{N}$. Por ejemplo, para el haplotipo A_1B_1 sería $E_{1,1} = \frac{440 \times 560}{980} = 251{,}43$ Por lo tanto, si lo calculamos para las celdas restantes, los esperados serían

Gametos	B_1	B_2	**Marginal**
A_2	308,57	231,43	540
Marginal	560	420	980

Ahora, aplicando la ecuación (5.19),

$$\chi^2_{gl} = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{5.20}$$

por lo que tenemos que calcular primero los valores de $\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$ y luego sumarlos. Para $i = 1,j = 1$ tenemos $\frac{(O_{1,1}-E_{1,1})^2}{E_{1,1}} = \frac{(300-251{,}43)^2}{251{,}43} = 9{,}38$. Para $i = 1,j = 2$ tenemos $\frac{(140-188{,}57)^2}{188{,}57} = 12{,}51$. Para $i = 2,j = 1$ tenemos $\frac{(260-308{,}57)^2}{308{,}57} = 7{,}65$. Finalmente, para $i = 2,j = 2$ tenemos $\frac{(280-231{,}43)^2}{231{,}43} = 10{,}19$. Sumando estos valores, tenemos el valor del estadístico

$$\chi^2 = 9{,}38 + 12{,}51 + 7{,}65 + 10{,}19 = 39{,}73 \tag{5.21}$$

Si ahora entramos en una tabla de valores del estadístico χ^2 con $gl = 1$, tenemos que el valor del estadístico correspondiente a $\alpha = 0{,}01$ es 6,635. Como nuestro resultado es claramente superior a este valor, es decir está a la derecha de este punto de corte, la diferencia de frecuencias entre lo observado y lo esperado (bajo la hipótesis de independencia de los loci) es estadísticamente significativa y podemos rechazar la hipótesis nula.

PARA RECORDAR

- Si x_1, x_2, x_3 y x_4 son las frecuencias de los gametos, con x_1 y x_4 la de los gametos en **fase de acoplamiento**, entonces el **desequilibrio de ligamiento** es

$$D = x_1x_4 - x_2x_3 \tag{5.22}$$

- Si $p_1 = x_1 + x_2$ y $p_2 = x_1 + x_3$ son las frecuencias de los alelos A_1 y B_1 respectivamente, entonces

$$x_1 = p_1p_2 + D$$
$$x_2 = p_1q_2 - D$$
$$x_3 = q_1p_2 - D$$
$$x_4 = q_1q_2 + D$$

- El rango de D posibles para cada par de loci depende de las frecuencias de los alelos en los mismos y en particular $D_{\text{mín}}$ y $D_{\text{máx}}$ quedan definidos por

$$D_{\text{mín}} = \text{máximo}(-p_1p_2, -q_1q_2)$$
$$D_{\text{máx}} = \text{mínimo}(p_1q_2, q_1p_2)$$

5.2 La evolución en el tiempo del desequilibrio de ligamiento

A esta altura podemos preguntarnos cómo evolucionará el desequilibrio de ligamiento con el tiempo, en nuestro caso en generaciones. Para eso podemos ver lo que ocurre en una generación. Si llamamos con x_1' a la frecuencia del gameto A_1B_1 en la siguiente generación, entonces de acuerdo con la ecuación (5.5) (poniendo «primas» donde corresponda)

$$x_1' = p_1'p_2' + D' = (1-r)(p_1p_2 + D) + rp_1p_2 \tag{5.23}$$

trabajando un poco la ecuación, para despejar D' (el desequilibrio de ligamiento en la siguiente generación) llegamos a

$$\begin{aligned} D' &= (1-r)(p_1p_2 + D) + rp_1p_2 - p_1'p_2' \Leftrightarrow \\ D' &= p_1p_2 - rp_1p_2 + D - rD + rp_1p_2 - p_1'p_2' \end{aligned}$$

Si asumimos que la frecuencia de los alelos no depende de sus combinaciones y que el equilibrio de Hardy-Weinberg no se ve afectado por la inclusión de otros loci, entonces podemos asumir $p_1' = p_1$ y $p_2' = p_2$, por lo que podemos simplificar la expresión anterior

$$\begin{aligned} D' = p_1p_2 + D - rD - p_1p_2 \;\; \therefore D' &= D - rD \\ D' &= D(1-r) \end{aligned}$$

Es decir, el desequilibrio de ligamiento en la siguiente generación será una fracción $(1-r)$ del desequilibrio de ligamiento en la anterior. El cambio en el desequilibrio de ligamiento será entonces la diferencia entre D' y D, es decir

$$\begin{aligned} \Delta_r D = D' - D = D(1-r) - D \;\; \therefore \\ \Delta_r D = -rD \end{aligned}$$

Claramente, si en cada generación el desequilibrio de ligamiento es el producto de $(1-r)$ por el desequilibrio en la anterior (que a su vez es la fracción $(1-r)$ de dos generaciones atrás), entonces luego de t generaciones será el producto de $(1-r)$, t veces por sí mismo. Es decir, será equivalente a $(1-r)^t$ multiplicado por el desequilibrio en la primera generación considerada, que llamamos D_0. Por lo tanto, el desequilibrio de ligamiento en la generación t será

$$D_t = (1-r)^t D_0 \tag{5.24}$$

Este patrón de decrecimiento es geométrico, y se aproxima razonablemente bien al patrón de decrecimiento exponencial, ya que $(1-r)^t \approx e^{-rt}$ (esta aproximación ya la hemos visto en capítulos previos). Por lo tanto, tenemos

$$D_t \approx e^{-rt} D_0 \tag{5.25}$$

En la figura 5.4 se puede apreciar la evolución del desequilibrio de ligamiento a medida que transcurren sucesivas generaciones, para distintos valores de r. Como $D_t = (1-r)^t D_0$, para independizar el comportamiento de las curvas del valor D_0 expresamos está gráfica considerando $D_0 = D_{\text{máx}} = 0{,}25$. En esta figura se aprecia que la fracción del desequilibrio de ligamiento que permanece con el paso de las generaciones decrece en general rápidamente, dependiendo de r: cuanto menor es la fracción de recombinantes entre los dos loci, más tarda en reducirse el desequilibrio. Esto es totalmente intuitivo, ya que cuanto menos probable sea la aparición de un evento de recombinación, menos probable es que los gametos recombinen y, por lo tanto, que se rompa el desequilibrio de ligamiento. El corte con la línea gris horizontal representa el número de generaciones necesario para reducir el desequilibrio a la mitad.

Veamos si podemos determinar analíticamente el número de generaciones necesarias para llegar a una proporción determinada de D_0 (por ejemplo, f). Tendremos $D_t = D_0 f$, con $0 < f < 1$. Si operamos con la ecuación (5.24), sustituyendo D_t, tenemos

$$D_t = D_0 f = (1-r)^t D_0 \;\therefore\; f = (1-r)^t \tag{5.26}$$

Aplicando logaritmo a ambos lados la ecuación (5.26) se transforma en

$$\ln(f) = \ln(1-r)^t = t\ln(1-r) \Leftrightarrow \tag{5.27}$$

$$t = \frac{\ln(f)}{\ln(1-r)} \tag{5.28}$$

El resultado de aplicar la ecuación (5.29) a distintas combinaciones de la fracción $f = \frac{D_t}{D_0}$ y r se puede apreciar en la tabla 5.1. Claramente, a medida que disminuye la fracción de gametos recombinantes r (de izquierda a derecha en la tabla), mayor se hace el número de generaciones necesarias para alcanzar una fracción determinada f. Por ejemplo, mientras que para reducir el desequilibrio de ligamiento inicial a la mitad con un $r = 0{,}50$ se precisa una generación, para lograrlo con un $r = 0{,}01$ se precisan ≈ 69 generaciones.

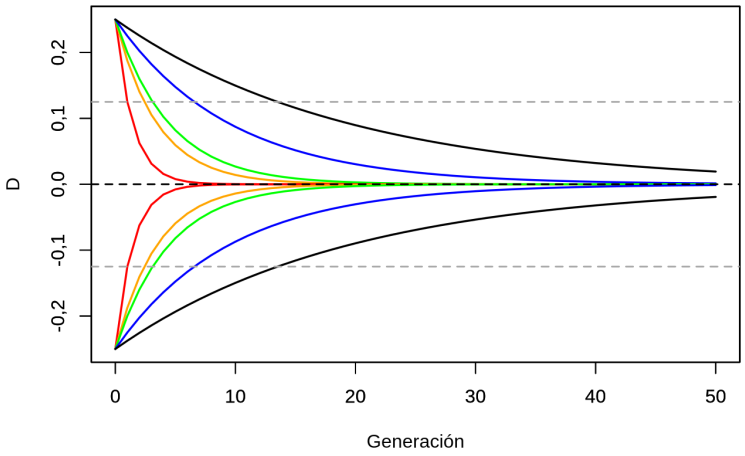


Figura 5.4. Evolución del desequilibrio de ligamiento respecto de $D_0 = D_{máx} = 0{,}25$. Rojo $r = 0{,}50$, anaranjado $r = 0{,}25$, verde $r = 0{,}20$, azul $r = 0{,}10$ y negro $r = 0{,}05$

Tabla 5.1. Número de generaciones requeridas para alcanzar una fracción del desequilibrio de ligamiento inicial (f), de acuerdo con la fracción recombinante (r)

	r=0,50	r=0,40	r=0,30	r=0,25	r=0,20	r=0,10	r=0,05	r=0,01
f=0,75	0,42	0,56	0,81	1,00	1,29	2,73	5,61	28,62
f=0,50	1,00	1,36	1,94	2,41	3,11	6,58	13,51	68,97
f=0,40	1,32	1,79	2,57	3,19	4,11	8,70	17,86	91,17
f=0,30	1,74	2,36	3,38	4,19	5,40	11,43	23,47	119,79
f=0,25	2,00	2,71	3,89	4,82	6,21	13,16	27,03	137,94
f=0,20	2,32	3,15	4,51	5,59	7,21	15,28	31,38	160,14
f=0,10	3,32	4,51	6,46	8,00	10,32	21,85	44,89	229,11
f=0,05	4,32	5,86	8,40	10,41	13,43	28,43	58,40	298,07

El comportamiento general de la función se puede apreciar en la figura 5.5. Claramente el comportamiento es aproximadamente exponencial en los dos ejes, lo cual hace extremadamente lenta la reducción final del desequilibrio de ligamiento si la tasa de recombinación es muy baja, algo que suele ocurrir entre loci muy cercanos físicamente.

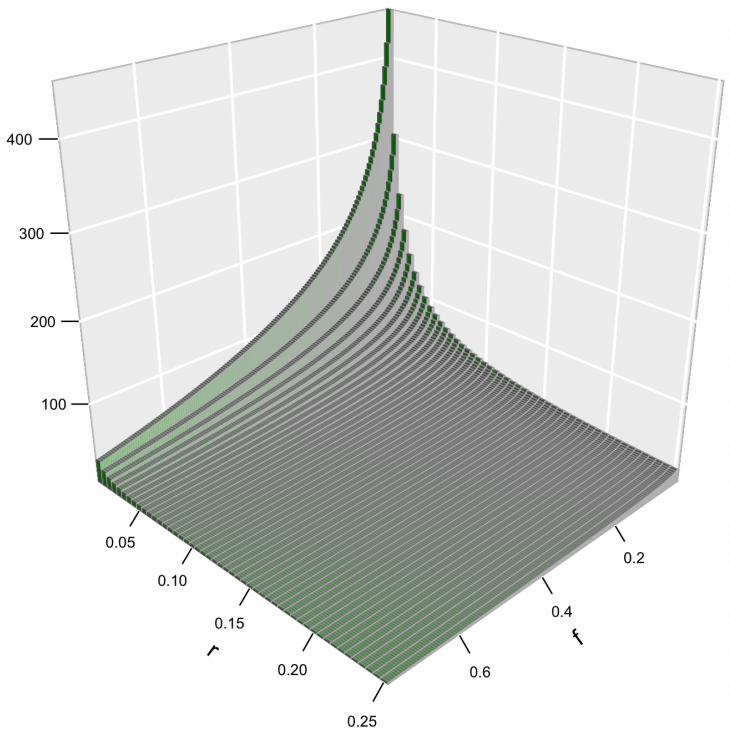


Figura 5.5. Superficie de la función bivariada que nos devuelve el número de generaciones requeridas para alcanzar una fracción $f = \frac{D_t}{D_0}$ del desequilibrio de ligamiento inicial en función de la fracción de recombinación r. Los límites fueron elegidos ($f \leqslant 0{,}75$, $r \leqslant 0{,}25$) para poder observar mejor el comportamiento general de la función (fuera de estos límites la superficie es mucho más plana).

Por último, la aproximación exponencial permite estimar el número t de generaciones necesarias para alcanzar una fracción f del desequilibrio de ligamiento inicial D_0

$$D_t = D_0 f \approx e^{-rt} D_0 \Leftrightarrow$$
$$ln(f) = -rt \Leftrightarrow$$

$$t = \frac{\ln(f)}{-r} \tag{5.29}$$

PARA RECORDAR

- Si partimos de un desequilibrio de ligamiento D_0 en la generación 0, el desequilibrio en la generación t será

$$D_t = (1-r)^t D_0 \tag{5.30}$$

- Esto muestra un comportamiento de función geométrica, que se aproxima razonablemente bien por una exponencial ya que $(1-r)^t \approx e^{-rt}$, por lo que

$$D_t \approx e^{-rt} D_0 \tag{5.31}$$

- Si $f = \frac{D_t}{D_0}$ es la fracción del desequilibrio de ligamiento inicial que permanece después de un tiempo, el número de generaciones requerido para alcanzarlo es

$$t = \frac{\ln f}{\ln(1-r)} \tag{5.32}$$

- Nuevamente, una aproximación del número de generaciones requerido, a partir de la aproximación exponencial es

$$t = \frac{\ln f}{-r} \tag{5.33}$$

5.3 Otras medidas de asociación

Como vimos previamente, en la ecuación (5.15) el rango de valores posibles de D es claramente dependiente de las frecuencias de los alelos en ambos loci, ya que este rango se encuentra dado por

$$\begin{cases} D_{\text{mín}} = \text{máximo}(-p_1p_2, -q_1q_2) \\ D_{\text{máx}} = \text{mínimo}(p_1q_2, q_1p_2) \end{cases} \tag{5.34}$$

Una alternativa para hacer D independiente de las frecuencias es dividir su valor por su máximo (o mínimo) valor posible en relación con las frecuencias (dependiendo del signo se utiliza $D_{\text{máx}}$ o $D_{\text{mín}}$). Es decir,

$$D' = \frac{D}{D_{\text{máx}}} \text{ si } D > 0$$
$$D' = \frac{D}{D_{\text{mín}}} \text{ si } D < 0$$

De esta manera, el estadístico D' nos permite comparar el desequilibrio de ligamiento de una manera independientemente de la historia de evolución de las frecuencias en distintas regiones del genoma o aún entre organismos diferentes. El rango de valores para D' es entonces $0 \leqslant D' \leqslant 1$, ya que tanto en el caso de que normalicemos según $D = D_{\text{máx}}$ o $D = D_{\text{mín}}$ tendremos $D' = 1$.

Otro estadístico usual es el r^2, definido como

$$r^2 = \frac{D^2}{p_1q_1p_2q_2} \tag{5.35}$$

Es importante tener en cuenta que en este contexto r **no se refiere a la tasa de recombinación**, la cual también fue denotada con r en secciones previas. r^2 es el cuadrado del coeficiente de correlación de Pearson, a veces llamado el coeficiente de determinación, que recordarás de los cursos de estadística. En nuestro contexto $r = \sqrt{r^2}$ es la correlación entre estados alélicos dentro del mismo gameto. Por otro lado, r^2 es útil para el cálculo de χ^2 a partir de las frecuencias de haplotipos ya que el valor numérico que usamos en el test es igual a r^2N. Igual que para D', $0 \leqslant r^2 \leqslant 1$, ya que se trata del cuadrado del coeficiente de correlación (que por definición está limitado a $-1 \leqslant r \leqslant 1$). Tanto D' como r^2 tienen signo positivo por definición, por lo que desafortunadamente no es posible tener el sentido del desequilibrio a partir de ellos (pero si a partir de D, del que ambos parten usualmente).

Ejemplo 5.3

A partir del resultado $\chi^2 = r^2N$ calcular el valor de este estadístico con los datos del ejemplo 5.1.

Usando las frecuencias calculadas para p_1,q_1, p_2, q_2 y el valor de D del ejemplo, sustituyendo, tenemos

$$\chi^2 = 0{,}05^2/(0{,}551 \times 0{,}449 \times 0{,}571 \times 0{,}429) \times (300 + 140 + 260 + 280) \approx 40{,}43 \quad (5.36)$$

que es un poco superior al resultado obtenido en el ejemplo 5.2. Si en lugar de eso partimos directamente del hecho de que $D = x_1x_4 - x_2x_3$, usando los números

$$\chi^2 = 980 \times \frac{\left[\frac{(300\times280)}{980^2} - \frac{(140\times260)}{980^2}\right]^2}{\frac{440}{980} \times \left(1 - \frac{440}{980}\right) \times \frac{560}{980} \times \left(1 - \frac{560}{980}\right)} = 39{,}73 \quad (5.37)$$

que es un valor idéntico al calculado usando el formato de tabla de 2×2, pero ahora por un procedimiento más abreviado.

La relación entre los dos estadísticos, D' y r^2 se puede observar en la figura 5.6.

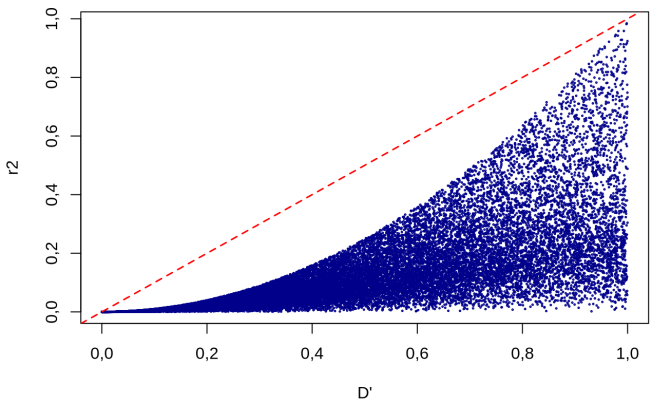


Figura 5.6. Relación entre los estadísticos D' y r^2 para realizaciones al azar de los 4 haplotipos en dos loci con dos alelos. Mientras que para valores bajos de r^2 es posible encontrar cualquier valor de D', no ocurre lo mismo en el sentido contrario (valores bajos de D' implican valores más bajos de r^2). Realización propia sobre idea en Hartl y Clark (2007).

La figura 5.6 representa 30.000 realizaciones al azar de las 4 frecuencias de los gametos, muestreados de una distribución uniforme y escalados (para que sumen a uno) en cada realización del experimento. A partir de esos muestreos se calcularon los dos estadísticos. Mientras que para valores bajos de r^2 se pueden encontrar todos los valores posibles de D', lo contrario no se cumple. Es más, los valores de r^2 son siempre menores que los de D', lo

que se aprecia claramente ya que los puntos están por debajo de la línea de identidad (diagonal en rojo). De hecho, r^2 puede tomar cualquier valor entre $0 \leqslant r^2 \leqslant (D')^2$ (recordemos que el cuadrado de un número entre cero y uno es menor que el número mismo).

Una diferencia importante entre el estadístico D' y el r^2 tiene que ver con la afectación de ambos frente a la historia que produce el desequilibrio de ligamiento. Mientras que r^2 es relativamente sensible al momento en que, por ejemplo, aparece una nueva variante, D' es relativamente independiente y su principal dependencia es la tasa de recombinación. Por ejemplo, un alelo que entra «temprano» en la historia evolutiva de la asociación tendrá, a medida que pase el tiempo, una alta correlación con el alelo en el otro locus con el que fue a dar en el inicio. En cambio, si la introducción de este alelo es «reciente» la correlación será baja. Eso explica en gran medida por qué para un valor determinado de D' el rango de valores posibles de r^2 es importante. Otras diferencias que favorecen el uso de r^2 sobre D' son que primero presenta mayor estabilidad en el caso de loci con alelos en baja frecuencia cuando los tamaños muestrales son pequeños, y que favorece la interpretación como la proporción explicada de varianza debida al desequilibrio de ligamiento en relación con la varianza explicada por **QTL** (*Quantitative Trait Loci*) (Van Inghelandt et al. 2011). Finalmente, existen otros estadísticos alternativos con mejores propiedades en ciertas condiciones o más fáciles de obtener. Por ejemplo, D. Gianola, Qanbari y Simianer (2013) introducen el estadístico ρ^2 que además de buenas propiedades de muestreo es mucho más sencillo de calcular que los estimados por máxima-verosimilitud o bayesianos de la correlación tetracórica (correlación para tablas de 2 × 2, por ejemplo).

PARA RECORDAR

- Dos medidas alternativas para el desequilibrio de ligamiento, menos sensibles a las frecuencias de los alelos son D' y r^2.
- El estadístico D' es el valor de D estandarizado al valor máximo (mínimo) posible, dependiendo del signo de D, es decir $D' = \frac{D}{D_{\text{máx}}}$ si $D > 0$; $D' = \frac{D}{D_{\text{mín}}}$ si $D < 0$.
- El estadístico r^2 es la correlación entre estados alélicos dentro del mismo gameto y se calcula como $r^2 = \frac{D^2}{(p_1 q_1 p_2 q_2)}$.
- Para cada valor de D' r^2 puede variar entre $0 \leqslant r^2 \leqslant (D')^2$, lo que refleja aspectos diferentes de ambos estadísticos: mientras D' es principalmente afectado por la tasa de recombinación, r^2 refleja también de alguna manera la historia de los alelos.
- El valor de r^2 es útil para calcular la significación estadística del desequilibrio de ligamiento, ya que $\chi^2 = r^2 N$ (con N = número de gametos).

5.4 Selección en modelos de dos loci

En esta sección vamos a estudiar qué ocurre con la selección cuando tenemos dos loci en los que la tasa de recombinación entre ellos no es $r = \frac{1}{2}$ (es decir, cuando no podemos considerar a ambos loci como estadísticamente independientes). Nuestro estudio va a ser somero, ya que las complicaciones posibles son inmediatas y muchas, lo que escapa al alcance del presente libro. Para el modelado de este fenómeno vamos a suponer que entre los dos loci

la tasa de recombinación $0 \leqslant r < \frac{1}{2}$ (aunque teóricamente r puede ser mayor a $\frac{1}{2}$). La idea central de esta sección aparece representada en la figura 5.7.

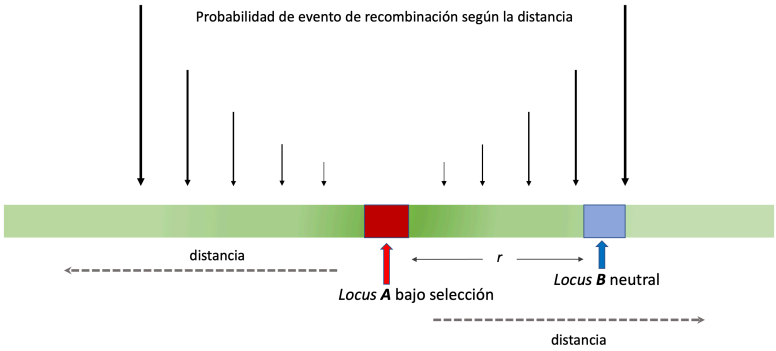


Figura 5.7. Ilustración del modelo de selección que vamos a analizar. Mientras que vamos a tratar con la selección en un solo locus (en rojo) de acuerdo con el modelo que vimos en el capítulo «Selección natural», observaremos qué ocurre en algún otro locus que se encuentre en desequilibrio de ligamiento con el primero. A medida que crece la distancia respecto al locus seleccionado, aumenta la probabilidad de recombinación y, por lo tanto, la pérdida del efecto asociado a la selección en el primer locus.

Tenemos un locus bajo selección, el locus **A**, y un locus «cercano» **B** «neutral» (o sea, que no está sujeto a selección). En el presente contexto el término *cercano* refiere a que $r < \frac{1}{2}$, es decir que ambos loci no se comportan desde el punto de vista de la recombinación como totalmente independientes. A medida que nos alejamos del locus bajo selección la probabilidad de que aparezca un quiasma entre el locus **A** y el **B** se incrementa (lo que representamos por flecha de tamaño creciente). En forma ingenua podríamos pensar que la dinámica del locus **A** sería la de selección de acuerdo con lo que ya vimos mientras que el comportamiento de **B** (no sujeto a la selección) sería completamente al azar y dominado exclusivamente por fenómenos como la mutación y la deriva. Sin embargo, en esta sección vamos a ignorar los efectos de estas dos últimas fuerzas evolutivas y vamos a estudiar si en ausencia de esas fuerzas las frecuencias de los alelos en **B** siguen imperturbables.

Tenemos dos loci con dos alelos cada uno, A_1 y A_2 en el locus **A**, B_1 y B_2 en el **B**. Imaginemos ahora que en lugar de asignar **fitness** a los gametos (como hicimos en el capítulo «Selección Natural» vamos a asignarle ese **fitness** a los genotipos (es decir, a la combinación específica de gametos). Para ello vamos a hacer una matriz con las frecuencias de los distintos genotipos luego del apareamiento aleatorio y la selección por viabilidad. El cuadro siguiente resume estos resultados. La primera columna y el encabezado representan los gametos, mientras que cada celda de la matriz representa la frecuencia esperada para dicho genotipo (el constituido por el gameto de fila con el de columna).

Gametos	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_2	–	$x_2^2w_{22}$	$2x_2x_3w_{23}$	$2x_2x_4w_{24}$
A_2B_1	–	–	$x_3^2w_{33}$	$2x_3x_4w_{34}$
A_2B_2	–	–	–	$x_4^2w_{44}$

El número de parámetros de **fitness** es ahora de $n(n+1)/2 = 4 \times 5/2 = 10$. La lógica de la notación es sencilla. Las frecuencias de los gametos (haplotipos) son x_1, x_2, x_3 y x_4, con los subíndices haciendo referencia a los gametos A_1B_1, A_1B_2, A_2B_1 y A_2B_2. Entonces w_{ij} es el **fitness** asociado a la combinación del gameto i con el gameto j. En la matriz (cuadro) de arriba tenemos las posibles combinaciones de gametos y los **fitness** asociados a cada combinación (fila i, columna j). La matriz sería simétrica (por eso la ausencia de valores en el triángulo inferior izquierdo), los valores ausentes están contemplados en su simétrico (de ahí que exista un factor 2 cuando $i \neq j$).

Si recordamos, por ejemplo, x_1 es la frecuencia del gameto A_1B_1. Veamos cuántos de esos aparecen en la siguiente generación (x_1'), luego de un apareamiento al azar. Analicemos la primera fila del cuadro anterior que corresponde a este gameto. Por ejemplo, la primera fila con la primera columna representa un genotipo A_1B_1/A_1B_1 que en la meiosis producirá gametos. Más allá de que haya o no recombinación todos los gametos serán A_1B_1 en este caso, por lo que la frecuencia de este genotipo en la edad adulta y produciendo gametos será $x_1^2 w_{11}$. El caso de la primera fila con la segunda columna es aún más ilustrativo. El genotipo A_1B_1/A_1B_2, más allá de la recombinación o ausencia de ella, producirá la mitad de los gametos A_1B_1 y la otra mitad A_1B_2. La frecuencia de este genotipo que llega a la reproducción es $2x_1x_2w_{12}$, por lo que la mitad de nuestro interés (la que genera A_1B_1) será $\frac{1}{2}2x_1x_2w_{12} = x_1x_2w_{12}$. Para la tercera fila el genotipo a considerar es A_1B_1/A_2B_1, que producirá la mitad de los gametos A_1B_1 y la mitad A_2B_1, por lo que su aporte a x_1' será $x_1x_3w_{13}$ (por razonamiento análogo al planteado en el caso anterior). Finalmente, de esta fila solo nos queda el genotipo A_1B_1/A_2B_2, que solo producirá gametos A_1B_1 en el caso de que no haya recombinación, que serán en ese caso la mitad. Como la probabilidad de no-recombinación es $(1-r)$, entonces todo junto es $\frac{1}{2}(1-r)2x_1x_4w_{14} = (1-r)x_1x_4w_{14}$. De todos los otros genotipos en el resto de las filas, el único que puede producir gametos A_1B_1 es el A_1B_2/A_2B_1 (fila $i = 2$, columna $j = 3$), pero solamente si ocurre recombinación, lo que acontece con probabilidad r, en cuyo caso la mitad de los gametos serán del tipo A_1B_1. Poniendo todo junto para este caso, obtenemos $\frac{1}{2}r(2x_2x_3w_{23}) = rx_2x_3w_{23}$.

Ahora, combinando los resultados de todas las posibilidades de obtener gametos A_1B_1 en la siguiente generación y recordando de dividir por el **fitness medio** ($\bar{w}$), tenemos que la frecuencia de este gameto en la siguiente generación será

$$x_1' = \frac{x_1^2 w_{11} + x_1x_2w_{12} + x_1x_3w_{13} + (1-r)x_1x_4w_{14} + rx_2x_3w_{23}}{\bar{w}} \tag{5.38}$$

Antes de continuar veamos el concepto de **fitness marginal** de un gameto. El mismo es el fitness medio de todos los genotipos en los que participa ese gameto, justo antes de que la selección ocurra, es decir

$$\bar{w}_i = \frac{\sum_{j=1}^{4} x_i x_j w_{ij}}{x_i} = \sum_{j=1}^{4} x_j w_{ij} \tag{5.39}$$

y, por lo tanto, el **fitness medio** de todos los gametos, el promedio de los **fitness marginales**, es

$$\bar{w} = \sum_{i=1}^{i=4} x_i \bar{w}_i \tag{5.40}$$

Prosigamos con el caso de nuestro gameto A_1B_1. En ese caso, la ecuación (5.39) se transforma en

$$\bar{w}_1 = x_1w_{11} + x_2w_{12} + x_3w_{13} + x_4w_{14} \tag{5.41}$$

Ahora asumamos además que $w_{14} = w_{23}$, lo cual implica que no hay un efecto diferencial *cis-trans*, es decir que el **fitness** depende de la combinación de alelos, pero no de cómo ellos están organizados en los cromosomas homólogos.[3] Si hacemos explícita la multiplicación en el numerador de (5.38) tenemos, sustituyendo además w_{23} por w_{14}

$$x_1' = \frac{x_1^2w_{11} + x_1x_2w_{12} + x_1x_3w_{13} + x_1x_4w_{14} - rx_1x_4w_{14} + rx_2x_3w_{23}}{\bar{w}} \Leftrightarrow$$

$$x_1' = \frac{x_1^2w_{11} + x_1x_2w_{12} + x_1x_3w_{13} + x_1x_4w_{14} - r(x_1x_4 - x_2x_3)w_{14}}{\bar{w}} \Leftrightarrow$$

$$x_1' = \frac{x_1(x_1w_{11} + x_2w_{12} + x_3w_{13} + x_4w_{14}) - rDw_{14}}{\bar{w}} \Leftrightarrow$$

$$x_1' = \frac{x_1\bar{w}_1 - rDw_{14}}{\bar{w}} \tag{5.42}$$

Veamos ahora el cambio en las frecuencias de los gametos, es decir Δx_i, que lo obtenemos como de costumbre, $\Delta x_i = x_i' - x$. Por ejemplo, para el gameto A_1B_1, el cambio de frecuencia será

$$\Delta x_1 = x_1' - x_1 = \frac{x_1\bar{w}_1 - rDw_{14}}{\bar{w}} - x_1 \therefore= \frac{x_1\bar{w}_1 - rDw_{14}}{\bar{w}} - x_1\frac{\bar{w}}{\bar{w}}$$

$$\Delta x_1 = \frac{x_1\bar{w}_1 - rDw_{14} - x_1\bar{w}}{\bar{w}} \therefore$$

$$\Delta x_1 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D}{\bar{w}} \tag{5.43}$$

En principio, podemos aplicar la misma lógica para el resto de las ecuaciones, lo que nos lleva a

$$\begin{cases} \Delta x_1 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D}{\bar{w}} \\ \Delta x_2 = \frac{x_2(\bar{w}_2 - \bar{w}) + rw_{14}D}{\bar{w}} \\ \Delta x_3 = \frac{x_3(\bar{w}_3 - \bar{w}) + rw_{14}D}{\bar{w}} \\ \Delta x_4 = \frac{x_4(\bar{w}_4 - \bar{w}) - rw_{14}D}{\bar{w}} \end{cases} \tag{5.44}$$

Finalmente, podemos interpretar los resultados de las ecuaciones (5.43) y (5.45) mediante una aproximación (recordemos la ecuación (5.3))

$$\Delta x_i \approx \Delta_s x_i + \Delta_r x_i \tag{5.45}$$

[3] Esta aproximación es bastante razonable en general, pero no siempre. Un contraejemplo es el caso de reguladores de expresión en *cis*.

Es decir, el cambio en la frecuencia de los gametos es aproximadamente la suma del cambio debido a la selección y el cambio debido a la recombinación. Más aún, el conflicto entre las direcciones de ambos aportes es evidente (notemos los signos de cada término en la ecuación): la selección tiende a favorecer la asociación de los alelos en gametos, mientras que la recombinación actúa destruyendo esa asociación.

Selección en un locus: impacto en loci en desequilibrio de ligamiento

Un ejemplo de la aplicación del análisis previo es el de un locus bajo selección en desequilibrio de ligamiento con un locus cercano, neutral este último (es decir, sin ventaja selectiva ninguno de sus alelos o combinaciones) aparece en el libro de John H. Gillespie (2004). ¿Qué ocurrirá con las frecuencias alélicas bajo estas condiciones? Utilizaremos la misma notación del capítulo «Selección natural». Para el primer locus, los **fitness relativos** quedan definidos por 1, $1-hs$ y $1-s$, y entonces nuestro cuadro de **fitness** w_{ij} estará dado por

Gametos	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_2	–	1	$1-hs$	$1-hs$
A_2B_1	–	–	$1-s$	$1-s$
A_2B_2	–	–	–	$1-s$

La matriz anterior es fácil de entender. Como solo el primer locus se encuentra bajo selección, todos los genotipos que compartan los mismos dos alelos en el locus **A** tendrán el mismo **fitness**. Por ejemplo, A_1B_1/A_1B_1 será equivalente a A_1B_1/A_1B_2 y a A_1B_2/A_1B_2, ya que en los tres casos se trata de genotipos homocigotas A_1/A_1 para el primer locus. Con este cuadro y los resultados previos vamos a ver qué ocurre con las frecuencias alélicas luego de una generación de selección.

Calculemos primero los **fitness marginales** y el **fitness medio**. Para los **fitness marginales** usamos la ecuación (5.39)

$$\bar{w}_i = \frac{\sum_{j=1}^{4} x_i x_j w_{ij}}{x_i} = \sum_{j=1}^{4} x_j w_{ij} \tag{5.46}$$

que, sustituyendo los w_{ij} por los valores en la matriz de arriba, nos da para $\bar{w}_1$

$$\bar{w}_1 = x_1 w_{11} + x_2 w_{12} + x_3 w_{13} + x_4 w_{14} =$$
$$x_1 \cdot (1) + x_2 \cdot (1) + x_3 \cdot (1-hs) + x_4 \cdot (1-hs) \;\; \therefore \bar{w}_1 = (x_1 + x_2) + (1-hs) \cdot (x_3 + x_4)$$

pero $x_1 + x_2 = p_1$, mientras que $x_3 + x_4 = q_1$, por lo que

$$\bar{w}_1 = p_1 + (1-hs)q_1 = p_1 + q_1 - q_1 hs \Leftrightarrow \bar{w}_1 = 1 - q_1 hs \tag{5.47}$$

En forma similar, teniendo en cuenta que w_{21} es en realidad equivalente al w_{12}, porque en nuestra matriz es un requisito que $i \leqslant j$, entonces

$$\bar{w}_2 = x_1 w_{12} + x_2 w_{22} + x_3 w_{23} + x_4 w_{24} = (x_1 + x_2) + (x_3 + x_4)(1 - hs) \therefore$$
$$\bar{w}_2 = 1 - q_1 hs$$

Por otro lado, para $\bar{w}_3$ tenemos

$$\bar{w}_3 = x_1 w_{13} + x_2 w_{23} + x_3 w_{33} + x_4 w_{34} = (x_1 + x_2)(1 - hs) + (x_3 + x_4)(1 - s) \ \therefore$$
$$\bar{w}_3 = p_1 - p_1 hs + q_1 - q_1 s = p_1 + q_1 - p_1 hs - q_1 s = 1 - p_1 hs - q_1 s \qquad (5.48)$$

Finalmente, para $\bar{w}_4$ tenemos

$$\bar{w}_4 = x_1 w_{14} + x_2 w_{24} + x_3 w_{34} + x_4 w_{44} = (x_1 + x_2)(1 - hs) + (x_3 + x_4)(1 - s) \ \therefore$$
$$\bar{w}_4 = p_1 - p_1 hs + q_1 - q_1 s = p_1 + q_1 - p_1 hs - q_1 s = 1 - p_1 hs - q_1 s \qquad (5.49)$$

Resumiendo, tenemos que $\bar{w}_1 = \bar{w}_2 = 1 - q_1 hs$ y $\bar{w}_3 = \bar{w}_4 = 1 - p_1 hs - q_1 s$.

Para calcular el **fitness medio** usamos la ecuación (5.40)

$$\bar{w} = \sum_{i=1}^{i=4} x_i \bar{w}_i \qquad (5.50)$$

y, sustituyendo en ella los valores de los **fitness marginales** obtenidos antes, nos dejan

$$\bar{w} = x_1 \bar{w}_1 + x_2 \bar{w}_2 + x_3 \bar{w}_3 + x_4 \bar{w}_4 =$$
$$x_1(1 - q_1 hs) + x_2(1 - q_1 hs) + x_3(1 - p_1 hs - q_1 s) + x_4(1 - p_1 hs - q_1 s)$$
$$(x_1 + x_2)(1 - q_1 hs) + (x_3 + x_4)(1 - p_1 hs - q_1 s) \ \therefore$$
$$\bar{w} = p_1(1 - q_1 hs) + q_1(1 - p_1 hs - q_1 s) = p_1 - p_1 q_1 hs + q_1 - p_1 q_1 hs - q_1^2 s \ \therefore$$
$$\bar{w} = 1 - 2p_1 q_1 hs - q_1^2 s \qquad (5.51)$$

Este **fitness medio** es equivalente al que vimos en el capítulo anterior, cosa que no sorprende ya que en última instancia estamos trabajando con selección en un locus mientras consideramos que el otro es neutral. Veamos ahora el cambio en la frecuencia del alelo A_1. Como $p_1 = x_1 + x_2$, tenemos que $\Delta p_1 = \Delta x_1 + \Delta x_2$. A partir de las ecuaciones (5.43) y (5.45), tenemos

$$\Delta x_1 + \Delta x_2 = \frac{x_1(\bar{w}_1 - \bar{w}) - r w_{14} D}{\bar{w}} + \frac{x_2(\bar{w}_2 - \bar{w}) + r w_{14} D}{\bar{w}} \Leftrightarrow$$

$$\Delta x_1 + \Delta x_2 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D + x_2(\bar{w}_2 - \bar{w}) + rw_{14}D}{\bar{w}} \Leftrightarrow$$

$$\Delta x_1 + \Delta x_2 = \frac{x_1(\bar{w}_1 - \bar{w}) + x_2(\bar{w}_2 - \bar{w})}{\bar{w}} \tag{5.52}$$

Como $x_1 + x_2 = p_1$ y $\bar{w}_1 = \bar{w}_2$, sustituyendo en la ecuación (5.52) tenemos que

$$\Delta x_1 + \Delta x_2 = \frac{p_1(\bar{w}_1 - \bar{w})}{\bar{w}} = \Delta p_1 \tag{5.53}$$

Sustituyendo ahora por los correspondientes valores de **fitness marginal** $\bar{w}_1$ y **fitness medio** $\bar{w}$ en la ecuación (5.53), tenemos

$$\Delta p_1 = \frac{p_1[(1 - q_1hs) - (1 - 2p_1q_1hs - q_1^2s)]}{\bar{w}} \Leftrightarrow \Delta p_1 = \frac{p_1[1 - q_1hs - 1 + 2p_1q_1hs + q_1^2s]}{\bar{w}} \Leftrightarrow$$

$$\Delta p_1 = \frac{p_1[-q_1hs + 2p_1q_1hs + q_1^2s]}{\bar{w}} \Leftrightarrow \Delta p_1 = \frac{p_1q_1s[-h + p_1h + p_1h + q_1]}{\bar{w}} \Leftrightarrow$$

$$\Delta p_1 = \frac{p_1q_1s[-h + p_1h + (1 - q_1)h + q_1]}{\bar{w}} \Leftrightarrow \Delta p_1 = \frac{p_1q_1s[-h + p_1h + h - hq_1 + q_1]}{\bar{w}} \Leftrightarrow$$

$$\Delta p_1 = \frac{p_1q_1s[p_1h + q_1(1 - h)]}{\bar{w}} \tag{5.54}$$

El resultado final en (5.54) es idéntico al cambio en frecuencia $\Delta_s p$ que obtuvimos en el capítulo «Selección natural», cuando desarrollamos el caso para un solo locus. Esto tampoco nos debería extrañar, ya que continuamos trabajando considerando que la selección opera en un solo locus (el A). Pasemos ahora a analizar el efecto en el segundo locus (que es neutral). Para observar qué ocurre respecto al cambio de frecuencia en este locus debemos recordar que $p_2 = x_1 + x_3$.

Luego, partir de las ecuaciones (5.45)

$$\Delta p_2 = \Delta x_1 + \Delta x_3 = \frac{x_1(\bar{w}_1 - \bar{w}) - rDw_{14}}{\bar{w}} + \frac{x_3(\bar{w}_3 - \bar{w}) + rDw_{14}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{x_1(\bar{w}_1 - \bar{w}) + x_3(\bar{w}_3 - \bar{w}) - rDw_{14} + rDw_{14}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{x_1(\bar{w}_1 - \bar{w}) + x_3(\bar{w}_3 - \bar{w})}{\bar{w}}$$

Como $p_2 = x_1 + x_3$, tenemos

$$\Delta p_2 = \frac{x_1\bar{w}_1 + x_3\bar{w}_3 - (x_1 + x_3)\bar{w}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{x_1\bar{w}_1 + x_3\bar{w}_3 - p_2\bar{w}}{\bar{w}} \tag{5.55}$$

De acuerdo con las ecuaciones (5.5) y (5.1), $x_1 = p_1p_2 + D$ y $x_3 = q_1p_2 - D$, mientras que las ecuaciones (5.47) y (5.49) indican que $\bar{w}_1 = 1 - q_1hs$ y $\bar{w}_3 = 1 - p_1hs - q_1s$, por lo que sustituyendo en (5.55) tenemos

$$\Delta p_2 = \frac{(p_1p_2 + D)[1 - q_1hs] + (q_1p_2 - D)[1 - p_1hs - q_1s] - p_2\bar{w}}{\bar{w}} =$$

$$\frac{p_1p_2 - p_1p_2q_1hs + D - Dq_1hs + q_1p_2 - p_1p_2q_1hs - p_2q_1^2s - D + Dp_1hs + Dq_1s - p_2\bar{w}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{p_1p_2 - 2p_1p_2q_1hs - Dq_1hs + q_1p_2 - p_2q_1^2s + Dp_1hs + Dq_1s - p_2\bar{w}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{p_2(p_1 + q_1) - 2p_1p_2q_1hs - Dq_1hs - p_2q_1^2s + Dp_1hs + Dq_1s - p_2\bar{w}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{p_2[1 - 2p_1q_1hs - q_1^2s] - Dq_1hs + Dp_1hs + Dq_1s - p_2\bar{w}}{\bar{w}} \tag{5.56}$$

Pero de acuerdo con la ecuación (5.51) tenemos que $\bar{w} = 1 - 2p_1q_1hs - q_1^2s$, por lo que sustituyendo en la ecuación (5.56) tenemos ahora

$$\Delta p_2 = \frac{p_2\bar{w} - Dq_1hs + Dp_1hs + Dq_1s - p_2\bar{w}}{\bar{w}} \Leftrightarrow$$

$$\Delta p_2 = \frac{-Dq_1hs + Dp_1hs + Dq_1s}{\bar{w}} \Leftrightarrow \tag{5.57}$$

$$\Delta p_2 = \frac{Ds[p_1h - q_1h + q_1]}{\bar{w}}$$

Finalmente, como $-q_1h + q_1 = q_1(1 - h)$, sustituyendo en (5.57) tenemos

$$\Delta p_2 = \frac{Ds[p_1h + q_1(1 - h)]}{\bar{w}} \tag{5.58}$$

El resultado de la ecuación (5.58) es de fundamental importancia. Si $D \neq 0$, entonces pese a que el locus B no está directamente bajo ninguna presión selectiva, la frecuencia del alelo B_1 cambiará de acuerdo con la ecuación (5.58). Si $D > 0$, entonces crecerá la frecuencia del alelo B_1 que se encuentra en **fase de acoplamiento**, y si $D < 0$ entonces decrecerá B_1 (al aumentar A_1), o lo que es lo mismo aumentará B_2 que se encuentra en **fase de repulsión.**

PARA RECORDAR

- El cambio en las frecuencias de los gametos en una generación está dado por las ecuaciones

$$\Delta x_1 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D}{\bar{w}}$$

$$\Delta x_2 = \frac{x_2(\bar{w}_2 - \bar{w}) + rw_{14}D}{\bar{w}}$$

$$\Delta x_3 = \frac{x_3(\bar{w}_3 - \bar{w}) + rw_{14}D}{\bar{w}}$$

$$\Delta x_4 = \frac{x_4(\bar{w}_4 - \bar{w}) - rw_{14}D}{\bar{w}}$$

- El cambio en las frecuencias de los gametos en una generación está dado por las ecuaciones

$$\Delta x_1 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D}{\bar{w}}$$

$$\Delta x_2 = \frac{x_2(\bar{w}_2 - \bar{w}) + rw_{14}D}{\bar{w}}$$

$$\Delta x_3 = \frac{x_3(\bar{w}_3 - \bar{w}) + rw_{14}D}{\bar{w}}$$

$$\Delta x_4 = \frac{x_4(\bar{w}_4 - \bar{w}) - rw_{14}D}{\bar{w}}$$

- En el modelo en que un locus se encuentra bajo selección pero no los loci vecinos (en desequilibrio de ligamiento), el cambio de frecuencia del alelo A_1 será igual que en modelo de un solo locus bajo selección, es decir

$$\Delta p_1 = \frac{p_1 q_1 s[p_1 h + q_1(1-h)]}{\bar{w}} \tag{5.59}$$

- Bajo este mismo modelo, el cambio de frecuencia del alelo B_1 en el locus que no está bajo selección directa (segundo locus) está dado por la ecuación

$$\Delta p_2 = \frac{Ds[p_1 h + q_1(1-h)]}{\bar{w}} \tag{5.60}$$

En todos los casos, el fitness medio está dado por $\bar{w} = 1 - 2p_1 q_1 hs - q_1^2 s$

5.5 Arrastre genético (*genetic hitchhiking* o *genetic draft*)

A esta altura te estarás preguntando ¿qué consecuencias tiene el fenómeno descrito al final de la sección anterior? Desde un punto de vista evolutivo, el fenómeno anterior tiene la enorme importancia de que aún las regiones del genoma que no se encuentran bajo selección directa por ninguna característica (tanto a nivel nucleotídico, aminoacídico o en ningún otro nivel) y que por lo tanto podrían considerarse **neutrales** pueden estar sujetas a las consecuencias de la selección si están asociadas (en desequilibrio de ligamiento) con algún locus bajo selección. Una analogía para esto sería que para las regiones que están cerca del locus bajo selección los alelos viajan (se da un cambio en sus frecuencias alélicas) «sin pagar» (el precio de la selección), y por lo que al fenómeno anterior se lo conoce como ***hitchhiking*** genético (del inglés, «hacer dedo» o *autostop*). El término fue acuñado por John

Maynard Smith[4] y John Haigh (Maynard Smith and Haigh 1974), pese a no ser los primeros en trabajar la idea. De cualquier manera, pese al atractivo del nombre, la idea continuó relativamente desatendida hasta que el artículo de John Gillespie (2000) resaltó su posible papel como la fuerza estocástica dominante sobre la deriva.

Claramente, a partir de lo anterior, un locus bajo selección puede arrastrar todo un segmento del cromosoma en que se encuentra en la dirección en que opera la selección sobre dicho locus. Para ello es necesario que exista un desequilibrio de ligamiento con las regiones asociadas estadísticamente, usualmente las colindantes. En el caso de que el desequilibrio de ligamiento inicial sea relativamente importante, con una tasa de recombinación r baja con el coeficiente de selección s, como es poco probable la aparición de recombinación en ese fragmento entonces el alelo más asociado en el segundo locus al alelo favorecido en el primero seguirá su camino. Esto, como una función continua de la distancia, implica que todo lo que esté asociado en otras posiciones del genoma a ese alelo favorecido sufrirá cambios en la misma dirección. Si el destino del alelo favorecido es la fijación, las regiones del genoma que estén suficientemente ligadas al mismo tendrán también una ventaja para ir hacia la fijación que no ganaron por derecho propio. El fenómeno dependerá del alcance de la relación r/s. La consecuencia directa de esto es la pérdida de heterocigosidad, o el aumento de homocigosidad en los loci asociados estadísticamente.

El caso típico es aquel en que una nueva mutación que llega a un locus (llamémosle A), que hasta el momento no tenía alelos (al menos alelos que no fuesen neutrales). Este nuevo alelo A_1 se encontrará en frecuencia $\frac{1}{2N}$ y, por lo tanto, los alelos viejos (todos los que no son el A_1) se encontrarán en frecuencia $1-(\frac{1}{2N})$. Supongamos que el locus A se encuentra relativamente cerca de otro, (B), con dos alelos B_1 y B_2 selectivamente neutros. Nuestro nuevo alelo A_1 le confiere ventajas a los genotipos en los que se encuentra, por lo que siguiendo el modelo que vimos en la sección previa estos tendrán un fitness relativo de 1 (A_1A_1), $1-hs$ (A_1A_2) y $1-s$ (A_2A_2). ¿Qué ocurrirá en el locus B si la fracción de recombinantes entre A y B es r? En principio, el alelo A_1 puede haber caído en un cromosoma que tenía el alelo B_1 en el otro locus (con probabilidad p_2), o en un cromosoma conteniendo el alelo B_2 (con probabilidad $q_2 = 1-p_2$). En función de esto se puede calcular el desequilibrio de ligamiento D, así como el cambio en frecuencia p_2 (y obviamente en p_1). Sin embargo, no existe una forma fácil y cerrada de describir la trayectoria de esta evolución a lo largo de las generaciones, siendo la aproximación más fácil la simulación en una computadora.

En la figura 5.8 podemos ver el resultado de la reducción de heterocigosidad al simular la situación anterior para 3 tamaños poblacionales, de $N = 100$, $N = 1000$ y $N = 5000$ individuos, con un coeficiente de selección $s = 0{,}1$ y variando r desde 0 a 0,001 para variar la relación r/s (con signo negativo a la izquierda del locus seleccionado) y con frecuencias iniciales en el locus B (antes de la mutación en el A) de $p_2 = q_2 = \frac{1}{2}$. Como $p_2 = q_2 = \frac{1}{2}$, es indistinto el análisis de las dos situaciones (es decir, el estado del alelo en el locus B próximo a donde surge A_1). A medida que el tamaño poblacional aumenta, menos se extiende el efecto de la reducción de heterocigosidad con la distancia. Esto se podría asociar directamente con el valor del desequilibrio de ligamiento ($D = x_1x_4 - x_2x_3$), el cual se

[4] John Maynard Smith (6 enero de 1920 – 19 abril de 2004) fue un teórico de la biología evolutiva y la genética matemática. Estudió en Eton, un colegio de la elite británica y frente a la ausencia de una educación formal en ciencia se basó en los trabajos de J. B. S. Haldane (quien también había asistido a Eton). Prosiguió sus estudios en ingeniería aeronáutica en el Trinity College, donde se graduó. Después de finalizada la II Guerra Mundial abandonó la ingeniería para estudiar biología y genética, guiado por el propio Haldane. Entre sus principales contribuciones se encuentra la aplicación de la «teoría del juego» a la evolución, junto con George R. Price, en particular de las estrategias evolutivamente estables.

puede calcular directamente a partir $x_1 = \frac{1}{2N}$, $x_2 = 0$, $x_3 = 1 - \frac{1}{2N}$, $x_4 = 0{,}5$, donde modelamos la situación planteando arbitrariamente que A_1 surge en proximidad de un alelo B_1.[5] Notemos que esto implica $D = x_1x_4 - x_2x_3 = (\frac{1}{2})(\frac{1}{2N}) - 0 \times (1 - \frac{1}{2N}) = \frac{1}{4N}$. En efecto, a medida que aumenta N se incrementa el denominador y decrece el desequilibrio inicial.

La otra cosa importante que apreciamos en la figura es que la reducción no opera en forma lineal con la distancia, al menos cuando esta es medida a partir de la tasa de recombinación (en la figura $s = 0{,}1$ se mantuvo constante, por lo que el cambio en r/s se debe solo al cambio de r en nuestro ejemplo). Más allá del cambio en el desequilibrio de ligamiento inicial, en las tres gráficas de la figura se observa un patrón similar: en el entorno muy cercano al locus bajo selección se observa una reducción de la heterocigosidad muy importante, pero esta tiende a desaparecer rápidamente a medida que nos alejamos del mismo. De hecho, ni siquiera para el tamaño poblacional más pequeño el efecto se extiende más allá del entorno $\left|\frac{r}{s}\right| < 0{,}10$. Esto puede parecer, en principio, un efecto muy pequeño. Sin embargo, en humanos un estimado general es del orden de 1 Mb/cM (centi-Morgan,[6]) es decir 1 millón de nucleótidos por cada $0{,}01 = 1\,\%$ de recombinación. Tomando en cuenta este número (que, repetimos, es una primera aproximación) en el gráfico de la derecha, notemos que el corte con la línea verde (reducción del $25\,\%$) ocurre en $r/s = 0{,}042$, y como $s = 0{,}10$ entonces la tasa de recombinación es de $r = 0{,}0042 = 0{,}42\,\%$. Traducido a nucleótidos, esta estimación nos da en el entorno de 400 kb (400 mil bases). Dicho de otra forma, si tenemos una nueva mutación con coeficiente de selección $s = 0{,}10$, a 400 kb del locus donde ocurre la misma se esperaría ver una reducción de la heterocigosidad del $25\,\%$. Si pudiésemos afinar más en la gráfica (nosotros lo hicimos ya en forma numérica), a los 100 kb del locus seleccionado, con el mismo coeficiente de selección $s = 0{,}10$ y el mismo tamaño poblacional ($N = 5000$), la reducción de la heterocigosidad es $> 70\,\%$, mientras que a 10 kb la reducción es $> 95\,\%$. Obviamente un tamaño poblacional de $N = 5000$ es extremadamente pequeño en poblaciones humanas modernas, aunque no necesariamente en el pasado.

Las tasas de recombinación varían entre géneros, y aún dentro de los géneros entre especies, entre cromosomas y más aún dentro de los mismos en distintas regiones. Existen estudios variados sobre la tasa de recombinación a estas distintas escalas. Por ejemplo, en un mapa de ligamiento del genoma bovino, Arias y colaboradores concluyen que para los cromosomas autosomales la distancia genética por cM es de 1,25 Mb, o al revés $\approx 0{,}8$ Mb/cM (Arias et al. 2009). Estimaciones más precisas para el genoma humano son del orden de 0,75 Mb/cM, lo que coincide razonablemente con el genoma bovino del que no estamos tan lejos en términos evolutivos. Kong y colaboradores estimaron que el total del genoma humano en mujeres abarca 4460 cM, mientras que el de los varones es de 2590 cM (Kong et al. 2002). Sardell y Kirkpatrick encuentran que en general, en distintas especies, los ***hotspots*** de recombinación se encuentran distribuidos hacia los **telómeros**, mientras que en hembras la distribución es más uniforme (Sardell and Kirkpatrick 2020). Si bien se desconoce el mecanismo molecular que se encuentra por detrás de este fenómeno, se sabe que los paisajes de la cromatina

[5] Recordemos que esto no afecta nuestro análisis, ya que se plantea una situación con igual frecuencia de alelos B_1 y B_2. Si A_1 surgiera en proximidad del otro alelo, solo deberíamos cambiar la notación para realizar un análisis análogo.

[6] Recuerda de tus cursos de genética que Sturtevant definió la unidad en un mapa genético como la distancia entre genes a la cual, en promedio, se observa un 1% de productos de recombinación meiótica. En honor a Thomas Hunt Morgan, dicha unidad se define como 1 centi-Morgan, pudiendo ser extendida a dos puntos cuales quiera en un cromosoma (correspondan o no a genes).

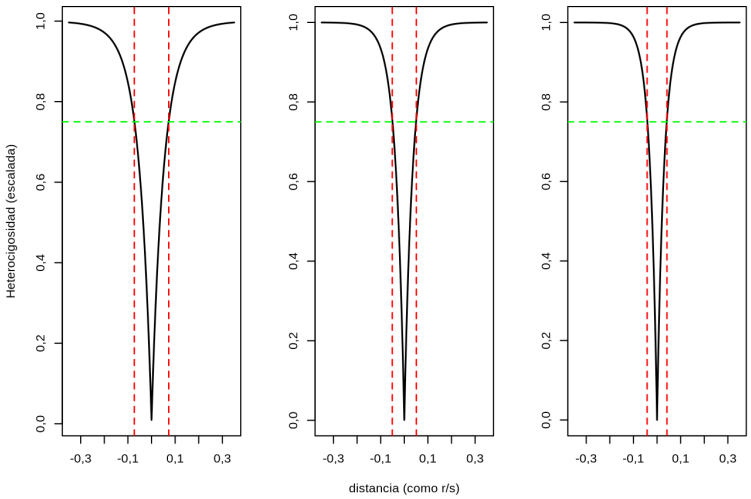


Figura 5.8. Heterocigosidad final (relativa a la inicial) en función de la distancia a un locus bajo selección direccional. La distancia al locus bajo selección se muestra como r/s a ambos lados del mismo (el cero corresponde al locus bajo selección). Los parámetros para esta simulación fueron, $N = 100$, $N = 1000$ y $N = 5000$ individuos (de izquierda a derecha), $h = \frac{1}{2}$, $s = 0{,}1$ y variando r desde 0 a 0,001 para variar la relación r/s (con signo negativo a la izquierda del locus seleccionado). Se aprecia claramente que el efecto fuerte del *hitchhiking* es bastante local, del entorno de $|r/s| < 0{,}05$, aunque esto puede significar miles de pares de bases. En rojo la distancia (como r/s) en la que se alcanza una reducción del 25 % de la heterocigosidad (es decir, la misma es el 75 % del valor inicial).

(*chromatin landscapes*, en inglés) juegan un rol importante. Kong y colaboradores, utilizando datos de 15.257 pares padre-hijo, encontraron que aproximadamente 15 % de los *hotspots* son específicos de cada sexo (Kong et al. 2010). Por otro lado, Baudat y colaboradores demostraron que el gen **PRDM9** es un determinante mayor de los sitios de recombinación en humanos y en ratón, con sus variantes alélicas asociadas a variación en el número de los mismos, lo que contribuiría a explicar variaciones dentro y posiblemente entre especies (Baudat et al. 2010).

Si bien hasta ahora hemos visto el rol de dos fuerzas estocásticas en el cambio de las frecuencias alélicas, su importancia relativa está sujeta a diferentes factores. En el caso de la deriva genética la dependencia es inversa respecto al tamaño poblacional efectivo, el factor del que depende. A tamaños poblacionales efectivos moderados su importancia es baja. Por el contrario, en el ***genetic draft*** existe una dependencia respecto al coeficiente de selección en el locus bajo selección y al desequilibrio de ligamiento entre este y los loci adyacentes (factor que a su vez tiene una cierta dependencia con el tamaño poblacional, como vimos). Sin embargo, a medida que crece el tamaño poblacional se llega a un punto donde el ***draft*** domina al ***drift*** como fuerza estocástica.

PARA RECORDAR

- El efecto de ***genetic draft*** o ***hitchhiking*** es el que corresponde a un locus bajo selección que en razón del desequilibrio de ligamiento con loci vecinos los arrastra en el proceso selectivo, más allá de que estos sean selectivamente «neutros».
- Eventualmente, si la relación r/s es relativamente grande, se producirán recombinaciones entre los loci (el locus seleccionado y el neutral), lo que llevará a un equilibrio distinto a la fijación en el locus neutral.
- Cuando aparece un nuevo alelo ventajoso en un locus que hasta el momento solo presentaba alelos neutrales, el fenómeno de ***hitchhiking*** suele implicar una reducción importante de la heterocigosidad en la vecindad del locus seleccionado. La importancia y extensión de este efecto depende de la relación r/s y del tamaño poblacional.
- En general resulta difícil o imposible encontrar soluciones explícitas a varios de los problemas asociados al ***draft***, por lo que una buena aproximación es la simulación computacional.

5.6 Causas del desequilibrio de ligamiento

Hasta ahora hemos manejado el desequilibrio de ligamiento como una condición mecanísticamente asociada a la cercanía física de los loci en el mismo cromosoma y, por lo tanto, a la probabilidad de que alelos en estos compartan destino en los gametos. En este sentido, vimos la influencia de la selección en un locus y una fracción de recombinantes menor a $\frac{1}{2}$ con otros loci como la causa de «arrastre» de esos otros loci y, en ciertas condiciones, el desarrollo de un desequilibrio de ligamiento. Cuando arriba un nuevo alelo en un locus determinado siempre lo hace en alguna configuración particular con los loci vecinos al mismo (es decir, una combinación particular de alelos en estos). Ya estudiamos previamente su devenir si el nuevo alelo tiene ventajas selectivas. Sin embargo, la selección no es la única causa posible asociada a una fracción de recombinación menor a $\frac{1}{2}$; en principio, en poblaciones

pequeñas el efecto de la deriva puede ocasionalmente incrementar la frecuencia del alelo nuevo y aún llevarlo a la fijación (aunque con baja probabilidad). ¿Cómo sucede esto? Dado que el alelo se origina para alguna configuración particular de alelos en otros loci, si el alelo incrementa su frecuencia por azar (es decir, deriva genética), en la medida que exista una ausencia relativa de recombinación también se espera un aumento de la frecuencia para los alelos de los otros loci que comparten configuración con él. Esto, al aumentar la correlación de estados alélicos compartidos, incrementará el valor de r^2 y también de D.

Si en lugar de considerar el fenómeno del desequilibrio de ligamiento desde el punto de vista de los mecanismos de transmisión de la información genética lo consideramos simplemente desde el punto de vista estadístico como la **no-independencia** de las frecuencias entre loci, entonces desaparece la necesidad de una explicación estrictamente asociada a estos mecanismos. Un ejemplo claro de esto es cuando aparecen correlaciones entre estados alélicos en diferentes cromosomas. En este caso ya no podemos usualmente apelar a la cercanía física, y es necesario invocar otros mecanismos. Uno de los procesos poblacionales que pueden explicar este «desequilibrio» (en realidad, apenas la **no-independencia**) es la existencia de determinadas estructuras poblacionales, como veremos a continuación.

Desequilibrio debido al mestizaje

Supongamos que tenemos dos poblaciones de la misma especie que han evolucionado en paralelo durante un tiempo relativamente importante respecto al tamaño poblacional como para que la deriva pueda hacer su trabajo en el cambio de las frecuencias alélicas. Supongamos también que vamos a estudiar dos loci bialélicos en particular en ambas poblaciones, que además se encuentran en equilibrio de Hardy-Weinberg. Luego del análisis genómico logramos determinar las frecuencias de los **4** haplotipos posibles, que son las que aparecen en el siguiente cuadro (bajo las columnas **Población 1** y **Población 2**):

Gameto/estadístico	Frecuencia	Población 1	Población 2	Población mezcla
A_1B_2	x_2	0,09	0,09	0,09
A_2B_1	x_3	0,09	0,09	0,09
A_2B_2	x_4	0,01	0,81	0,41
p_1	$x_1 + x_2$	0,90	0,10	0,50
p_2	$x_1 + x_3$	0,90	0,10	0,50
D	–	0	0	0,16
$D_{\max}$	–	0,09	0,09	0,25
D	–	0	0	0,64
r^2	–	0	0	0,41

A partir de las frecuencias de los haplotipos podemos calcular las frecuencias de los alelos, como ya lo hemos hecho antes. Por ejemplo, la frecuencia del alelo A_1 la calculamos como $p_1 = x_1 + x_2$. Como vemos a partir de las líneas de p_1 y p_2 en el cuadro de arriba, mientras que en la **Población 1** la frecuencia del alelo A_1 es $p_1 = 0{,}9$, en la **Población 2** $p_1 = 0{,}10$. Curiosamente (en realidad, a fin de mostrar el punto con número claros) para el alelo B_1 (del segundo locus) las frecuencias son también $p_2 = 0{,}9$ para la **Población 1** y $p_2 = 0{,}1$ para la **Población 2**.

Veamos ahora qué ocurre con el desequilibrio de ligamiento dentro de ambas poblaciones. Si recordamos que $D = x_1x_4 - x_2x_3$, haciendo los cálculos tenemos que para la **Población 1**, $D = 0{,}81 \times 0{,}01 - 0{,}09 \times 0{,}09 = 0$, mientras que para la **Población 2**, $D = 0{,}01 \times 0{,}81 - 0{,}09 \times 0{,}09 = 0$. Es decir, llegamos a la conclusión de que dentro de ambas poblaciones el desequilibrio de ligamiento para estos dos loci es cero. Como $D = 0$, entonces también lo será $D' = \frac{D}{D_{\text{máx}}}$ (no importa si usamos $D_{\text{máx}}$ o $D_{\text{mín}}$ en el denominador, ya que el numerador es cero) y también tenemos $r^2 = \frac{D^2}{p_1q_1p_2q_2} = 0$. Solo para chequear que no se trata de un artefacto producto de que dadas las frecuencias sea imposible que se encuentren en desequilibrio de ligamiento, calculamos el $D_{\text{máx}} = mn(p_1q_2, q_1p_2)$ para ambas poblaciones, que es igual a 0,09 en ambas.

Hasta aquí nada sorprendente. Veamos qué ocurre si ahora formamos una nueva población, mezclando la mitad de los individuos de la **Población 1** con la mitad de los individuos de la **Población 2**. Los resultados de mezclar estas dos poblaciones se encuentran en la última columna del cuadro de arriba. Como mezclamos mitad y mitad los individuos, las frecuencias de los haplotipos y de los alelos serán simplemente la media aritmética de las frecuencias en ambas poblaciones. Por ejemplo, para la frecuencia del haplotipo A_1B_1 tenemos $x_1 = \frac{0{,}81+0{,}01}{2} = 0{,}41$, mientras que las del A_1B_2 es de $x_2 = \frac{0{,}09+0{,}09}{2} = 0{,}09$ (el mismo cálculo para los restantes). Lo mismo para las frecuencias alélicas, donde $p_1 = \frac{0{,}9+0{,}1}{2} = 0{,}5$ y $p_2 = \frac{0{,}9+0{,}1}{2} = 0{,}5$. Veamos entonces qué ocurre con el desequilibrio de ligamiento. Usando las frecuencias de los haplotipos calculadas para la población **mezcla**, tenemos que $D = x_1x_4 - x_2x_3 = 0{,}41 \times 0{,}41 - 0{,}09 \times 0{,}09 = 0{,}16$. Este es un resultado por lo menos sorprendente, considerando que en ambas poblaciones ambos loci se encontraban en equilibrio de ligamiento. Si ahora calculamos los otros dos estadísticos de desequilibrio para esta población **mezcla**, tenemos $D' = \frac{0{,}16}{0{,}25} = 0{,}64$ y $r^2 = \frac{0{,}16^2}{0{,}5 \times 0{,}5 \times 0{,}5 \times 0{,}5} = 0{,}41$. O sea, en la nueva población no solo tenemos un desequilibrio de ligamiento distinto de cero, sino que además el mismo es del 64 % de su valor máximo posible para las frecuencias alélicas en ambos loci. ¿De dónde sale este desequilibrio?

Si prestamos atención al cuadro veremos que las diferencias de frecuencias en los alelos **entre** poblaciones es muy importante para ambos loci. Pese a esto, dentro de cada población las frecuencias de los haplotipos simplemente siguen lo esperado para la independencia (por ejemplo, p_1p_2 para el haplotipo A_1B_1). Sin embargo, pese a que en cada población uno de los haplotipos en **fase de acoplamiento** presenta una frecuencia muy alta, el descenso en la frecuencia del otro haplotipo en la misma fase es tan importante como para hacer que el producto de las frecuencias de los gametos en **fase de acoplamiento** sea idéntico al de las frecuencias de los gametos en **fase de repulsión** (es decir, $x_1x_4 = x_2x_3$). En nuestro caso $0{,}81 \times 0{,}01 = 0{,}081$ compensa perfectamente $0{,}09 \times 0{,}09 = 0{,}081$. Ahora, cuando mezclamos las dos poblaciones que tienen cada una frecuencia alta para uno de los dos gametos en **fase de acoplamiento** (pero A_1B_1 en una y A_2B_2 en la otra), el promedio son dos frecuencias relativamente altas de los dos gametos en **fase de acoplamiento** (0,41 para ambos, en nuestro caso). Es decir, la frecuencia alta en A_1B_1 no es acompañado por una reducción suficiente en A_2B_2. Ya que, en nuestro caso, el promedio de las frecuencias de los gametos en **fase de repulsión** se mantuvo en 0,09, se tiene que $x_2x_3 = 0{,}09 \times 0{,}09 = 0{,}081$. Para que existiese equilibrio de ligamiento en la nueva población mezcla x_1x_4 también debería ser 0,081, cosa que no ocurre ya que $0{,}41 \times 0{,}41 = 0{,}1681$, y de ahí el desequilibrio de ligamiento observado.

Es de notar que al invocar este mecanismo no hicimos uso de la fracción de recombinantes ni en forma explícita ni implícita, por lo que el mismo proceso podría haberse adjudicado a dos loci que se encontrasen a suficiente distancia para que se cumpla $r = \frac{1}{2}$ e incluso, con alguna modificación leve al ejemplo, que estuviesen en cromosomas diferentes.

Desequilibrio inducido por el sistema de apareamiento

En algunas especies de plantas la autofertilización es moneda corriente, variando la tasa de la misma. Cuando se produce autofertilización se produce una reducción de los heterocigotas, ya que la misma es un caso extremo de ***inbreeding***, concepto que veremos en más detalle en el capítulo «Apareamientos no-aleatorios». La lógica es igual sencilla: los loci en los que una planta dada son heterocigotas tienen una probabilidad de 50 % en autofertilización de dejar descendencia heterocigota, mientras que el otro 50 % será homocigota. Los homocigotas de la nueva generación solo podrán dejar descendencia homocigota, pero los heterocigotas tendrán nuevamente un 50 % de chance de dejar heterocigotas (es decir, 50 % del 50 % de la generación anterior). Claramente, de continuar esa progresión la fracción de heterocigotas decaerá en función de $\left(\frac{1}{2}\right)^t$, siendo t el número de generaciones.

Supongamos ahora dos loci que son homocigotos en la misma planta. Los eventos de recombinación entre ellos para la producción de gametos serán completamente irrelevantes (para estos dos loci) ya que los gametos producidos siempre serán del mismo tipo (porque para cada loci hay un solo alelo, dado que son homocigotos). Más aún, el efecto de la autofertilización disminuye el número de dobles heterocigotas, A_1B_2 y A_2B_1, fundamentales para reducir el desequilibrio de ligamiento. Por lo tanto, en función de la proporción de autofertilización esperamos una reducción de la eficiencia en el proceso que elimina el desequilibrio. Pese a la importancia de este fenómeno en el mantenimiento del desequilibrio de ligamiento, aún pequeñas tasas de ***outcrossing*** (es decir, cruzamiento con otras plantas no emparentadas) reduce rápidamente el efecto ya que quiebra los bloques haplotípicos formados. Un **bloque haplotípico** es una serie de variantes (por ejemplo, **SNP**) que se encuentran en fuerte desequilibrio de ligamiento entre todos ellos, siendo usualmente definido por un valor de r^2 (todos los marcadores tienen al menos ese valor con los adyacentes). Al romperse un bloque por un evento de recombinación, a ambos lados se sigue manteniendo el desequilibrio pero ya no entre las variantes de lados diferentes. A modo de ejemplo, en *Arabidopsis thaliana*, una planta modelo para la experimentación que normalmente experimenta más del 95 % de autofertilización, los bloques haplotípicos son del orden de las 100 kb (Nordborg et al. 2002), mientras que en especies como el maíz (que normalmente no experimenta autofertilización) los bloques haplotípicos son del orden de algunas decenas de miles de bases (Van Inghelandt et al. 2011).

El cromosoma Y

En mamíferos el cromosoma Y representa una situación particular. Al no tener un homólogo para recombinar este proceso no forma parte del repertorio de mecanismos moleculares en su evolución, a diferencia del cromosoma X que sí puede recombinar con su homólogo en el caso de las hembras. Esto hace del cromosoma Y un cromosoma muy particular. De hecho, se trata de un cromosoma con muy pocos genes, habiendo perdido más del 95 % de los mismos durante su evolución.

La ausencia de recombinación (en realidad hay un 5 % del mismo que se recombina con el X), como ya vimos, implica la imposibilidad de «barajar» los alelos en diferentes loci para que el tamiz de la selección pueda «elegir» los mejores. Es decir, si una nueva mutación beneficiosa aparece en un locus determinado, su suerte será siempre acompañar a la configuración de alelos en otros loci del gameto en el que apareció. Si está configuración era ya beneficiosa, la misma posiblemente incrementará el **fitness** del portador, pero si la configuración era detrimental el riesgo de perder esta mutación es grande. A esto se suma que el cromosoma Y está sujeto a una mayor tasa de deriva genética que el resto, puesto que su aporte es solo $\frac{1}{4}$ del aporte del resto (es la mitad de la información de los gametos sexuales en los machos, que son la mitad de los individuos, $(\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{4}$).

La imposibilidad de recombinar impide en principio eliminar por selección los errores generados en la replicación, a lo cual se suma alta tasa mutacional de este cromosoma en particular. Eso llevaría a una situación catastrófica, con acumulación continua de errores y casi sin posibilidad de liberarse de ellos. Sin embargo, existe un mecanismo que en principio funcionaría en el cromosoma Y y que permite la reparación de algunos de ellos. Este mecanismo se llama **conversión génica**, y fue propuesto para el cromosoma Y por Rozen y colaboradores (Rozen et al. 2003). Su funcionamiento se basa en la existencia de secuencias palindrómicas (una secuencia repetida, pero en el sentido contrario), las cuales permitirían la recombinación del cromosoma Y con sí mismo. Estos autores plantean que 8 palíndromos ocuparían la cuarta parte de la eucromatina de la región específica del macho (**MSY**, 95 % del largo total) en el cromosoma humano y que unos 600 nucleótidos por recién nacido habrían experimentado recombinación Y-Y.

Desequilibrio debido a recombinación reducida

Los rearreglos cromosómicos son una de las fuentes importantes de evolución a nivel genómico. Una de ellas se relaciona con la inversión de segmentos cromosómicos alrededor de los centrómeros, por ejemplo, en las moscas *Drosophila pseudoobscura* y *D. subobscura*. La inversión de un segmento de cromosoma en general inhibe la recombinación en él, por lo que todos sus genes están de alguna manera ligados físicamente. Igual que en el caso del cromosoma Y, esto conlleva la imposibilidad de «barajar» los alelos entre loci buscando las mejores combinaciones, pero al mismo tiempo las mutaciones beneficiosas que se acumulen tampoco serán separadas por eventos de recombinación. Finalmente, el mecanismo de **conversión génica** permite de alguna manera el flujo de información de un cromosoma al otro, sin recurrir a la recombinación, erosionando la estructura fija del bloque invertido.

PARA RECORDAR

- El desequilibrio de ligamiento entre loci vecinos es un fenómeno cuya ocurrencia es esperable por los mecanismos subyacentes a la transmisión de la información genética. Los eventos de recombinación, que ocurren en forma aleatoria a lo largo de los cromosomas (aunque no con distribución uniforme), tienden a romper ese desequilibrio.
- Más allá de las causas mecanísticas directas, como la vecindad en el cromosoma, otras veces el desequilibrio de ligamiento es producto de fenómenos a nivel poblacional, como, por ejemplo, cuando existe una **estructuración de las poblaciones** de la misma especie. Si existen diferencias importantes en las

frecuencias alélicas entre las subpoblaciones (producto, por ejemplo, de la deriva), aunque no exista desequilibrio de ligamiento en cada una de ellas, al considerarlas en su conjunto (una mezcla, por ejemplo), **sí se puede observar desequilibrio de ligamiento** bajo determinadas condiciones.

- El cromosoma Y presenta una situación particular ya que en la mayor parte de su extensión no tiene posibilidad de recombinar con un homólogo (por ejemplo, 95 % en humano, el restante 5 % con el X). Esto lleva a la imposibilidad de «barajar» los alelos en diferentes loci y, por lo tanto, el éxito de los nuevos alelos depende en gran medida de la «suerte» con la que cuenten al arribar a una determinada configuración de alelos en el cromosoma. El mecanismo de **conversión génica** Y-Y (o sea, dentro del mismo cromosoma) permite reducir el número de errores que se acumulan en el mismo por su escasa probabilidad de recombinación y de **conversión génica** con el homólogo.
- En algunas especies, como las moscas *Drosophila pseudoobscura* y *D. subobscura*, la inversión de un segmento de cromosoma en torno al centrómero es frecuente. Esto inhibe la recombinación en ese segmento y todos los genes dentro del mismo van de alguna manera asociados en su destino, de forma análoga a lo que ocurre con el cromosoma Y. En este caso, sin embargo, es posible el mecanismo de **conversión génica** con el homólogo, que no depende de la recombinación, permitiendo un flujo de información entre ambos cromosomas y erosionando la estructura de bloque en el segmento.

5.7 Conclusión

Si dos o más loci se encuentran segregando de forma independiente en una población, podemos recurrir a las frecuencias alélicas en ellos para calcular las frecuencias esperadas para combinaciones alélicas. En las poblaciones naturales esta expectativa de independencia no se cumple, observándose en muchos casos, en mayor o menor medida, un grado de asociación entre los alelos de diferentes loci. Diferentes métricas nos permiten cuantificar este fenómeno (tales como D, el coeficiente D' o el estadístico r^2). Los motivos detrás del fenómeno de desequilibrio de ligamiento pueden ser variados, siendo quizá el más evidente la proximidad física (inversamente proporcional a la recombinación entre dos loci). La selección natural, e incluso la deriva genética, son fuerzas evolutivas que tenderán a generar este fenómeno de no-independencia. Según la historia poblacional, pueden tender a existir diferentes combinaciones alélicas con mayor o menor frecuencia, afectando así el fenotipo de los individuos portadores y las trayectorias evolutivas observadas. En contraposición, la recombinación gamética es una fuerza que tenderá a favorecer la independencia entre loci, disminuyendo la magnitud del desequilibrio de ligamiento con el tiempo.

5.8 Actividades

5.8.1 Control de lectura

1. Tanto la selección natural como la recombinación pueden influir en las frecuencias observadas para combinaciones de alelos de dos o más loci. Describe cómo la combinación de estas fuerzas influye en las frecuencias haplotípicas observadas.

2. El valor de desequilibrio de ligamiento D es dependiente de las frecuencias alélicas presentes en una población. Discute: *a)* el rango de valores que se pueden observar en función de las frecuencias presentes en una población; *b)* la utilidad del coeficiente D' al momento de comparar la magnitud de desequilibrio de ligamiento observado en distintas poblaciones o condiciones.

3. Desde un punto de vista estadístico el desequilibrio de ligamiento implica la no-independencia entre las frecuencias alélicas entre loci. Nombra al menos un escenario biológico que genere este fenómeno (además de los mecanismos de transmisión de la información genética).

4. Describe el fenómeno denominado *genetic hithchiking*. ¿Aumenta o disminuye su relevancia a medida que aumenta el tamaño de la población considerada?

5. Las ecuaciones para calcular los máximos y mínimos teóricos esperados para el desequilibrio de ligamiento D son diferentes según exista un exceso de gametos en fase de acoplamiento ($D > 0$) o en fase de repulsión ($D < 0$). ¿A qué se debe esta diferencia? (Pista: Ten en cuenta cómo se expresan las frecuencias haplotípicas esperadas en función de las frecuencias alélicas y el valor D).

5.8.2 ¿Verdadero o falso?

1. Las tasas de recombinación varían entre especies de un mismo género. Esto se debe a que son constantes y características para cada cromosoma de las especies.

2. Las frecuencias alélicas de los loci que no tienen consecuencias directas sobre el *fitness* de los organismos portadores en un ambiente dado (es decir, loci selectivamente neutros) pueden apartarse de las frecuencias esperadas debido a los efectos de la selección.

3. La evolución del desequilibrio de ligamiento D depende de la tasa de recombinación r. Esta dependencia está dada por la ecuación $r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}$.

4. El fenómeno de *genetic hithchiking* lleva a una reducción abrupta de la heterocigosidad en los loci cercanos al locus seleccionado. Esta disminución de la heterocigosidad se extiende de forma no lineal alrededor de este locus.

5. La estructuración poblacional puede generar desequilibrio de ligamiento, aun cuando las subpoblaciones consideradas se encuentren en equilibrio de ligamiento.

Solución

1. **Falso**. Las tasas de recombinación son altamente variables, pudiendo diferir entre especies de un mismo género. No obstante, las tasas de recombinación no son características y constantes en cada cromosoma de una especie; diferentes zonas de

un mismo cromosoma pueden presentar grandes diferencias en sus tasas de recombinación (pensemos, por ejemplo, en la existencia de *hotspots* de recombinación).

2. **Verdadero.** Un ejemplo claro es el fenómeno denominado *genetic hitchhiking*: si un alelo selectivamente neutro se encuentra ligado físicamente a otro que posee una variante alélica que se ve seleccionada, su frecuencia aumentará junto a la de este alelo.

3. **Falso.** El desequilibrio de ligamiento D si presenta una dependencia con la tasa de recombinación r, pero no corresponde a la ecuación presentada (que corresponde al estadístico de asociación entre alelos r^2). El desequilibrio de ligamiento D cambia en magnitud según $D_t = (1-r)^t D_0$ (donde D_t es el desequilibrio luego de t generaciones, partiendo de un valor D_0), donde claramente se observa la dependencia con la tasa de recombinación r.

4. **Verdadero.** En un escenario de *genetic hitchhiking* las frecuencias alélicas presentes en loci vecinos al locus seleccionado también aumentan su frecuencia. Existirá por lo tanto un aumento de la homocigosidad en dichos loci, o lo que es igual una disminución de la heterocigosidad. Simulaciones computacionales muestran que este descenso de la heterocigosidad es no-lineal, y que su extensión en torno al locus seleccionado depende tanto de la relación r/s como del tamaño poblacional.

5. **Verdadero.** La estructuración de una población en subpoblaciones implica una ruptura del régimen de apareamientos aleatorios de la población global. De esta forma las frecuencias esperadas para las combinaciones de alelos de diferentes loci para cada población pueden desviarse de lo esperado para la población general, llevando a la no-independencia para los estados alélicos de estos loci (es decir, desequilibrio de ligamiento). Un caso extremo que grafica esto es el de una población que se estructura en dos poblaciones, cada una de las cuales fijando un haplotipo distinto (por ejemplo, una solo presenta el haplotipo A_1B_2 y la otra el A_2B_1).

5.8.3 Ejercicios

Ejercicio 5.1

Dos loci bialélicos A y B se encuentran a una distancia estimada de 2,3 centi-Morgans. En un momento dado las frecuencias alélicas para ambos alelos son las siguientes:

$$\begin{cases} f(A_1B_1) = 0{,}09 \\ f(A_1B_2) = 0{,}51 \\ f(A_2B_1) = 0{,}21 \\ f(A_2B_2) = 0{,}19 \end{cases}$$

a. Determina si las frecuencias genotípicas son las esperadas bajo equilibrio de Hardy-Weinberg. Discute sus resultados teniendo en cuenta el valor del desequilibrio de ligamiento D.

b. Determina las frecuencias genotípicas esperadas en la siguiente generación, asumiendo que ambos loci evolucionan bajo régimen neutral.

Solución

a. La estimación de las frecuencias alélicas esperadas bajo equilibrio de Hardy-Weinberg requiere tener en cuenta los valores de las frecuencias alélicas de A_1, A_2, B_1 y B_2 (las cuales llamaremos de ahora en más p_1, q_1, p_2 y q_2). Notemos que

$$\begin{cases} p_1 = f(A_1B_1) + f(A_1B_2) = 0{,}09 + 0{,}51 = 0{,}60 \\ q_1 = f(A_2B_1) + f(A_2B_2) = 0{,}21 + 0{,}19 = 0{,}40 \\ p_2 = f(A_1B_1) + f(A_2B_1) = 0{,}09 + 0{,}21 = 0{,}30 \\ q_2 = f(A_1B_2) + f(A_2B_2) = 0{,}51 + 0{,}19 = 0{,}70 \end{cases}$$

Bajo equilibrio Hardy-Weinberg se asume independencia en la segregación de los loci, por lo que las frecuencias genotípicas esperadas corresponden al producto de las frecuencias alélicas correspondientes. Las frecuencias genotípicas esperadas (que denotaremos con subíndice $_{HW}$) son por lo tanto:

$$\begin{cases} f(A_1B_1)_{HW} = p_1p_2 = 0{,}6 \times 0{,}3 = 0{,}18 \\ f(A_1B_2)_{HW} = p_1q_2 = 0{,}6 \times 0{,}7 = 0{,}42 \\ f(A_2B_1)_{HW} = q_1p_2 = 0{,}4 \times 0{,}3 = 0{,}12 \\ f(A_2B_2)_{HW} = q_1q_2 = 0{,}4 \times 0{,}7 = 0{,}28 \end{cases}$$

Se observa por lo tanto una desviación sustancial de las frecuencias esperadas bajo segregación independiente de los loci (es decir, existe desequilibrio de ligamiento). La magnitud del desequilibrio de ligamiento es equivalente a la magnitud de esta diferencia. Considerando la frecuencia de gametos en fase de acoplamiento con haplotipo A_1B_1 podemos calcular D, según

$$f(A_1B_1) = f(A_1B_1)_{HW} + D \Rightarrow 0{,}09 = 0{,}18 + D$$

por lo que $D = 0{,}09 - 0{,}18 = -0{,}09$. Notemos que $D < 0$ implica que existe un déficit en estos gametos en fase de acoplamiento, el cual debería ser compensado por un exceso de gametos en fase de repulsión A_1B_2. Vemos que el valor esperado $f(A_1B_2)_{HW} = 0{,}42$ y el valor observado para gametos en fase de repulsión $f(A_1B_2) = 0{,}51$, siendo la diferencia $D = 0{,}09$ y con signo opuesto al observado para gametos en fase de acoplamiento. Se puede corroborar lo mismo para los gametos en fase de acoplamiento y repulsión A_2B_2 y A_2B_1.

b. Por efecto de la recombinación las frecuencias haplotípicas cambiarán, disminuyendo generación a generación el valor de desequilibrio de ligamiento. Se espera un cambio de magnitud rD para cada haplotipo de una generación a la siguiente, donde el signo estará determinado según el haplotipo se encuentre en exceso o déficit según lo esperado bajo equilibrio de ligamiento. Por lo tanto, teniendo en cuenta que los loci se encuentran a una distancia de 2,3 centi-Morgas (es decir, la tasa de recombinación es $r = 2{,}3/100 = 0{,}023$) los cambios esperados para las frecuencias haplotípicas serán

$$\begin{cases} \Delta_r f(A_1B_1) = -0{,}023 \times (-0{,}09) = 0{,}00207 \\ \Delta_r f(A_1B_2) = -0{,}023 \times (0{,}09) = -0{,}00207 \\ \Delta_r f(A_2B_1) = -0{,}023 \times (0{,}09) = -0{,}00207 \\ \Delta_r f(A_2B_2) = -0{,}023 \times (-0{,}09) = 0{,}00207 \end{cases}$$

Por lo tanto, las frecuencias genotípicas esperadas en la siguiente generación son

$$\begin{cases} f(A_1B_1) = 0{,}09 + 0{,}00207 = 0{,}09207 \\ f(A_1B_2) = 0{,}51 + (-0{,}00207) = 0{,}50793 \\ f(A_2B_1) = 0{,}21 + (-0{,}00207) = 0{,}20793 \\ f(A_2B_2) = 0{,}19 + 0{,}00207 = 0{,}19207 \end{cases}$$

Ejercicio 5.2

Dos loci se encuentran segregando en una población de *D. melanogaster* bajo un desequilibrio de ligamiento de $D = 0{,}36$. Se logra determinar que luego de 20 generaciones este valor a $D = 0{,}20$.

a. Estima la tasa de recombinación entre los dos loci.

b. Estima el número de generaciones esperado para que el desequilibrio de ligamiento D pase a ser la mitad de su valor inicial.

Solución

a. El desequilibrio de ligamiento D entre dos loci decae en función de la tasa de recombinación r entre ellos luego de considerar t generaciones según

$$D_t = (1 - r)^t D_0$$

donde D_0 es el desequilibrio de ligamiento a tiempo inicial y D_t el observado luego de este intervalo generacional.

Tenemos entonces

$$0{,}20 = (1 - r)^{20} \times 0{,}36$$

$$\frac{0{,}20}{0{,}36} = (1 - r)^{20} \Rightarrow 1 - r = \sqrt[20]{\frac{0{,}20}{0{,}36}}$$

$$r = 1 - \sqrt[20]{\frac{0{,}20}{0{,}36}}$$

$$r = 1 - \sqrt[20]{\frac{0{,}20}{0{,}36}}$$

Por lo tanto, se deduce que existe una tasa de recombinación de $r \approx 0{,}029$ entre ambos loci.

b. La relación $D_t = (1 - r)^t D_0$ aproxima razonablemente bien a la función exponencial $D_t \approx e^{-rt} D_0$. A partir de este resultado, se puede estimar el número de generaciones

requerido para que se alcance una fracción $f = \frac{D_t}{D_0}$ del valor inicial de desequilibrio de ligamiento D_0 según la ecuación

$$t = \frac{ln(f)}{ln(1-r)}$$

En el caso planteado se tiene $f = 0{,}5$, y anteriormente estimamos $r = 1 - \sqrt[20]{\frac{0{,}20}{0{,}36}}$, por lo que se tiene

$$t = \frac{ln(0{,}5)}{ln(1-(1-\sqrt[20]{\frac{0{,}20}{0{,}36}}))} = \frac{ln(0{,}5)}{ln(\sqrt[20]{\frac{0{,}20}{0{,}36}})}$$

$$t \approx 23{,}6$$

Por lo tanto, se estima se requieren aproximadamente 24 generaciones para que el valor de desequilibrio de ligamiento D decaiga a la mitad de su valor inicial.

Ejercicio 5.3

Una población de 50 individuos diploides se encuentra segregando dos loci, A y B, los cuales presentan una baja tasa de recombinación. Inicialmente solo segregan tres alelos en la población, que llamaremos A_1, B_1 y B_2. En un momento dado surge un nuevo alelo por una mutación espontánea $A_1 \rightarrow A_2$. Los conteos haplotípicos en dicho momento son

A_1B_1	A_1B_2	A_2B_1	A_2B_2
50	49	1	0

a. Calcula el valor de D' y r^2 en esta situación, discutiendo en función de los máximos y mínimos teóricos esperados para dichos estadísticos.

b. Un tiempo después, algunos organismos portadores de los haplotipos A_1B_2 y A_2B_1 migran hacia otra región geográfica, fundando una nueva población. Algunas generaciones después se realiza un muestreo, observándose los siguientes conteos haplotípicos en la población resultante.

A_1B_1	A_1B_2	A_2B_1	A_2B_2
5	43	46	6

Recalcula los valores de D' y r^2, discutiendo las diferencias observadas en la nueva población respecto al momento en que surge el haplotipo A_2B_1 en la población original.

Solución

a. Notemos que al surgir una nueva variante alélica (en este caso, A_2) se observará siempre desequilibrio de ligamiento, ya que hasta que ocurra un evento de recombinación una de las cuatro variantes haplotípicas posibles no será observada. En este caso puntual, existe un

déficit de gametos con haplotipo A_2B_2 (en fase de acoplamiento). Las frecuencias esperadas (en función de D y las frecuencias alélicas) y las frecuencias haplotípicas observadas son

Haplotipo	Frecuencia esperada	Frecuencia observada
A_1B_1	$p_1p_2 + D$	$\frac{50}{50+49+1+0} = 0{,}50$
A_1B_2	$p_1q_2 - D$	$\frac{49}{50+49+1+0} = 0{,}49$
A_2B_1	$q_1p_2 - D$	$\frac{1}{50+49+1+0} = 0{,}01$
A_2B_2	$q_1q_2 + D$	$\frac{0}{50+49+1+0} = 0$

donde p_1, q_1, p_2, q_2 son las frecuencias de los alelos A_1, A_2, B_1 y B_2 respectivamente. Recordemos que las frecuencias alélicas son calculables a partir de las haplotípicas, siendo

$$\begin{cases} p_1 = f(A_1B_1) + f(A_1B_2) = 0{,}50 + 0{,}49 = 0{,}99 \Rightarrow q_1 = 1 - p_1 = 1 - 0{,}99 = 0{,}01 \\ p_2 = f(A_1B_1) + f(A_2B_1) = 0{,}50 + 0{,}01 = 0{,}51 \Rightarrow q_2 = 1 - p_2 = 1 - 0{,}51 = 0{,}49 \end{cases}$$

Podemos entonces calcular D con base en el déficit de frecuencia que observamos para el haplotipo A_2B_2,

$$\begin{aligned} f(A_2B_2) = q_1q_2 + D \Rightarrow D &= f(A_2B_2) - q_1q_2 \\ D &= 0 - q_1q_2 \\ D &= -0{,}01 \times 0{,}49 \\ D &= -0{,}0049 \end{aligned}$$

donde vemos que $D < 0$, acorde a lo esperado cuando existe un exceso de gametos en fase de repulsión (o lo que es igual, déficit de gametos en fase de acoplamiento).

En este caso $D' = \frac{D}{D_{\text{mín}}}$, donde

$$\begin{aligned} D_{\text{mín}} &= \text{máximo}(-p_1p_2, -q_1q_2) \\ D_{\text{mín}} &= \text{máximo}(-0{,}99 \times 0{,}51, -0{,}01 \times 0{,}49) \\ D_{\text{mín}} &= \text{máximo}(-0{,}5049, -0{,}0049) \\ D_{\text{mín}} &= -0{,}0049 \end{aligned}$$

Por lo tanto, tenemos $D' = \frac{-0{,}0049}{-0{,}0049} = 1$. En otras palabras, se observa el máximo desequilibrio de ligamiento posible para las frecuencias alélicas dadas. Esto es lógico, ya que el desequilibrio de ligamiento no puede ser mayor que en la situación en la que ya no se observa un haplotipo (claramente, no pueden existir frecuencias haplotípicas negativas). A esta situación se le suele denominar *desequilibrio de ligamiento completo.*

¿Qué sucede con el estadístico r^2? Recordemos que $r^2 = \frac{D^2}{p_1q_1p_2q_2} \Rightarrow r^2 = \frac{(-0{,}0049)^2}{0{,}99 \times 0{,}01 \times 0{,}51 \times 0{,}49} = 0{,}0097$, siendo su mínimo teórico de 0 y su máximo de 1. Por

lo tanto, mientras que D' se encuentra en su máximo teórico, el estadístico r^2 se encuentra cerca de su mínimo teórico. ¿A qué se debe esto? Recordemos que este estadístico refiere a la correlación entre los estados alélicos; podríamos pensarlo como la capacidad de predecir el estado alélico de un locus dada la información del estado alélico en el otro. En este caso particular, si bien el alelo A_2 se encuentra exclusivamente asociado al alelo B_2, el alelo A_1 se encuentra asociado casi por igual a los alelos B_1 y B_2, por lo que la correlación entre estados alélicos entre los loci es muy baja.

b. Para la nueva población fundada tenemos

Haplotipo	Frecuencia.esperada	Frecuencia.observada
A_1B_1	$p_1p_2 + D$	$\frac{5}{5+43+46+6} = 0{,}050$
A_1B_2	$p_1q_2 - D$	$\frac{43}{5+43+46+6} = 0{,}43$
A_2B_1	$q_1p_2 - D$	$\frac{46}{5+43+46+6} = 0{,}46$
A_2B_2	$q_1q_2 + D$	$\frac{6}{5+43+46+6} = 0{,}060$

Tenemos por lo tanto ahora $p_1 = 0{,}05 + 0{,}43 = 0{,}48 \Rightarrow q_1 = 1 - 0{,}48 = 0{,}52$ y $p_2 = 0{,}05 + 0{,}46 = 0{,}51 \Rightarrow q_2 = 1 - 0{,}51 = 0{,}49$. Calculemos D a partir de las frecuencias del haplotipo A_1B_2,

$$\begin{aligned} f(A_1B_2) = p_1q_2 - D \Rightarrow D &= p_1q_2 - f(A_1B_2) \\ D &= (0{,}48 \times 0{,}49) - 0{,}43 \\ D &= -0{,}1948 \end{aligned}$$

Nuevamente, $D < 0$ es coherente con el exceso de haplotipos en fase de repulsión observado. En este caso

$$\begin{aligned} D_{\text{min}} &= \text{máximo}(-p_1q_1, -p_2q_2) \\ D_{\text{min}} &= \text{máximo}(-0{,}48 \times 0{,}51, -0{,}52 \times 0{,}49) \\ D_{\text{min}} &= \text{máximo}(-0{,}2448, -0{,}2548) \Rightarrow \\ D_{\text{min}} &= -0{,}2448 \end{aligned}$$

por lo que $D' = \frac{-0{,}1948}{-0{,}2448} = 0{,}80$. Por otro lado $r^2 = \frac{D^2}{p_1q_1p_2q_2} \Rightarrow r^2 = \frac{(-0{,}1948)^2}{0{,}48 \times 0{,}52 \times 0{,}51 \times 0{,}49} = 0{,}61$. Vemos que en este caso sigue existiendo un desequilibrio de ligamiento considerable (80 % del máximo esperado para las frecuencias alélicas existentes), aunque menor al máximo teórico dado que se observan todos los haplotipos posibles. En cuanto al estadístico r^2, vemos que la correlación entre los alelos ahora es considerablemente mayor al de la población inicial. Esto es atribuible al efecto fundador. La nueva población se fundó con individuos que únicamente portaban los haplotipos A_2B_1 y A_1B_2, existiendo una correlación perfecta entre los estados alélicos de ambos loci. Si bien la recombinación generó otras combinaciones alélicas (disminuyendo la correlación de $r^2 = 1$ observable en la generación fundadora), al momento de muestreo de haplotipos la correlación sigue siendo apreciable (60 % del máximo teórico).

Ejercicio 5.4

Poco después de que se redescubrieran las leyes de Mendel, William Bateson, Edith Saunders y Reginald Punnett se dispusieron a estudiar la herencia de caracteres en *Lathyrus odoratus*. En particular, se centraron en el estudio de dos caracteres: color de flor (alelo V dominante, color violeta; alelo v recesivo, color rojo) y forma del grano de polen (alelo A dominante, forma alargada; alelo a recesivo, redondo).

Los investigadores realizaron cruces entre líneas homocigotas VV/AA (flor violeta, grano de polen alargado) y vv/aa (flor roja, **grano de polen** redondo).

La distribución de la generación resultante fue la siguiente

Fenotipo	Genotipo	Número observado
Flor violeta; grano alargado	$V_/A_$	1528
Flor violeta; grano redondo	$V_/aa$	106
Flor roja: grano alargado	$vv/A_$	117
Flor roja; grano redondo	vv/aa	381
		Total : 2132

¿Se encuentran los loci segregando de forma independiente?

Solución

Bajo equilibrio de Hardy-Weinberg la hipótesis nula es que los loci segregan de forma independiente. Realizando un doble cuadro de Punnett (recuerda tus cursos de Genética básica y la sección de ejemplos del capítulo «Variación y equilibrio de Hardy-Weinberg») se puede concluir que la proporción esperada para los cuatro fenotipos posibles es de $9:3:3:1$ (correspondientes a los genotipos $V_/A_;\ V_/aa;\ vv/A_; vv/aa$, respectivamente). Por lo tanto, podemos deducir los números esperados para cada fenotipo en una progenie de 2132 individuos, siendo

$$\begin{cases} N_{V_/A_} = 2132 \times (\frac{9}{16}) \approx 1199 \\ N_{V_/aa} = 2132 \times (\frac{3}{16}) \approx 400 \\ N_{vv/A_} = 2132 \times (\frac{3}{16}) \approx 400 \\ N_{vv/aa} = 2132 \times (\frac{1}{16}) \approx 133 \end{cases}$$

Notemos que existe un exceso de fenotipos donde se dan las combinaciones alélicas de los fenotipos parentales (es decir, con genotipos que incluyen combinación de alelos V y A o v y a), y un déficit para los fenotipos que involucran las otras combinaciones alélicas posibles. Esto parece indicar la existencia de desequilibrio de ligamiento (es decir, existencia de asociación estadística entre alelos).

Mediante un test de χ^2 podemos determinar si las diferencias observadas son estadísticamente significativas al valor de $\alpha = 1\,\%$.

Tenemos $\chi^2 = (\frac{(1528-1199)^2}{1199}) + (\frac{(106-400)^2}{400}) + (\frac{(117-400)^2}{400}) + (\frac{(381-133)^2}{133}) = 969{,}02$, lo cual es estadísticamente significativo a $\alpha = 1\%$ para $gl = 1$. Por lo tanto, se rechaza la hipótesis nula de independencia para la segregación de los loci considerados.

Ejercicio 5.5

(Inspirado en Ma et al. 2019)

La raza de ganado bovino holstein-frisia ha sido sometida a un intenso proceso de selección en América del Norte, a efectos de aumentar las tasas de producción o modificar el contenido nutricional de la leche producida. El locus DGAT1 codifica para la enzima diacilglicerol O-aciltransferasa 1, y se ha visto asociado a la relación grasa-proteína de la leche vacuna, así como al volumen de producción. Adyacente al locus DGAT1 se encuentra el locus SCRT1, el cual asumiremos es selectivamente neutro en el contexto dado.

a. En una población segregan dos alelos en el locus DGAT1 ($DGAT1_A$ y $DGAT1_G$, frecuencias $p_1 = 0{,}75;\ q_1 = 0{,}25$ respectivamente) y dos alelos en el locus SCRT1 ($SCRT1_T$ y $SCRT1_G$, frecuencias $p_2 = 0{,}45;\ q_2 = 0{,}55$ respectivamente). El locus DGAT1 se encuentra bajo un régimen de selección direccional donde el alelo $DGAT1_A$ se ve seleccionado, con $s = 0{,}2$. Se asume un modo de acción génica completamente aditivo. Si el desequilibrio de ligamiento entre los loci es $D = 0{,}22$. ¿Cuál es la frecuencia p_2' esperada para la siguiente generación?

b. En las cercanías de varios de los loci sujetos a selección direccional en líneas de la raza holstein se encuentran loci codificantes para proteínas relevantes para procesos como la espermatogénesis, la morfología y movilidad de esperma, y el desarrollo de ovocitos y embriones, entre otros. En un estudio comparativo entre líneas sujetas a selección y ganado silvestre, Ma y colaboradores observaron un aumento de variantes alélicas que son letales en homocigosidad para varios de estos loci (Ma et al. 2019). ¿Cómo explicas este fenómeno, teniendo en cuenta que tales variantes alélicas deberían estar sujetas a selección negativa? ¿Qué impacto esperas en la fecundidad de las líneas de ganado sujetas a selección conforme pasa el tiempo?

Solución

a. El cambio en la frecuencia del alelo p_2 del locus neutral SCRT se puede calcular utilizando la ecuación

$$\Delta p_2 = \frac{Ds[p_1 h + q_1(1-h)]}{\bar{w}}$$

donde $\bar{w} = 1 - 2p_1 q_1 hs - q_1^2 s$, y dado que el modo de acción génica planteado es completamente aditivo $h = 0{,}5$.

Remplazando en la ecuación con los valores dados se tiene por tanto

$$\Delta p_2 = \frac{0{,}22 \times 0{,}2 \times [0{,}75 \times 0{,}5 + 0{,}25 \times (1 - 0{,}5)]}{1 - 2 \times 0{,}75 \times 0{,}25 \times 0{,}5 \times 0{,}2 - 0{,}25^2 \times 0{,}2}$$
$$\Delta p_2 \approx 0{,}0232$$

Tenemos entonces

$$p'_2 = p_2 + \Delta p_2$$
$$p'_2 = 0{,}45 + (\frac{0{,}22 \times 0{,}2 \times [0{,}75 \times 0{,}5 + 0{,}25 \times (1 - 0{,}5)]}{1 - 2 \times 0{,}75 \times 0{,}25 \times 0{,}5 \times 0{,}2 - 0{,}25^2 \times 0{,}2})$$
$$p'_2 \approx 0{,}47$$

Vemos que, aun siendo un alelo neutral, se espera un aumento de la frecuencia para SCRT1_T debido al desequilibrio de ligamiento con un alelo seleccionado positivamente.

b. Dependiendo de la cercanía física y la tasa de recombinación entre los loci, puede existir un mayor o menor grado de desequilibrio de ligamiento entre los loci sujetos a selección y los loci codificantes para proteínas relevantes para los procesos de desarrollo mencionados. Si en los individuos seleccionados por poseer características deseadas (por ejemplo, mayor tasa de producción de leche) se encuentran ligados a los loci seleccionadas variantes alélicas letales desde el punto de vista del desarrollo (en heterocigosis), estas también aumentarán de frecuencia en la población. En otras palabras, las variantes alélicas letales para el desarrollo pueden aumentar en frecuencia en la población debido al fenómeno de *genetic hitchhiking* (igual que el alelo SCRT1_T en la primera parte de este ejercicio). A medida que aumentan en frecuencia los haplotipos que poseen estas variantes en la población, es de esperar que la fecundidad del ganado disminuya, ya que aumentan las probabilidades de que las combinaciones gaméticas sean homocigotas para los alelos letales, afectando o bien la producción o calidad del esperma, o el desarrollo embrionario. En efecto, el descenso en la fecundidad conforme se seleccionan líneas con mayor capacidad productiva es una problemática que ha sido observada en poblaciones de holstein-frisia.

Capítulo 6

Apareamientos no-aleatorios

Endijpué de tantos años, si tengo que elegir otra vez, la elijo a la Eulogia con los ojos cerrados. Porque si los abro elijo a otra.[1]

Inodoro Pereyra (el renegáu)

En los capítulos previos fuimos construyendo un panorama del cambio en las frecuencias alélicas y genotípicas bajo la acción de diferentes fuerzas evolutivas, esencialmente selección, deriva y mutación. En general, trabajamos bajo el supuesto de que en las poblaciones el apareamiento era completamente al azar y, por lo tanto, directamente ligado al equilibrio de Hardy-Weinberg (que analizamos en profundidad en el capítulo «Variación y equilibrio de Hardy-Weinberg»). Sin embargo, en muchas poblaciones tal suposición tiene poco fundamento, ya que existen estructuras de apareamiento que implican que la probabilidad de aparearse sea mayor entre ciertos individuos que entre otros. Por ejemplo, para los habitantes de los pequeños pueblos rurales, muchas veces separados entre ellos por la distancia o por barreras naturales, la posibilidad de encontrar pareja en otros puntos alejados del territorio se ve enormemente reducida y, por lo tanto, a lo largo de generaciones existirá una tensión entre emigrar (uno de los miembros de las parejas a formar) o encontrar pareja entre los individuos del pueblo, que de alguna manera serán parientes (Bourdieu 2002).

A nivel de la producción pecuaria, en muchos establecimientos la práctica común eran los rodeos cerrados, donde los reproductores machos se producían también en el establecimiento y, por lo tanto, eran hijos de alguna hembra del rodeo y hermano de varias de ellas. En el apareamiento «a campo», sin control de la reproducción para evitar que esos machos se apareen con sus madres, hermanas e hijas, el apareamiento entre parientes es inevitable. En plantas cuyos gametos tienen relativamente poca capacidad de dispersión a grandes distancias es poco razonable pensar que los individuos tengan la misma probabilidad de aparearse con los que lo rodean que con aquellos que se encuentran en extremos opuestos de la población.

Todos estos ejemplos y las situaciones que representan nos llevan a intentar entender qué ocurre cuando no podemos asumir las condiciones de **panmixia** y sus repercusiones a nivel de las frecuencias genotípicas. Para ello deberemos volver sobre un concepto central, el concepto de **identidad por ascendencia** para un par de alelos, ya que a partir de él vamos a desarrollar un marco para cuantificar el impacto del apareamiento no-aleatorio.

[1] Personaje del escritor rosarino Roberto Fontanarrosa, refiriéndose de forma humorística a su compañera de vida, la Eulogia.

OBJETIVOS DEL CAPÍTULO

☐ Introducir el concepto de identidad por ascendencia.

☐ Distinguir entre identidad por ascendencia e identidad de estado.

☐ Cómo esto afecta el equilibrio de Hardy-Weinberg,

☐ El impacto en el **fitness medio** de la población y el concepto de depresión endogámica

☐ Estudiar el rol de la estructura de poblaciones en la diferenciación genética y el papel de la migración en romper la misma.

☐ Discutir brevemente el rol de la estructuración poblacional, y el consecuente apareamiento no-aleatorio, en el surgimiento de nuevas especies. Se distinguirán distintos escenarios que reflejan este fenómeno en las poblaciones naturales.

6.1 El concepto de «identidad por ascendencia» (IBD)

El **coeficiente de endocría** se define usualmente como **la probabilidad de que los dos alelos presentes en un locus sean idénticos por ascendencia** (**IBD** por sus siglas en inglés, *Identity By Descent*). Eso nos lleva, obviamente, al concepto de identidad por ascendencia y a entender sus implicancias. Como ya vimos antes, los diferentes tipos y fuentes de variación en el ADN generan copias con diferencias respecto a la molécula original. Esto es lo que hemos usado como definición de alelo. Dos copias de una misma región o locus en el genoma pueden ser idénticas o diferentes. Si son diferentes al nivel que consideramos relevante, entonces directamente las consideramos como alelos diferentes. Sin embargo, aun siendo idénticas a nivel molecular, por ejemplo, podemos distinguir (en teoría) dos orígenes de esta similaridad: *a)* son idénticas porque derivan de exactamente la misma molécula en un ancestro reciente (que podríamos identificar) o *b)* son idénticas porque posiblemente derivan de la misma molécula en algún **ancestro remoto que no podemos identificar**. A la primera posibilidad le llamamos idéntico por ascendencia, mientras que a la segunda le llamaremos idénticos en estado (**IBS**, por sus siglas en inglés, *Identity By State*). Estos conceptos fueron ampliamente trabajados por el matemático francés Gustave Malécot,[2] y constituyen la base del tratamiento «moderno» del parentesco, aspectos en los que ahondaremos en este capítulo y más en el capítulo de «Parentesco y semejanza entre parientes».

[2] Gustave Malécot (28 diciembre de 1911 – noviembre de 1998) fue un matemático francés que dedicó una parte importante de su obra al tratamiento de los aspectos matemáticos de la genética de poblaciones y algunos de la genética cuantitativa, aunque también hizo contribuciones mayores en filosofía de la ciencia, estadística y la teoría de procesos estocásticos. Su tesis doctoral versó sobre la correlación entre parientes, trabajo que le permitió desarrollar el concepto de identidad por ascendencia. Les aportó una base probabilística a los desarrollos de Sewall Wright sobre las frecuencias genotípicas bajo apareamiento endogámico. Su obra se caracteriza por un profundo rigor matemático.

La diferencia entre **IBD** y **IBS** la podemos apreciar en dos ejemplos que aparecen en la figura 6.1. A la izquierda tenemos un individuo que deja dos descendientes, ambos heredando de él una copia del alelo azul. Claramente estamos en una situación donde los alelos azules en X e Y son idénticos por ascendencia, ya que pudimos identificar la copia ancestral (en A) desde la que derivan las dos copias en X e Y. Más aún, tanto el alelo azul en A con el alelo azul en X, como el alelo azul en A con el alelo azul en Y también son **IBD**, porque la copia en A es la considerada ancestral. A la derecha tenemos una situación de los alelos idénticos en estado. Los dos alelos en el individuo Z son idénticos, pero en la generación anterior, la generación de B y C, las dos copias de Z no provienen de una sola copia (es decir, no coalescen). En este caso, cada una de las copias en Z es **IBD** con el respectivo alelo azul del padre del que vino, pero no entre ellas.

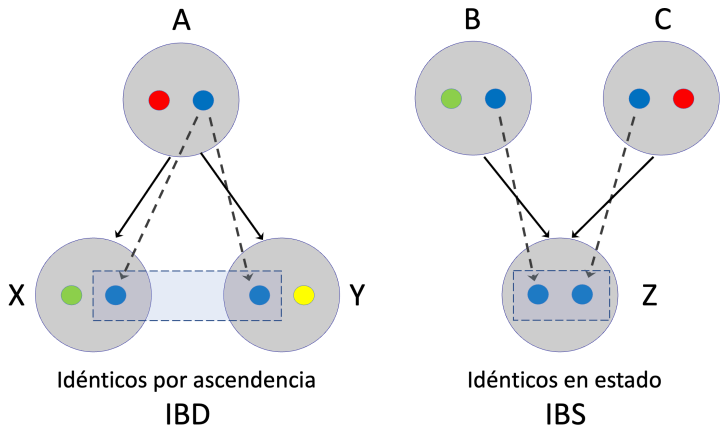


Figura 6.1. Las dos formas de identidad entre alelos. A la izquierda (identidad por ascendencia, IBD), los alelos azules en los individuos X e Y son idénticos porque descienden del **mismo alelo** en A. En cambio, a la derecha, los alelos azules en Z son idénticos en estado (IBS), porque no logramos identificar un ancestro desde donde provengan estas dos copias (que claramente debería estar en una generación anterior a B y C). Las líneas negras continuas identifican relaciones de ancestría **entre individuos**, las líneas negras punteadas reflejan relaciones **entre alelos**, mientras que los rectángulos representan los alelos considerados **IBD** o **IBS** según el caso.

Claramente, la diferencia entre considerar un par de alelos IBS o IBD es una cuestión del marco de tiempo definido, ya que, exceptuando el caso de mutaciones recurrentes, si dos moléculas (regiones) de ADN son idénticas es porque provienen de la misma molécula ancestral. Esto lo vimos cuando discutimos «El modelo coalescente». En la práctica, en el caso de pedigrís es usual resolver esto definiendo el límite como el del ancestro más antiguo reconocible de los individuos presentes y asumiendo que todos los individuos para los que se desconoce el ancestro descienden de diferentes padres. Esto es equivalente a decir que en los individuos de esa generación base $F = 0$.

Aprovechando la discusión previa podemos introducir un par de conceptos que nos serán útiles en el resto del capítulo y que revisitaremos en detalle en el capítulo «Parentesco y semejanza entre parientes». El primero es el concepto de ***coefficient of kinship*** (que se traduce usualmente como «coeficiente de parentesco»), y que denotaremos con f_{XY}, y corresponde

a la probabilidad de que dos alelos tomados al azar, uno del individuo X y otro del individuo Y, sean idénticos por ascendencia.

Veamos, por ejemplo, cómo hacer el cálculo de este coeficiente para la relación padre-hijo. Si volvemos a la figura 6.1, a la izquierda, y elegimos el hijo X por ejemplo, tenemos ilustrada la situación sobre la cuál realizar nuestros cálculos. Como hay dos alelos en A y dos alelos en X, existen 4 formas posibles de elegir un alelo en el padre y un alelo en el hijo: A_{rojo}-Y_{verde}, A_{rojo}-X_{azul}, A_{azul}-X_{verde}, A_{azul}-X_{azul}. De esas 4 formas posibles solo hay una en que los 2 alelos elegidos sean **IBD** (obviamente, la forma A_{azul}-X_{azul}). Es decir, $f_{AX} = \frac{1}{4}$.

Para el caso de medios hermanos (la relación entre X e Y en la figura), la situación es un poco más compleja, pero aún totalmente manejable. La probabilidad de que ambos hijos hereden el mismo alelo de A es un medio, ya que hay 4 posibilidades (que reciban X_{rojo}-Y_{rojo}, X_{rojo}-Y_{azul}, X_{azul}-Y_{rojo}, X_{azul}-Y_{azul}) y 2 de ellas son **IBD**. Pero ahora restrinjámonos a un caso en particular, el que contempla la herencia del alelo azul entre los hijos de A. Para este caso hay 4 combinaciones de alelos de un hermano con el otro: X_{verde}-Y_{azul}, X_{verde}-Y_{amarillo}, X_{azul}-Y_{azul}, X_{azul}-Y_{amarillo}. De estas solo una nos sirve como **IBD**, es decir $\frac{1}{4}$ de las posibilidades. Poniendo estos dos eventos independientes juntos (es decir, multiplicando las probabilidades), tenemos que $f_{XY} = (\frac{1}{2}) \times (\frac{1}{4}) = \frac{1}{8}$.

Para los casos más sencillos el cálculo es bastante obvio, pero la situación se puede complicar bastante en estructuras de pedigrí más complejas. En los casos que analizamos arriba, la situación solo involucraba las posibilidades de ningún alelo **IBD** o de un alelo **IBD**. Sin embargo, hay muchos casos donde los dos individuos pueden tener ambos alelos **IBD**. Un caso típico de esto son los hermanos enteros, que comparten padre y madre. En esos casos debemos considerar todas las posibilidades: 0, 1 o 2 alelos idénticos por ascendencia entre los dos individuos. Una forma de simplificar el procedimiento de cálculo es calcular explícitamente las probabilidades de cada uno de estos eventos, de donde será directamente posible sacar la media de alelos idénticos por **IBD**, que será la sumatoria entre 0 y 2 del número de alelos multiplicada por su probabilidad y dividida entre las cuatro formas diferentes de combinar los dos alelos en cada individuo. Si llamamos r_0, r_1 y r_2 a las probabilidades de 0, 1 o 2 alelos **IBD**, entonces el **coeficiente de parentesco** (*coefficient of kinship*) entre los individuos X e Y, f_{XY}, estará dado por

$$f_{XY} = (\frac{1}{4}) \times (0 \times r_0 + 1 \times r_1 + 2 \times r_2) = 0 \times r_0 + (\frac{1}{4}) \times r_1 + (\frac{1}{2}) \times r_2 \therefore \qquad (6.1)$$

$$f_{XY} = (\frac{1}{4}) \times r_1 + (\frac{1}{2}) \times r_2 \qquad (6.2)$$

En el cuadro siguiente recolectamos las probabilidades de compartir exactamente 0, 1 o 2 alelos **IBD**, así como los coeficientes (***coefficient of kinship***) para algunas relaciones usuales de parentesco

Parentesco	r_0	r_1	r_2	f_{XY}
Gemelos (monocigóticos)	0	0	1	$\frac{1}{2}$
Padre-Hijo	0	1	0	$\frac{1}{4}$
Abuelo-Nieto	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{8}$
Hermanos	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Medio-Hermanos	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{8}$
Primo-Hermanos	$\frac{3}{4}$	$\frac{1}{4}$	0	$\frac{1}{16}$

Por ejemplo, para calcular el coeficiente de parentesco entre Padre-Hijo basta fijarse que $r_1 = 1$ y $r_2 = 0$, por lo que

$$f_{PH} = (\frac{1}{4}) \times r_1 + (\frac{1}{2}) \times r_2 = (\frac{1}{4}) \times (1) + (\frac{1}{2}) \times (0) = \frac{1}{4} \tag{6.3}$$

Al mismo tiempo, para calcular el coeficiente de parentesco para hermanos enteros, notamos que $r_1 = \frac{1}{2}$ y $r_2 = \frac{1}{4}$, por lo que

$$f_{HE} = (\frac{1}{2}) \times r_1 + (\frac{1}{4}) \times r_2 = (\frac{1}{4}) \times (\frac{1}{2}) + (\frac{1}{2}) \times (\frac{1}{4}) = (\frac{1}{8}) + (\frac{1}{8}) = \frac{1}{4} \tag{6.4}$$

A la media del número de alelos **IBD** le llamamos $\bar{r}$ y es igual a

$$\bar{r} = 0 \times r_0 + 1 \times r_1 + 2 \times r_2 = r_1 + 2 \times r_2 \tag{6.5}$$

y una forma alternativa de medir el parentesco entre los individuos es el ***coefficient of relatedness***, r, (que lamentablemente también tiene como traducción «coeficiente de parentesco») y que es igual a la mitad de la media del número de alelos **IBD.** Es decir,

$$r = \frac{\bar{r}}{2} = (\frac{1}{2}) \times r_1 + r_2 \tag{6.6}$$

Como tanto f_{XY}, como r_{XY} (notemos que ahora usamos como suscrito el identificador de los individuos X e Y) son funciones lineales del número medio de alelos, tienen un relacionamiento directo entre ellas. De hecho,

$$f_{XY} = \frac{1}{2} r_{XY} \tag{6.7}$$

o lo que es lo mismo

$$r_{XY} = 2 f_{XY} \tag{6.8}$$

Para hacer las cosas un poquito más confusas aún, luego de ver el capítulo «El modelo genético básico» vamos a poder descomponer los efectos genéticos en varios componentes y, por lo tanto, vamos a llamar a_{XY} lo que en esta sección llamamos r_{XY}, para enfatizar que se trata del parentesco que atañe a los efectos genéticos aditivos.

PARA RECORDAR

- Llamamos a dos alelos idénticos por ascendencia (**IBD**) si provienen de una misma copia en un ancestro reciente (potencialmente identificable).
- En caso de que los alelos sean idénticos pero que no desciendan de la misma copia ancestral (en el marco de tiempo considerado), entonces llamamos a los mismos idénticos en estado (**IBS**).
- El ***coefficient of kinship***, que se traduce usualmente como coeficiente de parentesco, es la probabilidad de que dos alelos tomados al azar, uno del individuo X y otro del individuo Y, sean idénticos por ascendencia.
- Si r_0, r_1 y r_2 son las probabilidades de compartir exactamente 0, 1 o 2 alelos idénticos por ascendencia, entonces el ***coefficient of kinship*** entre los individuos X e Y queda definido por $f_{XY} = (\frac{1}{4}) \times r_1 + (\frac{1}{2}) \times r_2$.
- Una medida alternativa de parentesco es el ***coefficient of relatedness***, r, que es igual a la mitad de la media del número de alelos **IBD**, es decir, $r = \frac{\bar{r}}{2} = (\frac{1}{2}) \times r_1 + r_2$.
- La relación entre ambas formas de medir el parentesco es directa, y es igual a $r_{XY} = 2f_{XY}$.

6.2 Generalización de Hardy-Weinberg para apareamientos no-aleatorios

Basado en estos conceptos de IBD e IBS podemos definir el genotipo del locus como **autocigoto** si los alelos son réplicas de un único alelo en un ancestro común y **alocigoto** si no lo son. Es decir, más allá de que los dos alelos en un locus sean idénticos, ahora podemos distinguir un poco más, dependiendo si se trata de dos copias que provienen del mismo ancestro o de diferentes ancestros. Para entender por qué es importante esta distinción veamos qué ocurre en el caso, bastante extremo, de una especie que se reproduce mediante autofertilización (es decir, produce ella misma los gametos que generarán su descendencia). Supongamos, por ejemplo, un locus para el que nuestra especie (por ejemplo, una planta) es heterocigoto. Si cada planta deja en promedio otra planta como descendiente, esta planta hija podrá ser heterocigota para dicho locus, con probabilidad $\frac{1}{2}$ homocigota para uno de los alelos, con probabilidad $\frac{1}{4}$ u homocigota para el otro alelo, con probabilidad $\frac{1}{4}$. Si por azar (con probabilidad $\frac{1}{2}$) la planta hija es heterocigota, entonces a su vez, la descendencia de la misma será también heterocigota con probabilidad $\frac{1}{2}$. Es decir, a cada generación se pierde la mitad de los heterocigotas de la generación anterior, que irán a parar por lo tanto a los homocigotas.

Trasladado a la población de plantas y al conjunto de loci, el razonamiento anterior implica que ya no se cumplirán las proporciones esperadas en el equilibrio de Hardy-Weinberg. Es decir, habrá una falta sistemática de heterocigotas. ¿Las causas? Como no invocamos ninguna razón excepcional, salvo que en lugar de tener apareamientos aleatorios en la población

la reproducción se da por autofertilización, la causa principal debería ser esta. De hecho, como la evolución de los genotipos es al azar en cada planta y cada locus, la probabilidad de que un locus heterocigoto en una planta determinada termine en uno u otro de los homocigotos es idéntica y no se producirá un cambio en las frecuencias alélicas.

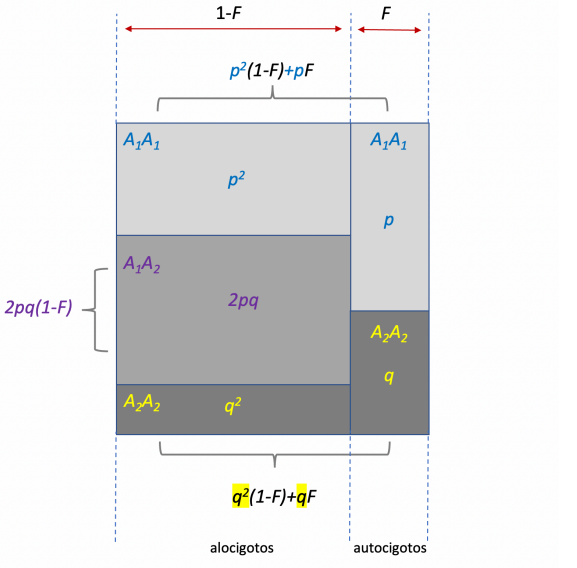


Figura 6.2. Representación de la distribución de frecuencias alélicas y genotípicas con el coeficiente de endocría F. A la derecha, la probabilidad de que un gen se vuelva **autocigótico** a causa de la endocría, y a la izquierda la probabilidad de que permanezca como **alocigótico** pese a la endocría. Los homocigotos se ven incrementados respecto al equilibrio de Hardy-Weinberg en proporción a las frecuencias de sus alelos correspondientes, p y q. Notemos la reducción de los heterocigotos (rectángulo gris central de la izquierda) en función de F. Figura propia a partir de idea en Hartl y Clark (2007).

Para demostrar lo anterior podemos recurrir a la representación que aparece en la figura 6.2. Supongamos por ahora que llamamos F a la probabilidad de que dos alelos en un locus sean IBD (por ejemplo, $F = \frac{1}{2}$ en el caso de autofertilización si asumimos que en la generación parental $F = 0$). Tenemos entonces dos posibilidades disjuntas: *a)* que los dos alelos en el locus sean IBD, con probabilidad F (la parte derecha del cuadrado en la figura), y *b)* que no sean IBD (pueden ser IBS o ser distintos alelos, parte izquierda del cuadrado), con probabilidad $1 - F$. En este último caso, el comportamiento de los alelos que formarán los cigotos es de equilibrio de Hardy-Weinberg, por lo que los aportes a los genotipos serán $p^2(1-F)$ para el A_1A_1, $2pq(1-F)$ para el A_1A_2 y $q^2(1-F)$ para el A_2A_2. En el primer caso, es decir cuando los dos alelos son IBD, como para que sean homocigotos los dos gametos deben ser iguales, entonces las frecuencias serán proporcionales a p y q (no a p^2 y q^2), por lo que el aporte a A_1A_1 será de pF y el aporte a A_2A_2 será de qF. Como se trata de dos eventos disjuntos podemos sumar las probabilidades.

Esto nos deja con las siguientes proporciones:

Genotipos	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Frecuencia con endogamia	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$
	p^2+pqF	$2pq-2pqF$	q^2+pqF
Esperado Hardy-Weinberg	p^2	$2pq$	q^2
Diferencia	pqF	$-2pqF$	pqF

Los resultados de la segunda línea salen simplemente del producto correspondiente en la primera línea. Por ejemplo, $p^2(1-F)+pF = p^2-p^2F+pF = p^2+F(p-p^2) = p^2+pqF$, ya que $pq = p(1-p) = p-p^2$. Más aún, si a estas frecuencias le restamos lo que esperamos del equilibrio Hardy-Weinberg, es decir si el apareamiento hubiese sido al azar entre todo el **pool de gametos**, entonces obtenemos la **diferencia** que aparece en la última línea. Claramente hay un enriquecimiento en los genotipos homocigotas, pqF para cada uno, a costa de los heterocigotas que pierden $pqF+pqF = 2pqF$ (el signo negativo en la última línea indica que pierden esa cantidad). Como $pqF+pqF-2pqF = 0$, la restricción de que deben sumar a 1 se mantiene ya que el balance del cambio neto respecto a HW es cero. Puesto de otra forma, $p^2+pqF+2pq-2pqF+q^2+pqF = p^2+2pq+q^2+(pqF-2pqF+pqF) = p^2+2pq+q^2 = 1$.[3]

Veamos qué ocurre ahora con las frecuencias alélicas. En todos los individuos homocigotas A_1A_1 los alelos serán A_1, mientras que en los individuos heterocigotas A_1A_2 la mitad de los alelos serán A_1. Teniendo en cuenta las frecuencias del cuadro anterior, la frecuencia de A_1 será

$$(p^2+pqF)+(\frac{1}{2})\times(2pq-2pqF) = p^2+pqF+pq-pqF = p^2+pq = p\times(p+q) = p \quad (6.9)$$

Es decir, la frecuencia del alelo A_1 no ha cambiado por la estructura reproductiva de la población, pese a que la distribución de los genotipos sí lo ha hecho. Obviamente que, como $p+q=1$, entonces $q = 1-p$ y como p sigue siendo la frecuencia de A_1, entonces q seguirá siendo la frecuencia para A_2 (es decir, la misma que si no existiese esa estructura reproductiva en la población).

Ahora, entendido el concepto de lo que implica para las poblaciones un apareamiento no aleatorio, podemos dar dos definiciones equivalentes de F:

- Probabilidad de que los alelos en un gen cualquiera de un individuo endogámico sean idénticos por ascendencia (IBD)
- Proporción de los loci en un individuo endogámico en que los alelos son idénticos por ascendencia.

[3] La persona atenta puede haber notado que uno de los ejercicios del capítulo «Deriva genética» lleva a demostrar estas relaciones, tomando como punto de partida la heterocigosidad esperada en una población finita y el coeficiente de fijación F.

PARA RECORDAR

- Cuando los apareamientos no son aleatorios dentro de la población hay una pérdida de heterocigotos y un aumento consiguiente de los homocigotos.
- El **coeficiente de endogamia** F es igual a la probabilidad de que los alelos en un gen cualquiera de un individuo endogámico sean idénticos por ascendencia (**IBD**).
- De forma equivalente, el **coeficiente de endogamia** es la proporción de los loci en un individuo endogámico en que los alelos son idénticos por ascendencia.
- Las nuevas frecuencias para una población con un coeficiente de endogamia F estarán dadas por

$$\begin{aligned} fr(A_1A_1) &= p^2(1-F)+pF \\ fr(A_1A_2) &= 2pq(1-F) \\ fr(A_2A_2) &= q^2(1-F)+qF \end{aligned} \tag{6.10}$$

6.3 F como correlación entre gametos unidos

Hasta ahora hemos trabajado con la noción de F en términos probabilísticos, ya que la definimos como la probabilidad de que los dos alelos en un gen cualquiera de un individuo endogámico sean IBD (o la proporción de loci con esas condiciones, lo cual es equivalente). Sin embargo, en la concepción original de Sewall Wright (1922), F fue definido como la correlación entre los gametos que se unen para formar un cigoto. Este enfoque es totalmente diferente, pero nos lleva (en general) a los mismos resultados que el enfoque probabilístico, aunque la interpretación es algo diferente, así como los límites. Cuando se trata de una probabilidad los límites están definidos por un valor entre 0 y 1, mientras que cuando se trata de un coeficiente de correlación, sus valores pueden variar en principio de -1 a 1.

Veamos entonces cómo arribamos a esta definición. Supongamos que mantenemos la condición de que con una probabilidad F los genes pueden transformarse en **autocigotos** a causa de la endocría. Supongamos además, sin pérdida de generalidad, que a los alelos les asignamos un valor numérico, en lugar de un código de letras. Es decir, el alelo A_1 pasa a ser representado por 1, mientras que A_2 por 0. En el cuadro siguiente podemos ver las posibilidades de formar cigotos a partir de los gametos de machos y de hembras, así como las frecuencias relativas de cada una de las combinaciones (las cuales dedujimos previamente). Vamos a usar x para el alelo del macho e y para el alelo que proviene de la hembra.

Gameto macho	Gameto hembra	Gametos unidos	Frecuencia relativa
$A_1(x=1)$	$A_1(y=1)$	$A_1A_1(1,1)$	p^2+pqF
$A_1(x=1)$	$A_2(y=0)$	$A_1A_2(1,0)$	$pq-pqF$
$A_2(x=0)$	$A_1(y=1)$	$A_2A_1(0,1)$	$pq-pqF$
$A_2(x=0)$	$A_2(y=0)$	$A_2A_2(0,0)$	q^2+pqF

Como recordarás de los cursos de estadística, la correlación de Pearson es la covarianza entre las dos variables dividida por el producto de los desvíos estándar de las mismas (o lo que es equivalente, la raíz cuadrada del producto de las varianzas). Nuestras variables x e y son los

estados alélicos en gametos de machos y de hembras, codificados en forma numérica. Para calcular la correlación, necesitamos entonces calcular varianzas y covarianzas de nuestras variables aleatorias. En particular, $\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}^2(x)$ y $\mathbb{C}\rtimes\gtrsim(xy) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)$, por lo que deberemos en primer lugar calcular las esperanzas de ambas variables, de los cuadrados de las mismas y del producto de ambas, ya que las precisamos para calcular varianzas y covarianzas. Para el caso de x, la esperanza estará dada por[4]

$$\mathbb{E}(x) = 1 \times (p^2 + pqF) + 1 \times (pq - pqF) + 0 \times (pq - pqF) + 0 \times (q^2 + pqF) \ \therefore$$
$$\mathbb{E}(x) = p^2 + pqF + pq - pqF = p^2 + pq = p \times (p + q) = p \tag{6.11}$$

De forma análoga, para y la esperanza estará dada por

$$\mathbb{E}(y) = 1 \times (p^2 + pqF) + 0 \times (pq - pqF) + 1 \times (pq - pqF) + 0 \times (q^2 + pqF) \ \therefore$$
$$\mathbb{E}(y) = p^2 + pqF + pq - pqF = p^2 + pq = p \times (p + q) = p \tag{6.12}$$

El otro componente de las varianzas es la esperanza de los cuadrados de las variables aleatorias. En el caso de x, la esperanza de x^2 será

$$\mathbb{E}(x^2) = 1^2 \times (p^2 + pqF) + 1^2 \times (pq - pqF) + 0^2 \times (pq - pqF) + 0^2 \times (q^2 + pqF) \ \therefore$$
$$\mathbb{E}(x^2) = p^2 + pqF + pq - pqF = p^2 + pq = p \times (p + q) = p \tag{6.13}$$

mientras que la de y^2 estará dada por

$$\mathbb{E}(y^2) = 1^2 \times (p^2 + pqF) + 0^2 \times (pq - pqF) + 1^2 \times (pq - pqF) + 0^2 \times (q^2 + pqF) \ \therefore$$
$$\mathbb{E}(y^2) = p^2 + pqF + pq - pqF = p^2 + pq = p \times (p + q) = p \tag{6.14}$$

En este punto tenemos todo para calcular ambas varianzas, la de x y la de y. Para el caso de x, poniendo juntos los resultados de las ecuaciones (6.11) y (6.13), será

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}^2(x) = p - p^2 = p(1 - p) = pq \tag{6.15}$$

[4] Notemos que esto equivale a sumar el resultado de la multiplicación de los valores de la columna «Gameto macho» de la tabla por la columna «Frecuencia relativa» para cada fila. Esto surge de la definición de la esperanza de un valor (promedio ponderado del mismo, es decir suma de los valores multiplicados por su probabilidad).

y para el caso de y, poniendo juntos los resultados de las ecuaciones (6.12) y (6.14), también será

$$\mathbb{V}(y) = \mathbb{E}(y^2) - \mathbb{E}^2(y) = p - p^2 = p(1-p) = pq \tag{6.16}$$

lo que no nos sorprende dada la simetría del problema (las frecuencias de los gametos son iguales en ambos sexos).

Como vimos antes, para calcular el coeficiente de correlación nos hace falta la covarianza entre ambas variables, mientras que para calcularla además de las esperanzas de ambas variables necesitamos la esperanza de su producto. En este caso, la esperanza es

$$\begin{aligned} \mathbb{E}(xy) = (1\times 1)\times(p^2+pqF) + (1\times 0)\times(pq-pqF) + \\ -(0\times 1)\times(pq-pqF) + (0\times 0)\times(q^2+pqF) \therefore \\ \mathbb{E}(xy) = p^2 + pqF \end{aligned} \tag{6.17}$$

Poniendo juntos los resultados de las ecuaciones (6.11), (6.12) y (6.17), tenemos que la covarianza está dada por

$$\mathrm{Cov}(xy) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y) = p^2 + pqF - p\times p = pqF \tag{6.18}$$

Finalmente, juntando ahora los resultados de la ecuación (6.18) (en el numerador) y de las ecuaciones (6.15) y (6.16) (para el denominador), llegamos a nuestro tan ansiado resultado

$$r_{GU} = \frac{Cov(xy)}{\sqrt{\mathbb{V}(x)\mathbb{V}(y)}} = \frac{pqF}{\sqrt{pq\times pq}} = \frac{pqF}{pq} = F \tag{6.19}$$

donde el subíndice $_{GU}$ corresponde a «gametos unidos».

En palabras, la correlación entre estados alélicos en los gametos es igual al coeficiente de endogamia F. Cuando no existe endogamia se tiene que $F = 0$, y en ese caso no existe correlación entre los estados alélicos en los gametos (es decir, un gameto con el alelo A_1 en machos no tendrá una «preferencia especial» por algún alelo en los gametos de las hembras). Dado que en este enfoque F es un coeficiente de correlación, en principio su valor podría ser negativo. No obstante, esto no ocurre en general en poblaciones naturales, por lo que las dos definiciones son equivalentes.

PARA RECORDAR

- Además del enfoque probabilístico para el coeficiente de endogamia F, existe un enfoque alternativo (que de hecho fue el original, propuesto por Sewall Wright (1922)) en el que dicho coeficiente se define como la correlación entre estados alélicos en los gametos que se unen en una población con endocría.
- La correlación entre estados alélicos en los gametos que se unen r_{GU} es igual al coeficiente de endocría, ya que para gametos x e y que se unen

$$r_{GU} = \frac{Cov(xy)}{\sqrt{\mathbb{V}(x)\mathbb{V}(y)}} = \frac{pqF}{\sqrt{pq \times pq}} = \frac{pqF}{pq} = F \tag{6.20}$$

6.4 Endocría y depresión endogámica

Como vimos en las secciones precedentes, la endocría (es decir, el apareamiento entre parientes) produce un aumento del número de homocigotas y una reducción consiguiente del número de heterocigotas. Como existe un cambio en la frecuencia de los genotipos respecto al apareamiento aleatorio, entonces resulta interesante analizar el impacto que esto produce en el **fitness** de la población. En el cuadro siguiente combinamos la información de las frecuencias genotípicas esperadas para un coeficiente de endocría F, que vimos en la sección «Generalización de Hardy-Weinberg para apareamientos no-aleatorios», con nuestro modelo de **fitness** para un locus, que vimos en la sección «Selección natural en el modelo de un locus con dos alelos».

Genotipos	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Frecuencia con endogamia	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$
	p^2+pqF	$2pq-2pqF$	q^2+pqF
Fitness	1	$1-hs$	$1-s$

El **fitness medio**, que denotaremos como $\bar{w}_F$ para marcar que se trata de una población bajo endogamia, será entonces la suma de los productos de la segunda y tercera línea del cuadro anterior. Es decir

$$\begin{aligned} \bar{w}_F &= (p^2+pqF)(1)+(2pq-2pqF)(1-hs)+(q^2+pqF)(1-s) \ \therefore \\ \bar{w}_F &= p^2+pqF+2pq-2pqF-2pqhs+2pqFhs+q^2+pqF-q^2s+pqFs \end{aligned} \tag{6.21}$$

Reagrupando y cancelando términos (por ejemplo, notando que $p^2+2pq+q^2=(p+q)^2=1^2=1$), tenemos que la ecuación (6.21) se transforma en

$$\bar{w}_F = 1-(2pqhs+q^2s)-(pqs-2pqhs)F \ \therefore \tag{6.22}$$

$$\bar{w}_F = 1-(2pqhs+q^2s)-2pqs\left(\frac{1}{2}-h\right)F \tag{6.23}$$

Para ver más claro el significado del apareamiento no aleatorio y su resumen numérico, el **coeficiente de endocría**, en el **fitness medio** de la población, podemos hacer los cambios de variables

$$a = (2pqhs + q^2 s) \tag{6.24}$$

y

$$b = 2pqs\left(\frac{1}{2} - h\right) \tag{6.25}$$

por lo que podemos escribir la ecuación (6.23) como

$$\bar{w}_F = 1 - a - bF \tag{6.26}$$

que es una ecuación lineal en F. En general, para muchos loci asociados al **fitness**, $0 < h < \frac{1}{2}$, por lo que el término $\left(\frac{1}{2} - h\right)$ será positivo, y como además $2pqs$ también lo es, entonces necesariamente el coeficiente $b > 0$. Como $0 \leqslant F \leqslant 1$, entonces el **fitness** medio, en el caso de que este dependa de un locus, decaerá en forma lineal con el coeficiente F. En este caso, a medida que aumenta el apareamiento entre parientes decrecerá el **fitness medio** de la población. Esto es una situación muy conocida por todos, y en general las especies han desarrollado mecanismos para evitar el apareamiento entre parientes muy cercanos. Por ejemplo, en la mayoría de las poblaciones humanas se ha desarrollado el tabú del incesto como forma de prevenir esta situación. Pese a esto, resulta notable que Yengo y colaboradores encuentran una prevalencia de eventos extremos de endogamia de $\sim 1/3652$ en individuos de ancestría europea, nacidos en el Reino Unido entre 1938 y 1967 (Yengo, Wray, and Visscher 2021). En las poblaciones de animales de cría esta suele ser también una variable de extremo interés a controlar, ya que produce el conocido fenómeno de la **depresión endogámica**, concepto veremos en mucho más detalle en el capítulo «Endocría, exocría, consanguinidad y depresión endogámica».

Una forma alternativa de ver el impacto de la endocría en el **fitness medio** de las poblaciones consiste en hacer la diferencia entre el **fitness** en condiciones de apareamiento aleatorio y en condiciones de endocría. De acuerdo con la ecuación (4.7), el **fitness medio** en condiciones de equilibrio Hardy-Weinberg estará dado por $\bar{w} = 1 - 2pqhs - q^2 s$, lo cual es equivalente a $\bar{w} = 1 - (2pqhs + q^2 s)$. Al mismo tiempo, de acuerdo con la ecuación (6.23), $\bar{w}_F = 1 - (2pqhs + q^2 s) - 2pqs\left(\frac{1}{2} - h\right)F$, por lo que la diferencia estará dada por

Fitness.medio	.
$\bar{w}_F$	$1 - (2pqhs + q^2 s) - 2pqs\left(\frac{1}{2} - h\right)F$
$\bar{w}$	$1 - (2pqhs + q^2 s)$
$\bar{w}_F - \bar{w}$	$-2pqs\left(\frac{1}{2} - h\right)F$

De acuerdo con la diferencia $\bar{w}_F - \bar{w}$, la disminución del valor de **fitness medio** que acontece por el apareamiento entre parientes será igual a $-2pqs\left(\frac{1}{2} - h\right)F = -aF$; es decir, la disminución en el fitness medio será lineal respecto al valor de F.

PARA RECORDAR

- El **fitness medio** de una población en la que el **coeficiente de endocría** es F estará dada por $\bar{w}_F = 1 - (2pqhs + q^2 s) - 2pqs\left(\frac{1}{2} - h\right)F$.
- Si definimos $a = (2pqhs + q^2 s)$ y $b = 2pqs\left(\frac{1}{2} - h\right)$, entonces el **fitness medio** quedará expresado como $\bar{w}_F = 1 - a - bF$.
- La disminución del valor de **fitness medio** que acontece por el apareamiento entre parientes, respecto al apareamiento bajo condiciones de equilibrio Hardy-Weinberg, será igual $-2pqs\left(\frac{1}{2} - h\right)F = -aF$. De otra forma, la disminución del **fitness medio** será directamente proporcional al valor de F.

6.5 Un caso extremo: la autogamia

En plantas es común encontrar el fenómeno de la **autogamia**, en el que la planta se reproduce consigo misma. Es decir, desde el punto de vista genético la **autogamia** consiste en la fusión de gametos femeninos y masculinos producidos por el mismo individuo. Esto producirá un incremento en el **coeficiente de endocría** de generación en generación. Como vimos en la sección «El concepto de 'identidad por ascendencia' (IBD)», el coeficiente de consanguinidad de un individuo es igual al *coefficient of kinship* entre sus padres. En el caso de tratarse de **autogamia** los padres son el mismo individuo (A), por lo que $F = \frac{1}{2}(1 + F_A)$, donde F_A es el coeficiente de consanguinidad de A. Multiplicamos por $(1 + F_A)$ porque si el individuo A es consanguíneo, los alelos en su descendencia tienen mayor probabilidad de ser **IBD**, exactamente en esa proporción (abundaremos en esto, así como una forma gráfica de calcularlo a partir de pedigrís en el capítulo «Parentesco y semejanza entre parientes»).

Para ver cómo evoluciona el coeficiente F a lo largo de las generaciones alcanza con ver que el sistema de apareamiento es regular. Es decir, para una generación t dada,

$$F_t = \frac{1}{2}(1 + F_{(t-1)}) \tag{6.27}$$

Si en lugar de F usamos el **coeficiente de panmixia**, que es $1 - F$, entonces lo anterior se puede transformar fácilmente multiplicando la ecuación (6.27) por -1 y sumando 1 a

$$(1 - F_t) = 1 - \frac{1}{2}(1 + F_{(t-1)}) = \frac{1}{2} - \frac{1}{2}F_{(t-1)} = \frac{1}{2}(1 - F_{(t-1)}) \tag{6.28}$$

Claramente, la ecuación (6.28) invita a seguir con la recurrencia. Por ejemplo, dos generaciones hacia atrás será

$$(1 - F_t) = \frac{1}{2}\left[\frac{1}{2}(1 - F_{(t-2)})\right] = (\frac{1}{2})^2(1 - F_{(t-2)}) \tag{6.29}$$

Con la misma lógica, t generaciones hacia atrás el **coeficiente de panmixia** será

$$(1 - F_t) = (\frac{1}{2})^t(1 - F_0) \tag{6.30}$$

con F_0 igual al coeficiente de endocría en la generación inicial. En el caso de que $F_0 = 0$, el cual establecemos sencillamente porque no lo conocemos y no podemos rastrear su valor, la ecuación (6.30) se reduce a

$$(1 - F_t) = (\frac{1}{2})^t(1 - F_0) = (\frac{1}{2})^t(1 - 0) = (\frac{1}{2})^t \tag{6.31}$$

Si volvemos a manejarnos con el coeficiente de endocría, basta volver a multiplicar por -1 y sumar 1 para llegar a

$$1 - (1 - F_t) = F_t = 1 - (\frac{1}{2})^t(1 - F_0) \tag{6.32}$$

que en el caso de $F_0 = 0$ se reduce a $F_t = 1-(\frac{1}{2})^t$. Como vimos antes, F nos permite seguir la evolución de la reducción en heterocigotas en función del aumento en los homocigotas. La evolución a lo largo de las primeras 10 generaciones se puede ver en la figura 6.3, donde a partir de un valor inicial de $F_0 = 0$ en una generación se alcanza $F_1 = 0{,}5$ y menos de 4 generaciones son suficientes para alcanzar un valor $F > 0{,}90$.

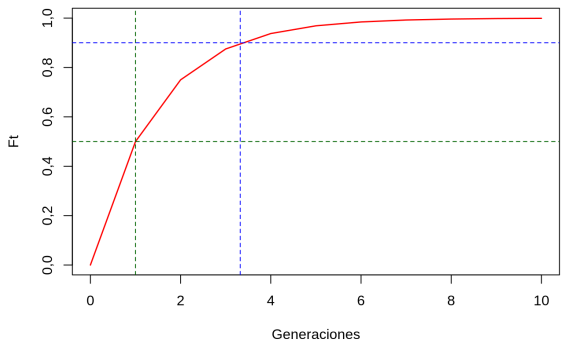


Figura 6.3. Evolución del coeficiente de endocría a lo largo de las generaciones para el caso de **autogamia**. A partir de un valor inicial de $F_0 = 0$, en una generación el coeficiente alcanza $F_1 = 0{,}5$ (líneas verdes) y en $\frac{\ln 0{,}1}{\ln 0{,}5} = 3{,}322 < 4$ generaciones alcanza el 90 % de su valor máximo.

La **autogamia** implica entonces un aumento del número de homocigotos y una disminución consiguiente de los heterocigotos, sin implicar un cambio de frecuencias alélicas. Sin embargo, cuando existe selección esto ya no se mantiene. Al mismo tiempo que se produce una reducción en el **fitness medio** de la población, como vimos en la sección «Endocría y depresión endogámica», la selección contra los recesivos deletéreos será mucho más eficiente ya que los mismos dejarán de estar en lo heterocigotos (invisibles o poco visibles para la selección) para pasar a estar mayormente en los homocigotos recesivos, que son vistos plenamente por la selección.

Al mismo tiempo, el desequilibrio de ligamiento tardará mucho más en romperse debido a la reducción en la recombinación, producto de la abundancia de homocigotos. Un ejemplo de especie de interés comercial que experimenta importante **autogamia** es la cebada (*Hordeum vulgare*). Al mismo tiempo, las especies que experimentan **autogamia** en forma parcial permiten hacer una selección extrema para uniformidad de características, mediante la explotación del mecanismo de **autogamia** durante varias generaciones, lo que implica transformar en homocigotos la mayor parte de los loci. Esto trae aparejado la evidente reducción del **fitness**, como ya vimos en la sección «Endocría y depresión endogámica». Sin embargo, si ahora cruzamos dos variedades altamente uniformes, el híbrido será uniforme también porque se trata de una $F1$ entre homocigotos, pero explotando el **vigor híbrido** (lo opuesto a la depresión endogámica y que veremos mucho más adelante en detalle) poseerá también características excelentes desde el punto de vista productivo.

PARA RECORDAR

- La **autogamia** es la reproducción del individuo consigo mismo y desde el punto de vista genético consiste en la fusión de gametos femeninos y masculinos producidos por el mismo individuo.
- En el caso de autogamia el coeficiente de endocría en la generación t, F_t dado que el valor inicial de la misma es F_0, es igual $1-(1-F_t)=F_t=1-(\frac{1}{2})^t(1-F_0)$.
- La autogamia incrementará rápidamente la frecuencia de homocigotos en la población, a costa de los heterocigotos. Esto provocará una disminución del **fitness medio** de la población, pero en el caso de alelos recesivos deletéreos también contribuirá a su disminución.

6.6 El coeficiente de endocría y los estadísticos F

Como ya vimos previamente en otros capítulos (por ejemplo, en el capítulo «Deriva genética»), las poblaciones de muchas especies se encuentran estructuradas en subpoblaciones, cada una de ellas con su propia dinámica evolutiva. Por ejemplo, dentro de una región tenemos diferentes cuencas, que a su vez tienen diversos afluentes, donde cada una de ellos posee una conexión con diferentes espejos de agua. Eso nos permite identificar, en forma arbitraria al principio, diferentes niveles dentro de una jerarquía: cuencas, dentro de ellas afluentes y dentro de los mismos los espejos de agua. La estructura puede ser también mucho más sencilla, con un par de niveles, por ejemplo, poblaciones aisladas (por una barrera física, por ejemplo) dentro de un conjunto regional de las mismas (es decir, un **ensemble** poblacional). En este sentido, en lo que nos atañe en el presente capítulo, la estructura reproductiva de cada población del **ensemble** puede seguir sus propios patrones, lo que se reflejará también en la relación entre frecuencias alélicas y genotípicas de cada una de ellas. Como vimos antes, el apartamiento del apareamiento al azar se refleja usualmente en una carencia de heterocigotos respecto a lo esperado para el equilibrio de Hardy-Weinberg. Es más, definimos un índice de fijación F como la probabilidad de encontrar en un locus dos alelos idénticos por ascendencia y vimos que la pérdida de heterocigotos era igual a $2pqF$. Sin embargo, al tener distintas poblaciones (y, por lo tanto, niveles, por ejemplo, individuos, poblaciones, **ensemble**) necesitamos de alguna forma entender a qué nivel se producen los apartamientos respecto al equilibrio Hardy-Weinberg. Para esto definiremos una serie de

estadísticos, primero para representar la heterocigosidad observada y esperada a distintos niveles, y luego unos estadísticos de fijación análogos a nuestro F.

En la figura 6.4 se observa una representación de una estructura poblacional de una especie diploide, donde diferentes poblaciones (enmarcadas por recuadros verdes) poseen una diferente proporción de alelos de dos colores, rojo y azul, así como distintas proporciones de los 3 genotipos. Todas estas poblaciones se encuentran agregadas en un gran **ensemble**, encuadrado en azul.

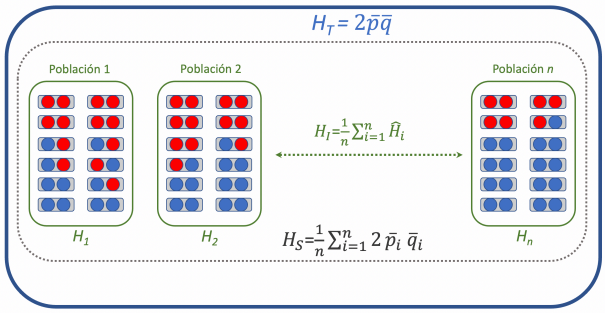


Figura 6.4. Estructuración de una gran población (marco en azul) como un **ensemble** de poblaciones (marcos en verde) y la distribución de la heterocigosidad observada y esperada a los distintos niveles de agregación, así como los correspondientes estadísticos.

Claramente, dentro de cada población es posible contar la proporción de individuos heterocigotas (si estos fuesen distinguibles por su fenotipo, o si los podemos identificar a nivel molecular) y, por lo tanto, si tenemos n poblaciones, para cada una de ellas obtendremos un estimador $\hat{H}_i$. Decimos que obtenemos un estimador, y por eso lo señalamos con el «gorro» (es decir, acento circunflejo:ˆ) porque normalmente solo tenemos una muestra de individuos de cada población. Una manera sencilla de resumir la información respecto a la proporción de heterocigotas de cada una de ellas será la media de la heterocigosidad, es decir

$$H_I = \frac{1}{n}\sum_{i=1}^{i=n} \hat{H}_i \qquad (6.33)$$

Se trata, en este caso, de una forma de resumir la información sobre la **heterocigosidad observada**, producto de los F a nivel individual (de ahí el subíndice I). Para cuantificar la reducción respecto a lo esperado al existir dicha estructura, en primer lugar debemos calcular una medida resumen de **lo esperado** en cada población si las mismas se aparearan (a la interna) al azar.

Otra vez, una medida sencilla lo constituye la media de lo esperado en cada una. Por lo tanto,

$$H_S = \frac{1}{n}\sum_{i=1}^{i=n} 2p_i q_i \tag{6.34}$$

ya que en cada una de las n poblaciones las frecuencias alélicas pueden ser diferentes (de ahí los términos p_i y $q_i = 1 - p_i$). Llegado este punto solo nos queda definir un nivel de heterocigosidad. Si ponemos todos los individuos juntos, sin importar a qué población pertenezcan, entonces la **heterocigosidad esperada global** es igual a

$$H_T = 2\bar{p}\bar{q} \tag{6.35}$$

Con todos estos estimadores de heterocigosidad en mano estamos ahora en condiciones de entender la estructura de la pérdida de heterocigosisidad (es decir, a qué nivel y en con qué magnitud se da la pérdida). El primer nivel, dentro de la estructura representada en la figura 6.4 es el que corresponde a la pérdida entre lo esperado y lo observado dentro de las poblaciones. Como nuestras medidas de heterocigosis esperada y observada son H_S y H_I respectivamente, la diferencia entre estas medidas (estandarizada por la esperada) será nuestro índice de fijación a nivel intra población. Es decir, nuestro índice de fijación F_{IS} será calculado según

$$F_{IS} = \frac{H_S - H_I}{H_S} \tag{6.36}$$

En este caso, las letras que aparecen en el subíndice representan lo individual o intrapoblación (I) versus lo esperado para la población (S, de subpoblación). Más aún, F_{IS} es exactamente igual al F definido previamente, en la medida en que la media represente adecuadamente lo que ocurre dentro de cada población.

En nuestra estructura, el siguiente nivel sería comparar lo esperado en la población total con lo esperado si se cumpliese el equilibrio de Hardy-Weinberg en cada una de las poblaciones. Una nota importante es que aún cumpliéndose el equilibrio de Hardy-Weinberg en cada población no necesariamente se cumple cuando consideramos a todos los individuos en conjunto. Un ejemplo claro de esto podemos verlo en la figura 6.5. Dos poblaciones de ovejas en lados opuestos de un río han fijado diferentes alelos para el color: a la izquierda blanco, a la derecha verde. Dentro de cada población las ovejas se aparean libremente, por lo que podemos esperar que se encuentren en el equilibrio de Hardy-Weinberg. El río es una barrera infranqueable para las ovejas, por lo que no hay migración entre ambas poblaciones. Si, mientras recorremos el río en bote, capturamos a las ovejas que pastan en la ribera (el mismo número en una ribera que en la otra), les extraemos una muestra y luego secuenciamos el locus «del color» (en principio no existe dicha cosa, la herencia del color en mamíferos es extremadamente compleja) vamos a llegar a unos números que nos llamarán

la atención. Mientras que las frecuencias del alelo blanco y del verde son iguales, es decir $p = q$, por lo que esperamos una frecuencia $2pq = 2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$ heterocigotos. Sin embargo, en nuestra muestra no hay ningún individuo heterocigoto ya que las ovejas son de dos poblaciones en las que se ha fijado el alelo para el color, por lo que al considerar todo el **ensemble** (a través de una muestra de ambas poblaciones) nos encontramos con un déficit enorme de heterocigotos. La razón es obvia: se trata de una forma de estructura de apareamiento no-completamente-al-azar (ya que los individuos solo se pueden aparear con los de su respectiva población) y, por lo tanto, existirá una carencia de heterocigotos, como ya vimos en la sección «Generalización de Hardy-Weinberg para apareamientos no-aleatorios». Más allá de ser obvia la razón, lo veremos en más detalle en la sección «El efecto Wahlund».

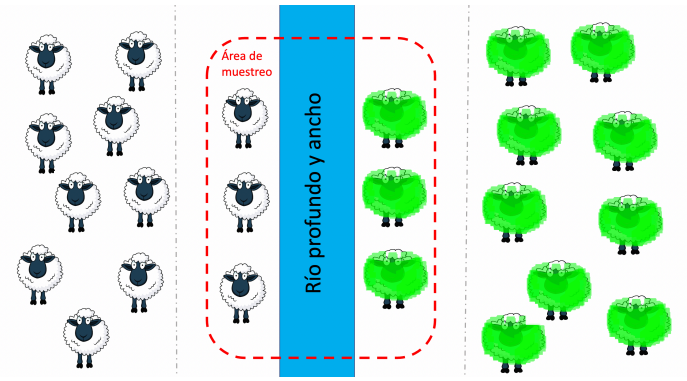


Figura 6.5. Reducción en el número de heterocigotos a causa de la estructuración en poblaciones disjuntas. Las ovejas a ambos lados del río, una barrera infranqueable para las mismas, han fijado alelos diferentes para el color de la lana. Si muestreamos las ovejas que aparecen en la ribera, aproximadamente igual número a ambos lados, obtendremos una frecuencia intermedia para los alelos blanco y verde, por lo que esperamos cerca del 50 % de heterocigotos ($2pq = 2\frac{1}{2}\frac{1}{2} = \frac{1}{2}$), pero en realidad no hay ningún individuo heterocigoto ya que los alelos están fijados en las dos poblaciones (diseño de las ovejas de http://cliparts.co/clipart/2377087 Fuente: Cliparts.co).

Como hicimos para F_{IS}, para definir el índice de fijación de las subpoblaciones respecto al total vamos a hacer la resta de la heterocigosidad total esperada respecto a la esperada para las poblaciones y estandarizamos dividiendo entre la heterocigosidad total esperada. Esta no es ni más ni menos que $2\bar{p}\bar{q}$, siendo $\bar{p}$ y $\bar{q} = 1-\bar{p}$ las frecuencias calculadas considerando todo el **ensemble** en su conjunto. Es decir, el índice de fijación de las subpoblaciones respecto al total será

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{6.37}$$

Finalmente, con la misma lógica podemos definir el índice de fijación entre lo que ocurre **dentro** de las poblaciones y el total como

$$F_{IT} = \frac{H_T - H_I}{H_T} \tag{6.38}$$

Todos estos índices deben ser entendidos en conjunto. De hecho, Sewall Wright planteó la relación entre ellos (1943; 1951) como

$$F_{ST} = \frac{(F_{IT} - F_{IS})}{(1 - F_{IS})} \tag{6.39}$$

Es posible llegar a otra forma de la relación entre los índices, también muy informativa. Operando sobre la ecuación (6.40), tenemos que

$$\begin{aligned}(1 - F_{IS}) \times F_{ST} = F_{IT} - F_{IS} \ \therefore \\ F_{IT} = F_{ST} + F_{IS} - F_{ST}F_{IS}\end{aligned} \tag{6.40}$$

Multiplicando por -1 y sumando 1 a ambos lados, tenemos

$$\begin{aligned}-F_{IT} = -F_{ST} - F_{IS} + F_{ST}F_{IS}F_{ST} + F_{IS} \Leftrightarrow \\ 1 - F_{IT} = 1 - F_{ST} - F_{IS} + F_{ST}F_{IS}F_{ST} + F_{IS}\end{aligned} \tag{6.41}$$

Pero $(1 - F_{ST})(1 - F_{IS}) = 1 - F_{ST} - F_{IS} + F_{ST}F_{IS}F_{ST} + F_{IS}$, por lo que poniendo todo junto, la ecuación (6.41) se transforma en

$$(1 - F_{IT}) = (1 - F_{ST})(1 - F_{IS}) \tag{6.42}$$

Por lo que vimos, F_{ST} representa la diferenciación genética a causa de la estructuración en poblaciones independientes. En general no resulta trivial interpretar su valor sin el apoyo de más evidencia proveniente de otras fuentes, pero como forma de orientación, Hartl y Clark proponen una serie de guías prácticas dadas por Sewall Wright que pueden ayudar a entender la situación (Hartl and Clark 2007; Sewall Wright 1978). Estas son

- Valores de $0 \leqslant F_{ST} < 0{,}05$ indicarían **poca** diferenciación genética.
- Valores de $0{,}05 \leqslant F_{ST} < 0{,}15$ indicarían diferenciación genética **moderada**.
- Valores de $0{,}15 \leqslant F_{ST} < 0{,}25$ indicarían **gran** diferenciación genética.
- Valores de $F_{ST} \geqslant 0{,}25$ indicarían **muy grande** diferenciación genética.

Sin embargo, como también mencionan Hartl y Clark (2007), el mismo Wright advierte que aún valores de F_{ST} menores a 0,05 podrían estar indicando una diferenciación genética significativa, por lo que la interpretación de estos valores no es inmediata (Sewall Wright 1978).

Los estadísticos F son muy usados para comprender la diversidad genética aún entre especies cercanas. Así, por ejemplo, Loaisiga y colaboradores en un estudio sobre la diversidad del teosinte (*Zea perennis*), una planta que junto con el maíz (*Zea mays*) forman el género *Zea*, comparan seis especies diferentes muy relacionadas entre sí en diferentes regiones de Centroamérica (Loaisiga et al. 2012). Como los híbridos F_1 de teosinte $\times$ maíz son fértiles, estas especies representan un reservorio importante de material genético para mantener la diversidad del maíz o aún para introducir variantes de genes importantes para la producción (por ejemplo, resistencia a patógenos, tolerancia a metales pesados, etcétera). Para estudiar la diversidad estos autores utilizan un juego de 21 microsatélites (SSR) en 120 individuos, detectando un total de 109 alelos diferentes. A partir del análisis de las frecuencias de los alelos y de los genotipos, calculan para el conjunto de las especies un $F_{IS} = 0{,}0815$, mientras que el $F_{ST} = 0{,}2017$. Con estos datos resulta claro que el apartamiento de Hardy-Weinberg dentro de las poblaciones es relativamente pequeño, mientras que la mayor parte de la diferenciación aparece entre especies (el equivalente a las poblaciones en nuestros razonamientos previos). De esta forma, el estudio sugiere que es fundamental la conservación de las diferentes especies ya que es donde reside la mayor parte de la diversidad.

Ejemplo 6.1

Consideremos un **ensemble** de poblaciones constituido por las 3 poblaciones que aparecen en la figura 6.4. Veamos si existe deficiencia de heterocigotos a algún nivel que nos permita sospechar apareamientos no-aleatorios dentro de las mismas o alguna estructuración geográfica.

Parámetro	Población 1	Población 2	Población 3	Media
fr(rojo)	$p_1 = \frac{13}{24} = 0{,}542$	$p_2 = \frac{12}{24} = 0{,}500$	$p_3 = \frac{7}{24} = 0{,}292$	$\bar{p} = \frac{32}{72} = 0{,}444$
$\hat{H}$	$H_1 = \frac{5}{12} = 0{,}417$	$H_2 = \frac{2}{12} = 0{,}167$	$H_3 = \frac{1}{12} = 0{,}083$	$\frac{8}{36} = 0{,}222$
$2p_iq_i$	$2\frac{13}{24}\frac{11}{24} = 0{,}497$	$2\frac{12}{24}\frac{12}{24} = 0{,}500$	$2\frac{7}{24}\frac{17}{24} = 0{,}413$	0,470

Primero vamos a calcular todos los índices de heterocigosidad. La heterocigosidad observada está dada por la ecuación (6.33), es decir (con $n = 3$)

$$H_I = \frac{1}{n}\sum_{i=1}^{i=n} \hat{H}_i = (\frac{1}{3}) \times (0{,}417 + 0{,}167 + 0{,}083) = 0{,}222 \tag{6.43}$$

que es el valor en la última columna del cuadro anterior, en la fila correspondiente a la heterocigosidad observada. Si en cada población los apareamientos fuesen al azar esperaríamos en cada una de ellas una proporción de heterocigotos igual a $2p_iq_i$, que es la que figura en la tercera línea del cuadro anterior.

Por lo tanto, la heterocigosidad media esperada, de acuerdo con la ecuación (6.34), es

$$H_S = \frac{1}{n}\sum_{i=1}^{i=n} 2p_i q_i = (\frac{1}{3}) \times (0{,}497 + 0{,}500 + 0{,}413) = 0{,}470 \tag{6.44}$$

Finalmente, la heterocigosidad total esperada, de acuerdo con la ecuación (6.35), es

$$H_T = 2\bar{p}\bar{q} = 2\bar{p}(1-\bar{p}) = 2 \times 0{,}444 \times (1 - 0{,}444) = 0{,}494 \tag{6.45}$$

A partir de estos datos podemos calcular ahora los índices de fijación correspondientes. El primero, el que hace a la estructura interna de las poblaciones está dado por la ecuación (6.36) y es, por lo tanto

$$F_{IS} = \frac{H_S - H_I}{H_S} = \frac{0{,}470 - 0{,}222}{0{,}470} = 0{,}528 \tag{6.46}$$

Basados en este número, podemos decir que se trata de un valor realmente alto y que por lo tanto **dentro** de las poblaciones (o alguna de ellas) el apareamiento se aparta claramente de lo esperable para las condiciones de **panmixia.**

Veamos que ocurre al segundo nivel, es decir el índice de fijación de las poblaciones respecto al total de los individuos, que podemos calcular a partir de la ecuación (6.37)

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{0{,}494 - 0{,}470}{0{,}494} = 0{,}0486 \tag{6.47}$$

En este caso, el índice de fijación es bastante menor. De acuerdo con las guías prácticas dadas por Wright (Sewall Wright 1978), se trata de un valor (apenas) menor a 0,05 y, por lo tanto, indicativo de poca diferenciación genética **entre** poblaciones.

El índice de fijación entre el nivel más y menos inclusivo está dado por la ecuación (6.38), es decir

$$F_{IT} = \frac{H_T - H_I}{H_T} = \frac{0{,}494 - 0{,}222}{0{,}494} = 0{,}551 \tag{6.48}$$

Para verificación, usando la ecuación (6.42), tenemos que

$$(1-F_{IT}) = (1-F_{ST})(1-F_{IS}) = (1-0{,}0486)(1-0{,}528) = 0{,}449 \Leftrightarrow F_{IT} = 1-0{,}449 = 0{,}551 \tag{6.49}$$

que es igual al valor que obtuvimos directamente. En nuestro caso, entonces, se trata de un **ensemble** de poblaciones en que la mayor parte del déficit de heterocigotos que hemos encontrado se debe un apareamiento no-aleatorio **dentro** de las poblaciones que lo constituyen y en mucho menor medida debido a la diferenciación genética **entre** poblaciones.

Una manera alternativa de pensar en F_{ST} es en términos de partición de la varianza en frecuencias de los alelos. En particular, si pensamos que cada población en un conjunto grande de poblaciones formando el **ensemble** tiene sus propias frecuencias p_i y q_i, la varianza en la frecuencia de p, $\sigma^2_{\mathbf{p}}$, estará dada por

$$\sigma^2_{\mathbf{p}} = \frac{1}{n}\sum_{i=1}^{i=n}(p_i - \bar{p})^2 \tag{6.50}$$

Lo mismo para q, cuya varianza será igual a la de p y, por lo tanto, la varianza en frecuencias igual a $\sigma^2_{\mathbf{p}}$. Para comparar esta varianza con la del nivel más inclusivo la dividimos entre $2\bar{p}\bar{q} = \bar{p}(1-\bar{p})$ y, por lo tanto,

$$F_{ST} = \frac{2\sigma^2_{\mathbf{p}}}{2\bar{p}(1-\bar{p})} = \frac{\sigma^2_{\mathbf{p}}}{\bar{p}(1-\bar{p})} \tag{6.51}$$

La derivación formal de este resultado lo veremos en la sección «El efecto Wahlund».
Pese a su importancia, los estadísticos F han estado sujetos a diferentes interpretaciones, sumado a diferentes formas de calcularlos y a partir de fuentes de información diferente, lo que ha generado no poca confusión. Una revisión importante del tema, con un enfoque en lo que pueden aportar diferentes partes del genoma se encuentra en la revisión de Holsinger y Weir (2009).

PARA RECORDAR

- Los individuos suelen pertenecer a poblaciones que pertenecen a estructuras jerárquicas, que se definen de manera algo arbitraria. Por ejemplo, individuos, dentro de poblaciones, que a su vez están dentro de un **ensemble** de poblaciones. La estructuración en grupos menores suele implicar una reducción en la heterocigosidad respecto a las condiciones de **panmixia**.
- Definimos tres estadísticos para heterocigosidad, que usaremos para calcular los estadísticos F. Para n poblaciones la **heterocigosidad media observada** es $H_I = \frac{1}{n}\sum_{i=1}^{n}\hat{H}_i$.
- A su vez, la **heterocigosidad media esperada dentro** de las poblaciones la calculamos como $H_S = \frac{1}{n}\sum_{i=1}^{n} 2p_iq_i$.
- Finalmente, la **heterocigosidad esperada global** es igual a $H_T = 2\bar{p}\bar{q}$.
- F_{IS}: es la correlación de alelos dentro de un individuo en relación con la población en la que se encuentra. A su vez, es la desviación media de las frecuencias de los genotipos con respecto a las expectativas de Hardy-Weinberg dentro de las poblaciones. $F_{IS} = \frac{H_S - H_I}{H_S}$.
- F_{ST}: Correlación de alelos elegidos al azar dentro de la misma población en relación con el **ensemble** de poblaciones; equivalentemente, la proporción de la diversidad genética debida a las diferencias de frecuencia alélica entre poblaciones. $F_{ST} = \frac{(F_{IT} - F_{IS})}{(1 - F_{IS})}$.
- F_{IT}: Correlación de alelos dentro de un individuo en relación con toda la población. A su vez, es la desviación de las frecuencias de los genotipos con respecto a las expectativas de Hardy-Weinberg para todo el **ensemble** de poblaciones. $F_{IT} = \frac{H_T - H_I}{H_T}$.
- f_X: Correlación de los alelos dentro de un individuo en relación con la subpoblación en la que se encuentra; equivalente a F_{IS}.

6.7 El efecto Wahlund

Hasta ahora hemos visto el fenómeno de la estructuración en poblaciones desde el lado de la pérdida de heterocigosis respecto a lo esperado bajo régimen de panmixia. Sin embargo, el fenómeno puede verse desde un lado en cierta forma opuesto, o complementario. Cuando dos o más poblaciones que se encontraban totalmente aisladas pasan a mezclarse y a reproducirse en condiciones de **panmixia**, entonces podemos ver el fenómeno opuesto: el exceso de heterocigotos respecto a la situación anterior (que era equivalente a juntar todos los individuos previo a la reproducción). Obviamente, como las frecuencias genotípicas deben sumar a 1, entonces el aumento de los heterocigotos debe ir acompañado de una reducción igual en los homocigotos. Este fenómeno se conoce como **efecto Wahlund** en honor del genetista sueco Sten Wahlund,[5] que lo describió por primera vez (Wahlund 1928).

Para hacer gráfico este fenómeno, pensemos en la situación descrita en la figura 6.5. Tenemos dos poblaciones de ovejas con diferentes alelos de color fijados en cada una y las mismas se encuentran aisladas por una barrera infranqueable, que lo constituye el **río profundo y ancho**. Este aislamiento nos lleva a que, si muestreamos ovejas en la ribera del río, pese a encontrar una frecuencia alélica intermedia para el locus (ficticio) del color no vamos a encontrar ningún heterocigoto. Supongamos ahora que en beneficio de los humanos de la zona se construye un puente sobre el río, como lo muestra la figura 6.6, de tal forma que las ovejas pueden cruzar libremente de un lado para el otro. Suponiendo que las mismas no se ven demasiado cohibidas por lo extraño de sus congéneres al otro lado del río, el próximo apareamiento será en condiciones de panmixia y como el número de animales homocigotos blancos y homocigotos verdes es igual, la frecuencia del alelo blanco será $\frac{1}{2}$, igual que la del verde y, por lo tanto, bajo equilibrio de Hardy-Weinberg esperamos ahora que la frecuencia de los heterocigotos sea de $2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$.

De hecho, la frecuencia de heterocigotos en el apareamiento bajo panmixia de la población mezcla siempre será mayor al correspondiente del promedio de las poblaciones separadas, a menos que las frecuencias alélicas sean idénticas en las dos poblaciones. Para demostrar esto, asumiendo que ambas poblaciones tienen igual tamaño, alcanza con probar que

$$2\bar{p}\bar{q} = 2\frac{p_1+p_2}{2}\left[1-\frac{p_1+p_2}{2}\right] \geqslant \frac{2p_1q_1+2p_2q_2}{2} = 2\frac{p_1(1-p_1)+p_2(1-p_2)}{2} \;\therefore \quad (6.52)$$

$$(p_1+p_2)\left[1-\frac{p_1+p_2}{2}\right] \geqslant p_1(1-p_1)+p_2(1-p_2) \;\therefore \quad (6.53)$$

$$(p_1+p_2)(2-p_1-p_2) \geqslant 2(p_1-p_1^2+p_2-p_2^2) \;\therefore \quad (6.54)$$

$$2p_1+2p_2-p_1^2-p_1p_2-p_1p_2-p_2^2 \geqslant 2p_1-2p_1^2+2p_2-2p_2^2 \;\therefore \quad (6.55)$$

$$p_1^2+p_2^2-2p_1p_2 \geqslant 0 \quad (6.56)$$

[5] Sten Gösta William Wahlund (6 marzo de 1901 - 6 julio de 1976) fue un genetista y político sueco que introdujo el concepto del cambio de frecuencias genotípicas bajo el efecto de la panmixia en poblaciones que se encontraban separadas.

Pero $p_1^2 + p_2^2 - 2p_1p_2 = p_1^2 - 2p_1p_2 + p_2^2 = (p_1 - p_2)^2$, por lo que la inecuación (6.56) se reduce a

$$(p_1 - p_2)^2 \geqslant 0 \tag{6.57}$$

lo que claramente es cierto porque el cuadrado de cualquier número real es mayor o igual a cero, lo que prueba nuestra conjetura de que habrá tantos o más heterocigotos en condiciones de panmixia que cuando consideramos las poblaciones separadas. Más aún, el único caso en que habrá las mismas frecuencias en ambos casos (antes y después de la panmixia) es cuando $p_1 = p_2$, ya que en ese caso $(p_1 - p_2)^2 = (p_1 - p_1)^2 = (0)^2 = 0$.

De hecho, los términos del lado izquierdo de la primera inecuación (6.56) representan la situación **luego** de la **panmixia**, mientras que los de la derecha representan la situación antes, por lo que si le restamos a los términos de la izquierda los de la derecha tendremos el Δ_H, es decir el cambio en la frecuencia de heterocigotos por el efecto Wahlund. Esto es lo que hicimos en la última línea de las ecuaciones (6.56), excepto que en la tercera línea multiplicamos ambos lados de la inecuación por 2, así que ahora debemos multiplicar por $\frac{1}{2}$. Por lo tanto, para dos poblaciones, el cambio de frecuencias viene dado por

$$\Delta_H = \frac{1}{2}(p_1^2 + p_2^2 - 2p_1p_2) = \frac{1}{2}(p_1 - p_2)^2 \tag{6.58}$$

A mayor diferencia en las frecuencias de las poblaciones, mayor será el efecto de unión y **panmixia**.

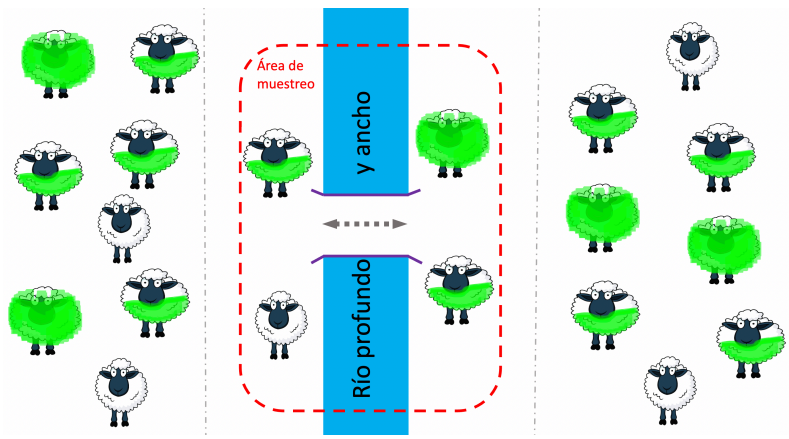


Figura 6.6. Efecto de Wahlund al establecerse una conexión masiva entre dos poblaciones completamente diferenciadas. Si bien previo a la existencia del puente ambas poblaciones se encontraban en equilibrio de Hardy-Weinberg, al estar fijados los alelos (diferentes) en ambas, no existían heterocigotos. Al aparecer la posibilidad de panmixia, las frecuencias alélicas se mantienen, pero ahora la mitad de las ovejas serán heterocigotas (diseño de las ovejas de http://cliparts.co/clipart/2377087 Fuente: Cliparts.co).

Este fenómeno también lo podemos interpretar a través de los diagramas de de Finetti, que ya vimos en el capítulo «Variación y equilibrio de Hardy-Weinberg». En efecto, como vimos antes y como se ve en la figura 6.7, dentro del triángulo que define el espacio posible de frecuencias genotípicas, el equilibrio de Hardy-Weinberg estará determinado por la curva que va desde el vértice aa hasta el vértice AA. Cualquier población que se encuentre en equilibrio de Hardy-Weinberg deberá estar sobre esta curva. Supongamos ahora que tenemos nuestras dos poblaciones separadas, con frecuencias del alelo dominante (este detalle no importa realmente) p_1 y p_2. Dadas estas frecuencias y al saber que dentro de ellas se cumplen las condiciones para el equilibrio de Hardy-Weinberg implica que basta cortar la curva con una línea perpendicular a la arista entre aa y AA por el punto p_i para determinar las frecuencias genotípicas en cada población antes del evento de panmixia. Para cualquier combinación lineal de las dos poblaciones, pero en particular para la mezcla en iguales proporciones, la frecuencia del conjunto antes de la reproducción será $\bar{p}$ y estará ubicada en $p_1 \leqslant \bar{p} \leqslant p_2$, mientras que las proporciones genotípicas estarán dadas por el corte de la perpendicular en $\bar{p}$ con el segmento de recta entre 1 y 2. Claramente, para cualquier par de puntos sobre la curva o, dicho de otra forma para cualquier par de frecuencias p_1 y p_2, el corte de la perpendicular por $\bar{p}$ con el segmento entre 1 y 2 estará siempre por debajo del corte por la misma perpendicular con la curva de equilibrio Hardy-Weinberg (por ejemplo, marcado como y en la figura). Como esas distancias representan en el diagrama de de Finetti la proporción de heterocigotos, entonces, cuando $p_1 \neq p_2$, el promedio de las poblaciones separadas tendrá siempre menos heterocigotos que la población mezclada y en panmixia.

La importancia del efecto Wahlund está directamente asociada al papel que jueguen los heterocigotos en las características de interés. Por ejemplo, en características donde el alelo recesivo es deletéreo, los heterocigotos tenderán a tener un fenotipo «normal» y, por lo tanto, la mezcla de poblaciones diferentes llevará a un descenso en el número de individuos con la patología. En seres humanos son comunes los ejemplos de enfermedades raras hereditarias que disminuyen la frecuencia en forma importante en los recién nacidos producto del mestizaje de poblaciones. Más aún, como veremos más adelante, la exocría (el fenómeno opuesto a la endocría) puede tener grandes ventajas en el mejoramiento genético que son explotadas a través de los cruzamientos y cuya importancia depende tanto de la ventaja de los heterocigotos como de la diferencia de frecuencias entre las poblaciones que se cruzan.

Una forma interesante y diferente de ver este fenómeno es a través de su relación con el índice de fijación a nivel de poblaciones, F_{ST}, que vimos antes. Supongamos que nuevamente estamos con dos poblaciones separadas con el mismo número de individuos en ambas, y que luego de un evento dado se juntan y se aparean bajo panmixia. Supongamos que, sin pérdida de generalidad, el alelo de frecuencia q_i es recesivo y que denotaremos a la frecuencia de los homocigotos recesivos con Q. Al mezclar las dos poblaciones y darse la reproducción bajo panmixia, vamos a observar un decrecimiento de en la frecuencia de homocigotos (tanto recesivos como dominantes) debido al efecto Wahlund. Como ambas poblaciones previo a la mezcla se encontraban en equilibrio Hardy-Weinberg, la frecuencia de homocigotos recesivos previo a la unión será el promedio de los mismos en ambas poblaciones, es decir $\mathrm{Q}_{\mathrm{separadas}} = \frac{q_1^2+q_2^2}{2}$. Luego de la mezcla, bajo condiciones de panmixia, la frecuencia de los homocigotos recesivos será $\mathrm{Q}_{\mathrm{unidas}} = \bar{q}^2$, con $\bar{q} = \frac{q_1+q_2}{2}$.

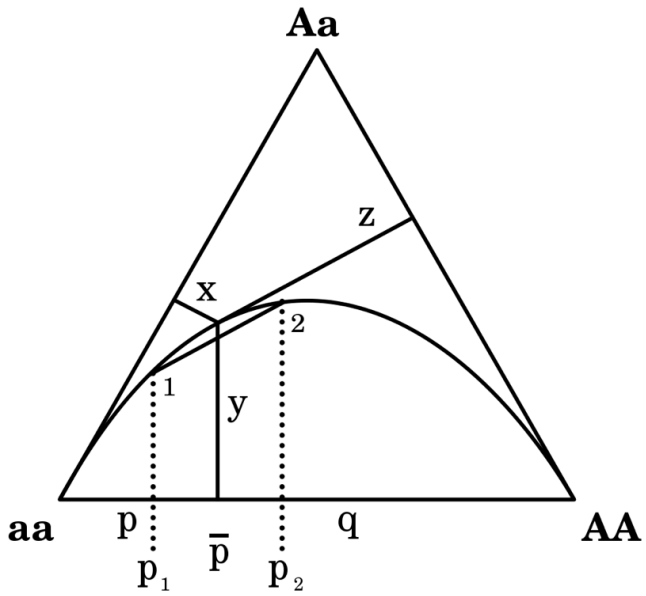


Figura 6.7. Diagrama de de Finetti ilustrando el efecto de Wahlund. Dos poblaciones separadas, con frecuencias p_1 y p_2, ambas en equilibrio de Hardy-Weinberg están por lo tanto sobre la curva que va desde *aa* a *AA* y que marca la región permitida para el equilibrio H-W. La proporción de cada uno de los genotipos al combinarlas (sin apareamiento) será el promedio y, por lo tanto, estará representado por el punto medio en la recta que las une. Sin embargo, si ocurre apareamiento entre ambas el punto estará sobre la curva de equilibrio, que claramente está por encima de la recta siempre y, por lo tanto, la frecuencia de los heterocigotas será mayor. Figura tomada de Wikipedia, CC BY-SA 2.5 (Archivo: De Finetti diagram.svg).

Por lo tanto, la diferencia (es decir, pérdida de homocigotos recesivos) será igual a

$$\begin{aligned} Q_{separadas} - Q_{unidas} &= \frac{q_1^2 + q_2^2}{2} - \bar{q}^2 = \frac{q_1^2 + q_2^2 - 2\bar{q}^2}{2} \Leftrightarrow \\ Q_{separadas} - Q_{unidas} &= \frac{(q_1^2 - \bar{q}^2) + (q_2^2 - \bar{q}^2)}{2} \end{aligned} \tag{6.59}$$

Sin embargo, operando con el numerador vamos a demostrar que $(q_1^2 - \bar{q}^2) + (q_2^2 - \bar{q}^2)$ es igual a $(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2$. Partiendo de esta última expresión pasamos a desarrollarla

$$\begin{aligned} (q_1 - \bar{q})^2 + (q_2 - \bar{q})^2 &= q_1^2 - 2q_1\bar{q} + \bar{q}^2 + q_2^2 - 2q_2\bar{q} + \bar{q}^2 \\ &= q_1^2 + q_2^2 + 2\bar{q}^2 - 2\bar{q}(q_1 + q_2) \end{aligned} \tag{6.60}$$

Pero $\bar{q} = \frac{(q_1+q_2)}{2} \Leftrightarrow (q_1 + q_2) = 2\bar{q}$, por lo que sustituyendo en la ecuación (6.60) tenemos

$$(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2 = q_1^2 + q_2^2 + 2\bar{q}^2 - 2\bar{q}(q_1 + q_2) \ \therefore \tag{6.61}$$

$$(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2 = q_1^2 + q_2^2 + 2\bar{q}^2 - 2\bar{q}(2\bar{q}) = q_1^2 + q_2^2 + 2\bar{q}^2 - 4\bar{q}^2 \ \therefore \tag{6.62}$$

$$(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2 = q_1^2 + q_2^2 - 2\bar{q}^2 = (q_1^2 - \bar{q}^2) + (q_2^2 - \bar{q}^2) \ \therefore \tag{6.63}$$

$$(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2 = (q_1^2 - \bar{q}^2) + (q_2^2 - \bar{q}^2) \tag{6.64}$$

Por lo tanto, conseguimos demostrar que el numerador de (6.59) se puede escribir como $(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2$, por lo que sustituyendo en dicha ecuación llegamos a que

$$Q_{separadas} - Q_{unidas} = \frac{(q_1 - \bar{q})^2 + (q_2 - \bar{q})^2}{2} = (\frac{1}{2}) \times (q_1 - \bar{q})^2 + (\frac{1}{2}) \times (q_2 - \bar{q})^2 \tag{6.65}$$

Claramente, la forma de la ecuación (6.65) es la de una varianza, es decir, la diferencia entre los valores y la media de los mismos, elevada al cuadrado y ponderada por la frecuencia (que en este caso es $\frac{1}{2}$ para cada población). Por lo tanto, podemos llamar a esto la varianza en homocigotos «recesivos» debido al pasaje a la panmixia, es decir

$$Q_{separadas} - Q_{unidas} = \sigma_q^2 \tag{6.66}$$

En palabras, el efecto de Wahlund provoca una disminución de la frecuencia de homocigotas recesivos luego de la mezcla de poblaciones y apareamiento bajo panmixia igual a la varianza de frecuencias alélicas (del alelo recesivo) entre poblaciones. Más aún, este mismo razonamiento lo podemos hacer para el homocigoto en el otro alelo (que notaremos P) y, por lo tanto, por analogía tenemos

$$P_{separadas} - P_{unidas} = \sigma_p^2 \tag{6.67}$$

Pero como $p + q = 1$, entonces $\sigma_p^2 = \sigma_q^2 = \sigma^2$, ya que $\mathbb{V}(x + k) = \mathbb{V}(k - x) = \mathbb{V}(x)$ para cualquier constante k. Como el efecto en la pérdida de homocigotos ahora es la suma de la pérdida en los «recesivos» y en los «dominantes», es decir la reducción total en homocigotos será igual a

$$\text{Red. homocigosis}_{\text{total}} = \sigma * p^2 + \sigma_q^2 = 2\sigma^2 \tag{6.68}$$

Claramente, la reducción en homocigotos tiene que ser acompañada por un incremento igual en los heterocigotos en la panmixia respecto a la situación de las dos poblaciones separadas, $H_T - H_S$. Es decir que

$$\text{Red. homocigosis}_{\text{total}} = 2\sigma^2 = H_T - H_S \tag{6.69}$$

Al mismo tiempo, bajo panmixia $H_T = 2\bar{p}\bar{q}$. Ahora, si vamos a nuestra definición original del coeficiente F_{ST}, ecuación (6.37), tenemos que

$$F_{ST} = \frac{H_T - H_S}{H_T} = \frac{2\sigma^2}{2\bar{p}\bar{q}} = \frac{\sigma^2}{\bar{p}(1 - \bar{p})} \tag{6.70}$$

que es el mismo resultado al que habíamos arribado en la ecuación (6.51). Esto nos deja con una alternativa razonable para la estimación del índice de fijación. Asumiendo tamaños similares de las poblaciones, alcanza con conocer las frecuencias de los alelos en las distintas poblaciones, con lo que podremos calcular tanto σ_p^2 como $\bar{p}$ y, por lo tanto, obtener el estimador

$$\hat{F}_{ST} = \frac{\hat{\sigma}_p^2}{\bar{p}(1 - \bar{p})} \tag{6.71}$$

Ejemplo 6.2

Una población donde segrega un **locus** bialélico A se estructura en un ensemble poblacional, pasando a conformarse por cuatro subpoblaciones de tamaño efectivo $N_e = 200$. Se asume que estas se reproducen bajo régimen de panmixia.

a. En un momento dado, las frecuencias para el alelo A_1 son $p_1 = 0{,}61$, $p_2 = 0{,}77$, $p_3 = 0{,}65$, $p_4 = 0{,}63$. Calcula la pérdida de heterocigosidad en la población total debido al estructuramiento de la misma.

b. ¿Cuántas generaciones son necesarias para que la divergencia genética entre poblaciones pueda ser considerada como muy grande?

Solución

a. Dado que las subpoblaciones tienen igual tamaño, la frecuencia alélica promedio $\bar{p} = \frac{0{,}61+0{,}77+0{,}65+0{,}63}{4} = 0{,}665$; esto implica que la heterocigosidad esperada en la población total es de $H = 2\bar{p}(1 - \bar{p}) = 2 \times 0{,}665 \times (1 - 0{,}665) = 0{,}44555$.

El promedio de heterocigosidad observado, sin embargo, es de $\bar{H} = [2 \times 0{,}61 \times (1-0{,}61) + 2 \times 0{,}77 \times (1-0{,}77) + 2 \times 0{,}65 \times (1-0{,}65) + 2 \times 0{,}63 \times (1-0{,}63)]/4 = 0{,}4378$. Por lo tanto, la estructuración poblacional implica una pérdida de frecuencia de heterocigotas de al momento $\Delta_H = 0{,}44555 - 0{,}4378 = 0{,}00775$ considerado.

b. La varianza para las frecuencias alélicas en un ensemble poblacional crece linealmente según (ver capítulo «Deriva genética»)

$$\sigma_p^2 = t \times \frac{\bar{p}(1-\bar{p})}{2N_e} \tag{6.72}$$

A su vez, la divergencia genética entre poblaciones se puede estimar con

$$\hat{F}_{ST} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})} \tag{6.73}$$

Siguiendo la propuesta de Sewall Wright, podemos considerar que la divergencia genética es muy alta cuando $F_{ST} \gg 0{,}25$. Combinando estos tres elementos, tenemos

$$\begin{aligned}
\hat{F}_{ST} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})} &\gg 0{,}25 \\
\frac{(t \times \frac{\bar{p}(1-\bar{p})}{2N_e})}{\bar{p}(1-\bar{p})} &\gg 0{,}25 \\
t \times \frac{\bar{p}(1-\bar{p})}{2N_e \times \bar{p}(1-\bar{p})} &\gg 0{,}25 \\
t \times \frac{1}{2N_e} &\gg 0{,}25 \\
t &\gg 2 \times N_e \times 0{,}25 \\
t &\gg 2 \times 200 \times 0{,}25 \\
t &\gg 100
\end{aligned} \tag{6.74}$$

Por lo tanto, solo considerando la acción de la deriva genética como fuente de varianza de frecuencias alélicas en el ensemble, se espera que luego de unas cien generaciones estas poblaciones posean una divergencia genética muy alta para el **locus** considerado.

PARA RECORDAR

- Se conoce como efecto Wahlund al fenómeno del incremento de la frecuencia de heterocigotas luego de la unión de las mismas y la reproducción bajo condiciones de **panmixia**, respecto al promedio de dos poblaciones antes de la unión.

- Para dos poblaciones, en un locus diploide con dos alelos, con frecuencias del alelo A_1 p_1 y p_2, ambas poblaciones en equilibrio Hardy-Weinberg, el aumento en la frecuencia de heterocigotas, Δ_H, será

$$\Delta_H = \frac{1}{2}(p_1^2 + p_2^2 - 2p_1p_2) = \frac{1}{2}(p_1 - p_2)^2 \tag{6.75}$$

- El efecto Wahlund nos permite obtener un estimador de la diferenciación entre poblaciones que consiste en dividir la varianza en frecuencias de un alelo en las poblaciones ($\hat{\sigma}_p^2$) entre el producto de la media de frecuencias y su complementario a 1 ($\bar{p}\bar{q} = \bar{p}(1 - \bar{p})$), es decir

$$\hat{F}_{ST} = \frac{\hat{\sigma}_p^2}{\bar{p}(1 - \bar{p})} \tag{6.76}$$

6.8 Subdivisión, migración y el modelo de islas

En las primeras secciones de este capítulo nuestro enfoque fue el de considerar el fenómeno del apareamiento no-aleatorio desde la perspectiva de las modificaciones de las frecuencias genotípicas a la interna de las poblaciones y en el surgimiento de una fuente importante de variación en estas frecuencias a partir de la estructuración en unidades aisladas, las poblaciones. Notablemente, analizamos el déficit en heterocigotos resultante del apareamiento no-aleatorio, tanto a nivel interno de las poblaciones como el que surge de la estructuración en poblaciones aisladas. En la sección «El efecto Wahlund» vimos el fenómeno del apareamiento no-aleatorio desde una perspectiva opuesta: cómo el apareamiento al azar de una mezcla de poblaciones nos llevaba a destruir esa estructuración previa y la vuelta al nivel de heterocigotos esperado, incrementando su frecuencia y disminuyendo consecuentemente la de los homocigotos. La sección presente está basada en el enfoque de Hartl y Clark (2007), que al mismo tiempo de mantenerse simple expone los principales conceptos con claridad. Un tratamiento bastante claro, pero mucho más extenso, con otros varios modelos considerados, aparece en el texto de Hamilton (2009), que recomendamos para quien quiera profundizar pero manteniendo la simplicidad matemática.

Claramente, la **mezcla total** de poblaciones es un evento relativamente infrecuente respecto a otras formas de **comunicación genética** entre poblaciones. Mucho más frecuentes son los procesos de migración de individuos entre poblaciones. A nivel de poblaciones humanas, todos conocemos los grandes procesos migratorios, entre ellos los relativamente recientes (a escala evolutiva) que permitieron la constitución actual de nuestro continente. En poblaciones animales de interés comercial, por ejemplo, el «mejoramiento» del ganado criollo existente hacia las distintas razas bovinas mediante la **migración** y apareamiento con individuos provenientes de poblaciones europeas es un fenómeno patente. De hecho, en aquellas especies en las que la **sustitución** de una población por otra resulta impracticable, la introgresión de material genético «deseable» suele ocurrir a partir de un mecanismo que es perfectamente asimilable a la migración de individuos de una población a otra (unidireccional, en estos casos). En general, tanto en poblaciones humanas como de otras especies, el patrón de flujo genético suele ser bastante más complicado que el envío de unos cuantos migrantes de una población hacia otra. Por ejemplo, en la historia de nuestro continente aún se especula sobre los tiempos de entrada de los primeros pobladores y si se trató de

uno o varios eventos independientes. Más aún, la llegada de los conquistadores europeos introdujo una importante asimetría en los apareamientos ya que por decenios solo llegaron hombres al continente, que procreaban con mujeres nativas, lo que produjo un fuerte sesgo en los cromosomas **Y** de los mestizos, así como en el **ADN mitocondrial** de todos los descendientes (ya que dicho material genético se hereda casi exclusivamente por vía materna). Esto nos lleva a su vez a que distintos tipos de marcadores moleculares cuentan partes distintas de la historia, y a que en general debamos interpretar con mucha precaución los datos genéticos ya que es difícil sintetizar historias complejas en unos pocos estimadores.

Una forma sencilla de entrar al tema del efecto de las migraciones en las frecuencias alélicas es a través del más sencillo de los modelos, como lo ilustra la figura 6.8.

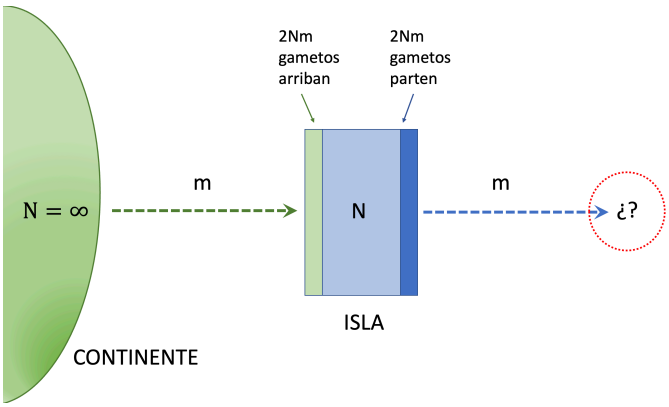


Figura 6.8. Modelo de isla para estudiar el efecto de la migración en el cambio de frecuencias alélicas. El continente posee una población de tamaño infinito, mientras que en la isla la población es de tamaño N. En cada generación una proporción m de los individuos de la isla es reemplazada por migrantes que vienen del continente y Nm individuos parten de la isla hacia lugar desconocido.

El centro de este modelo es una isla con una población de tamaño constante N que a cada generación recibe una proporción m de migrantes desde el continente, que asumimos de tamaño poblacional infinito (claramente, alcanza con que sea muy grande, pero el supuesto de infinito simplifica razonamientos). Obviamente, como el tamaño de la población de la isla es constante, el arribo de Nm inmigrantes desde el continente debe significar la partida de Nm (emigrantes) hacia algún lado (desconocido e irrelevante para nuestro modelo).

Una analogía física de este modelo se puede ver en la figura 6.9.

Un tanque muy grande con colorante azul (que en nuestra analogía hace el papel de **continente**) rebosa hacia un tanque más pequeño (también podría ser una canilla que pierde), debajo de este (que hace el papel de **isla**). El tanque que está debajo esta **inicialmente** lleno de colorante rojo. A medida que las gotas de colorante azul caen las mismas se mezclan con el colorante rojo del tanque inferior, por lo que el rebose de este tanque (que mantiene el nivel por el desagüe de la derecha) va cambiando de color: inicialmente rojo, luego violeta y yendo hacia el azul a medida la mayoría del colorante inicial ha sido reemplazado por el colorante azul que cae. Notemos que no hay una entrada de colorante rojo, mientras que el aporte de colorante azul a este tanque es infinito.

Volvamos al modelo ilustrado en la figura 6.8. Consideremos que estamos tratando nuevamente con un gen con dos alelos, A_1 y A_2 en un organismo diploide. Siguiendo la notación

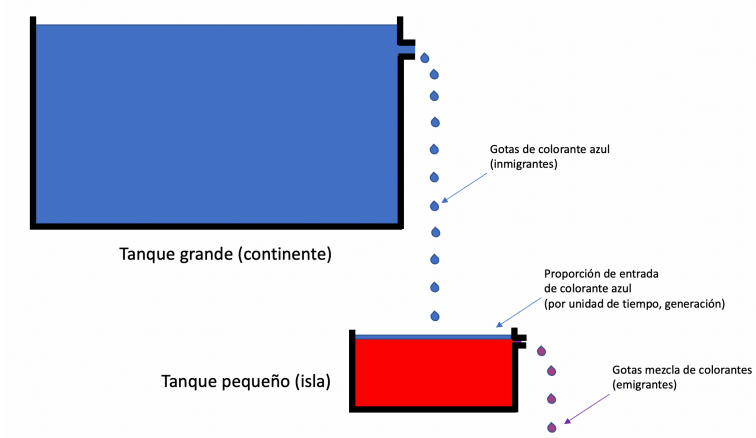


Figura 6.9. Analogía física para el modelo de continente-isla. Un tanque grande de colorante azul (análogo a la población en el continente) tiene un goteo permanente que cae en un tanque pequeño (isla) de colorante rojo. La salida del tanque pequeño será de la mezcla entre los dos colorantes (que asumimos instantánea), por lo tanto, de color rojo al principio, luego violeta y yendo hacia el azul a largo plazo.

de Hartl y Clark (2007), llamaremos p a la frecuencia actual de A_1 **en la isla**, mientras que usaremos $p^\star$ para la frecuencia del mismo alelo en el continente. Claramente, la frecuencia de A_1 en la siguiente generación **en la isla**, que denotaremos p', será igual a p por la fracción de individuos de la isla que no emigran, más la fracción de inmigrantes por $p^\star$, es decir

$$p' = (1-m)p + mp^\star \tag{6.77}$$

Si ahora le restamos $p^\star$ a ambos lados de la ecuación (6.77) y operamos, tenemos que

$$p' - p^\star = (1-m)p + mp^\star - p^\star = (1-m)p + p^\star(m-1) = (1-m)p - p^\star(1-m) \;\therefore$$
$$p' - p^\star = (1-m)(p-p^\star) \tag{6.78}$$

Claramente, la ecuación (6.78) se trata de una recurrencia, donde calculamos el valor de cada generación como el producto del término de interés en la generación anterior $(p-p^\star)$ por un factor constante, en este caso $(1-m)$. Cada generación hacia atrás volvemos a multiplicar por este factor constante. Por lo tanto, considerando t generaciones hacia atrás este término se transformará en $(1-m)^t$, multiplicado por la diferencia de frecuencias original entre la población de la isla p_0 y la del continente, $p^\star$ (es decir, $(p_0 - p^\star)$). Poniendo todo junto, tenemos ahora

$$p_t - p^\star = (1-m)^t(p_0 - p^\star) \Leftrightarrow \tag{6.79}$$

$$p_t = p^\star + (1-m)^t(p_0 - p^\star) \tag{6.80}$$

El resultado de la ecuación (6.80) es muy claro y ya hemos visto ecuaciones de forma similar. Se trata de la suma de un término constante, que es la frecuencia del alelo A_1 en la población del continente, $p^\star$, más una función de la diferencia entre el valor inicial de la isla p_0 respecto al continente. En función de que el número de generaciones t entra como exponente de una constante $0 \leqslant (1-m) < 1$ que multiplica a la diferencia de frecuencias iniciales, la forma del decaimiento del segundo término de la derecha en la ecuación (6.80) será geométrica, o aproximadamente exponencial. Es decir, la frecuencia del alelo A_1 en la isla se aproximará rápidamente al valor del continente, disminuyendo la diferenciación genética entre ambas poblaciones.

El modelo anterior, unidireccional, es extremadamente simplista y atiende unas pocas situaciones. Un modelo algo más general es el **modelo de islas**, que aparece representado en la figura 6.10. En este modelo todas las poblaciones (4 en nuestro ejemplo) poseen igual tamaño y aparecen totalmente conectadas en el intercambio de individuos (o gametos). De hecho, cada una de las d poblaciones (o *demes*, con $d = 4$ en nuestro ejemplo) intercambia en cada generación una proporción m/d con cada una de las otras (envía y recibe), incluida ella misma (simbolizado por las flechas recurrentes en color azul). Esto es equivalente a pensar que, en cada generación, cada una de las poblaciones aporta una proporción m de sus gametos a un **pool central** (marcado por el círculo rojo en el centro) y luego recibe de ese **pool** exactamente la misma proporción m de gametos, pero ahora **elegidos al azar del pool.**

Como el **pool central** está formado por muestreos representativos de las distintas poblaciones, la frecuencia del alelo A_1 en el mismo será $\bar{p}$ (el promedio de todas las poblaciones), y como además todos los gametos vuelven a alguna población, la media no cambiará en el tiempo. Supongamos ahora que atendemos la frecuencia del alelo A_1 en una cualquiera de las poblaciones en una generación t dada. La probabilidad de que ese alelo A_1 venga de la misma población en la generación anterior es el producto de la frecuencia de A_1 en la generación por la probabilidad de no haya migrado, que es $(1-m)$, más la probabilidad de que sea un migrante (que fue al pool y volvió) multiplicado por la frecuencia en el **pool**, que es $\bar{p}$. En conjunto, esto implica por lo tanto que

$$p_t - \bar{p} = (1-m)(p_{(t-1)} - \bar{p}) \tag{6.81}$$

De hecho, nosotros elegimos la generación t como referencia, pero podríamos haber elegido la generación $(t-1)$, por lo que, cambiando de símbolos, también podemos decir que

$$p_{(t-1)} - \bar{p} = (1-m)(p_{(t-2)} - \bar{p}) \tag{6.82}$$

En función de esto, sustituyendo el $(p_{(t-1)} - \bar{p})$ de la ecuación (6.81) por el lado derecho de la ecuación (6.82), tenemos

$$p_t - \bar{p} = (1-m)(1-m)(p_{(t-2)} - \bar{p}) = (1-m)^2(p_{(t-2)} - \bar{p}) \tag{6.83}$$

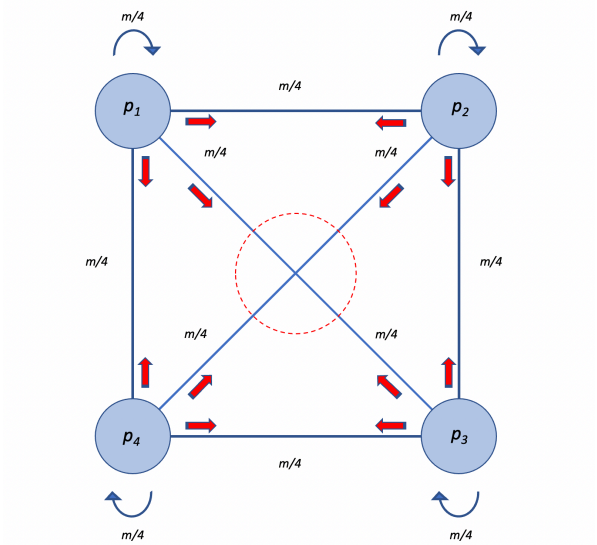


Figura 6.10. Modelo de islas, con **4** poblaciones. Las frecuencias del alelo A_1 de cada una aparece dentro de los círculos correspondientes. Cada población intercambia con las otras exactamente la misma cantidad de migrantes, en nuestro caso $m/4$ (incluyéndose ella misma), por lo que el modelo es equivalente a que cada población aporte una proporción m de gametos a un **pool** central y luego reciba del **pool** esa misma proporción m elegidos al azar. Elaboración propia sobre idea en Hartl y Clark (2007).

Sucesivamente, si seguimos t generaciones hacia atrás en esta recursión llegamos, obviamente, a que

$$p_t - \bar{p} = (1-m)^t(p_0 - \bar{p}) \Leftrightarrow p_t = \bar{p} + (1-m)^t(p_0 - \bar{p}) \tag{6.84}$$

Donde p_0 es la frecuencia inicial del alelo A_1 en la población particular que estamos considerando (alguna de las $d = 4$ *demes* que elegimos para estudiar, en nuestro caso). La ecuación (6.84) tiene una forma análoga a la de la ecuación (6.80), en el modelo isla-continente. La diferencia fundamental es que mientras que en el **modelo isla-continente** nos acercamos generación a generación a $p^\star$ (frecuencia del alelo A_1 en el continente); en el **modelo de islas** nos acercamos en cada **isla** (población o *deme*) a $\bar{p}$ (frecuencia media del alelo A_1 en el **ensemble** de poblaciones). En razón de la similitud, el comportamiento cualitativo de la reducción en la diferenciación será similar entre los dos modelos.

Que las migraciones contribuyen a romper la diferenciación genética entre poblaciones queda ya muy claro. La pregunta relevante es cómo cuantificamos este efecto. Como vimos previamente en «Homocigosidad y heterocigosidad», la recursión para obtener el valor de F en la generación t a partir del valor en la generación anterior $(t-1)$ dentro de una población cualquiera (por ejemplo, la de la isla), era

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{(t-1)} \tag{6.85}$$

como aparece en la figura 3.10. Es decir, la probabilidad de que sean **IBD** es la probabilidad de que sean idénticos en esa generación ($\frac{1}{2N}$), más la probabilidad de que sean distintos alelos muestreados en esta generación pero que esos alelos fueran idénticos en la anterior ($\left(1 - \frac{1}{2N}\right) F_{(t-1)}$). Consideremos ahora la situación con migración, asumiendo que podemos reemplazar F por F_{ST} llegado el momento (ya veremos por qué). En cualquiera de los dos casos que representan los términos de la derecha de la ecuación (6.85), los gametos será idénticos si ninguno de los dos alelos es reemplazado por un alelo inmigrante, lo que ocurre con probabilidad $(1-m)$ cada evento independiente, es decir $(1-m)(1-m) = (1-m)^2$. Por lo tanto, agregando este factor a cada uno de los eventos de la derecha, la ecuación (6.85) se transforma en

$$F_t = \left(\frac{1}{2N}\right)(1-m)^2 + \left(1 - \frac{1}{2N}\right)(1-m)^2 F_{(t-1)} \tag{6.86}$$

En el equilibrio, es decir cuando no cambie la diferenciación debida a la entrada de inmigrantes a la isla, F no cambiará más en el tiempo, por lo que $F_t = F_{(t-1)} = \hat{F}$. Sustituyendo F_t y $F_{(t-1)}$ por $\hat{F}$ en la ecuación (6.86) nos deja con la siguiente igualdad

$$\hat{F} = \left(\frac{1}{2N}\right)(1-m)^2 + \left(1 - \frac{1}{2N}\right)(1-m)^2 \hat{F} \tag{6.87}$$

Expandiendo los términos en $(1-m)^2$ y operando llegamos a que

$$\begin{aligned} \hat{F} &= \left(\frac{1}{2N}\right)(1 - 2m + m^2) + \left(1 - \frac{1}{2N}\right)(1 - 2m + m^2)\hat{F} \ \therefore \\ \hat{F} &= \frac{1}{2N} - \frac{2m}{2N} + \frac{m^2}{2N} + \hat{F} - 2m\hat{F} + m^2\hat{F} - \frac{1}{2N}\hat{F} + \frac{2m}{2N}\hat{F} - \frac{m^2}{2N}\hat{F} \ \therefore \end{aligned} \tag{6.88}$$

En general m es relativamente pequeño, por lo que m^2 es muy pequeño (el cuadrado de un número entre 0 y 1 es menor que el número mismo) y $m \ll N$, por lo que podemos despreciar los factores de orden m^2 y $\frac{m}{N}$ y la ecuación (6.88) queda aproximada según

$$\hat{F} \approx \frac{1}{2N} + \hat{F} - 2m\hat{F} - \frac{1}{2N}\hat{F} \ \therefore \tag{6.89}$$

$$\hat{F} - \hat{F} + 2m\hat{F} + \frac{1}{2N}\hat{F} \approx \frac{1}{2N} \ \therefore \tag{6.90}$$

$$2m\hat{F} + \frac{1}{2N}\hat{F} \approx \frac{1}{2N} \;\; \therefore \tag{6.91}$$

$$\hat{F}\left(2m + \frac{1}{2N}\right) \approx \frac{1}{2N} \;\; \therefore \tag{6.92}$$

$$\hat{F}\frac{1+4Nm}{2N} \approx \frac{1}{2N} \;\; \therefore \tag{6.93}$$

$$\hat{F}(1+4Nm) \approx 1 \tag{6.94}$$

Si bien $\hat{F}$ es el estimador de F_{IT}, asumiendo **panmixia** dentro de las poblaciones tenemos que $F_{IS} = 0$, y entonces

$$\begin{aligned}(1 - F_{IT}) = (1 - F_{ST})(1 - F_{IS}) = (1 - F_{ST})(1 - 0) = (1 - F_{ST}) \Leftrightarrow \\ (1 - F_{IT}) = (1 - F_{ST}) \Leftrightarrow F_{IT} = F_{ST}\end{aligned} \tag{6.95}$$

Por lo tanto, bajo **panmixia** dentro de las poblaciones, $F_{IT} = F_{ST}$ y, por lo tanto, $\hat{F}$ se transforma en un estimador de $\hat{F}_{ST}$, es decir

$$\hat{F}_{ST} \approx \frac{1}{1+4Nm} \tag{6.96}$$

La importancia de la ecuación (6.96) es evidente: la diferenciación genética entre poblaciones en el modelo de islas depende en forma inversa de $1 + 4Nm$. Recordemos que en nuestro modelo de isla-continente Nm es el número de inmigrantes que arriban a la isla en cada generación, y $\hat{F}$ solo depende de su producto, por lo que considerarlos en conjunto nos permite una interpretación independiente de todas las combinaciones de valores. Claramente, al depender en forma inversa de Nm, cuanto mayor la cantidad de inmigrantes por generación, menor será la diferenciación genética esperada. La figura 6.11 representa los valores de equilibrio esperados de acuerdo con el número de migrantes por generación. Claramente, el decaimiento de $\hat{F}_{ST}$ es muy rápido, y ya con solo un migrante por generación alcanza para mantener un valor de $\hat{F}_{ST} = 0{,}2$ (línea verde). Alcanza con 5 migrantes por generación para que el valor de $F_{ST} = 0{,}05$ (línea azul), que de acuerdo con las guías prácticas dadas por Wright para la interpretación de F_{ST} indicaría que por debajo de ese valor casi no existe diferenciación genética.

Ejemplo 6.3

Siete islas forman parte de un archipiélago, en donde viven subpoblaciones de tamaños similares de una especie de ave. Un estudio determina las frecuencias alélicas (p_i) para un **locus** marcador en varias de estas subpoblaciones, arrojando los siguientes valores: $p_1 = 0{,}55$, $p_2 = 0{,}68$, $p_3 = 0{,}70$, $p_4 = 0{,}61$, $p_5 = 0{,}72$, $p_6 = 0{,}58$, $p_7 = 0{,}82$.

¿El flujo de individuos entre las islas influye notablemente en el grado de diferenciación genética entre poblaciones de las islas?

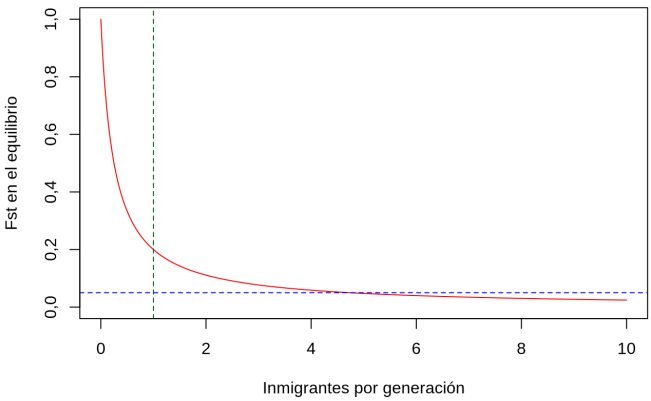


Figura 6.11. Valores de F_{ST} en el equilibrio en función del número de inmigrantes por generación. La línea verde representa el valor esperado de diferenciación genética en el equilibrio con un solo migrante por generación, que es de 0,2. La línea azul representa el valor $F_{ST} = 0{,}05$ que de acuerdo con las guías prácticas dadas por Wright representan situaciones donde casi no existe diferenciación genética.

Solución

Asumiendo que se cumple el modelo de islas, se tiene que el número de migrantes promedio por generación entre las mismas equivale a $\hat{F}_{ST} = \frac{1}{1+4Nm}$.

Dadas las frecuencias alélicas, podemos estimar $\hat{F}_{ST}$ según $\hat{F}_{ST} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})}$.

Con las frecuencias alélicas dadas, se tiene que $\bar{p} = \frac{0{,}55+0{,}68+0{,}70+0{,}61+0{,}72+0{,}58+0{,}82}{7} \approx 0{,}67$ (donde se asume que las subpoblaciones muestreadas tienen tamaños similares). Luego, se tiene que $\sigma_p^2 = \frac{1}{7}\sum_{i=1}^{i=7}(p_i - \bar{p})^2 = 0{,}0086619$. Esto implica $\hat{F}_{ST} \approx \frac{0{,}0086619}{0{,}67\times(1-0{,}67)} \approx 0{,}039$, por lo que siguiendo las guías prácticas propuestas por Sewall Wright nuestras estimaciones sugieren un bajo grado de diversificación genética entre las subpoblaciones.

Combinando ambas ecuaciones, podemos estimar el número de migrantes que podría estar involucrado en mantener a las poblaciones homogéneas a nivel genético

$$\hat{F}_{ST} = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})} = \frac{1}{1+4Nm} \Rightarrow 1 + 4Nm = \frac{\bar{p}(1-\bar{p})}{\sigma_p^2} \tag{6.97}$$

$$4Nm = \frac{\bar{p}(1-\bar{p})}{\sigma_p^2} - 1 \Rightarrow Nm = \frac{1}{4}\left(\frac{\bar{p}(1-\bar{p})}{\sigma_p^2} - 1\right) \tag{6.98}$$

$$Nm = \frac{1}{4}\times\left(\frac{\left(\left(\frac{0{,}55+0{,}68+0{,}70+0{,}61+0{,}72+0{,}58+0{,}82}{7}\right)\right)\left(1-\left(\frac{0{,}55+0{,}68+0{,}70+0{,}61+0{,}72+0{,}58+0{,}82}{7}\right)\right)}{0{,}0086619} - 1\right) \tag{6.99}$$

$$N_m \approx 6{,}2 \tag{6.100}$$

PARA RECORDAR

- En el modelo de isla-continente el flujo de información genética es unidireccional, del continente a la isla, a través de una proporción de m migrantes por generación. La frecuencia del alelo A_1 en la isla, al cabo de t generaciones estará dado por $p_t = p^\star + (1-m)^t(p_0 - p^\star)$, con p_0 la frecuencia inicial **en la isla** y $p^\star$ la frecuencia en el continente (que no cambia con las generaciones ya que su tamaño es infinito y además no recibe otro influjo).
- En el modelo de islas, el flujo de información genética es simétrico entre las d islas que constituyen el **ensemble** de poblaciones (incluyendo la propia isla). Esto es equivalente a que en cada generación todas las islas envían una proporción m de sus gametos a un **pool central** y que luego reciben esa misma proporción de gametos (pero de la mezcla), antes de la reproducción. En estas condiciones, la frecuencia del alelo A_1 en cualquiera de las islas estará dada por $p_t = \bar{p} + (1-m)^t(p_0 - \bar{p})$, con $\bar{p}$ la media de las frecuencias del alelo A_1 en las islas y p_0 la frecuencia inicial **en la isla** que estamos considerando.
- Las migraciones de individuos entre poblaciones contribuyen a romper el efecto de diferenciación genética entre ellas a causa de la evolución diferente de las frecuencias alélicas en ellas. En general alcanza con unos pocos migrantes por generación para reducir sustantivamente la diferenciación genética. La relación entre el número de migrantes Nm y el estadístico F_{ST} está dada por $\hat{F}_{ST} \approx \frac{1}{1+4Nm}$.

6.9 Mecanismos de especiación

Si bien actualmente la ciencia acepta la teoría evolutiva darwinista como la base de la existencia de las diferentes especies que habitan nuestro planeta, los mecanismos por detrás de formación de las nuevas especies continúa siendo un tópico álgido en el campo de la evolución. Es decir, bajo el concepto de que una especie es diferente de otra si es imposible obtener descendencia fértil de la unión de ambas, resulta claro que debemos dilucidar cuáles son los mecanismos que producen esas barreras reproductivas entre individuos hasta hace poco muy cercanos, y cuál es la importancia relativa de cada una de ellos.

En la figura 6.12 se muestran algunos de los mecanismos posibles de **especiación**.

De acuerdo con lo que vimos en la sección «Subdivisión, migración y el modelo de islas», aún un flujo bajo de migrantes por generación alcanza para reducir sustancialmente la diferenciación genética entre poblaciones, diferenciación sin la cual es imposible pensar en la formación de especies, al menos en los loci directamente relacionados al bloqueo reproductivo. Es decir, al menos para los loci involucrados en la formación de las **barreras precigóticas** (aquellas que tienen lugar antes o durante la fecundación) o de las **barreras poscigóticas** (las que afectan la viabilidad de los individuos producidos, por ejemplo, a partir de abortos espontáneos, esterilidad del híbrido, muerte prematura, híbridos débiles, etcétera) debe existir una diferenciación importante.

El primer mecanismo que vemos en la figura, a la izquierda, es el de la especiación **alopátrica**. En la misma, en determinado momento en la historia evolutiva de una especie surge algún tipo de barrera física que separa una parte de los individuos de los otros. Ejemplos

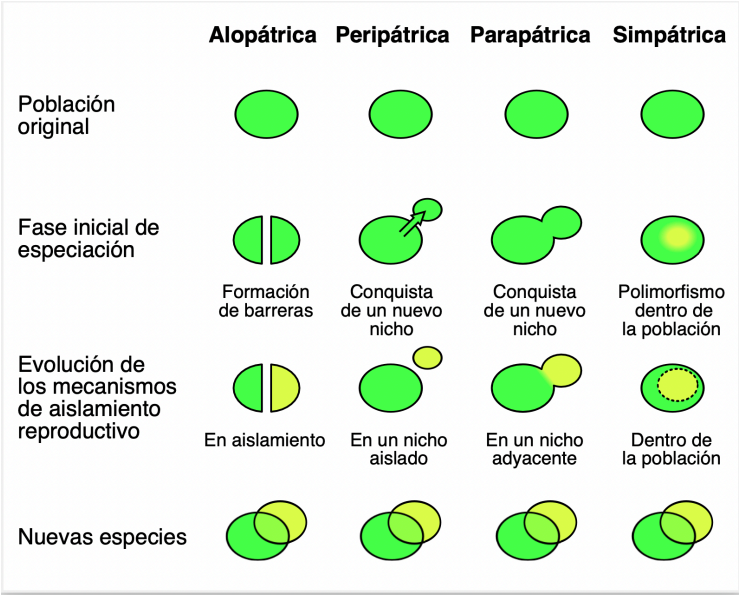


Figura 6.12. Mecanismos de especiación. Imagen tomada de Wikipedia, Speciation_modes.svg: Ilmari Karonen derivative work: Mircalla22 (talk) - Speciation_modes.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=8637025.

de aparición geológica de estas barreras puede ser el surgimiento de una cadena montañosa, la aparición de un cuerpo de agua en el medio del territorio original (por ejemplo, la aparición de islas como producto del ascenso del nivel del mar en los grandes deshielos), el surgimiento de una barrera de tierra para poblaciones acuáticas (por ejemplo, el istmo de Panamá) y cualquier otro evento que impida el intercambio por un tiempo suficiente de material genético. Las poblaciones en aislamiento unas de otras irán acumulando diferencias genéticas entre ellas, ya sea por los procesos estocásticos (deriva, mutación) como por procesos selectivos. Es importante notar que luego de generadas las barreras reproductivas no importa si las poblaciones vuelven a coexistir en el territorio (por ejemplo, por la unión de islas al volver a bajar el nivel del mar) ya que al no ser capaces de generar descendencia fértil no existirá flujo genético entre ellas. Un ejemplo muy claro de este proceso de formación de especies lo vimos en la sección «Selección direccional» y es el de los pinzones de Darwin, donde debido a la relativamente importante barrera física que constituye la distancia entre islas, el número de migrantes entre poblaciones es pequeño, dándose las condiciones para la acumulación de diferencias genéticas entre poblaciones. Otros ejemplos notables incluyen el **enanismo insular** y el **gigantismo insular**, fenómenos en los que los individuos de una especie que se encuentran aislados en una isla desarrollan un tamaño completamente fuera del rango normal para la especie y terminan generando barreras reproductivas físicas (biológicas), entre otras. Las razones que se postulan son variadas, pero en general se refieren a un proceso selectivo diferencial: a medida que los recursos naturales se agotan en la isla, animales más pequeños necesitan menores cantidades de alimento y de espacio habitable, por lo que se ven favorecidos (**enanismo insular**), mientras que en ausencia de limitantes en los recursos, mayores tamaños suelen competir mejor (**gigantismo insular**). Un ejemplo de **enanismo insular** lo constituye el **elefante pigmeo** de la isla de Malta (*Palaeoloxodon falconeri*, extinto), cuya altura a la grupa no sobrepasaba el metro en los machos (y menos en

las hembras). Los **dodos** (*Raphus cucullatus*, también extintos) en cambio constituyen un ejemplo de **gigantismo insular:** si bien evolucionaron a partir de la familia de las palomas, su altura era de cerca de 1 metro y pesaban más de 10 kilos.

La segunda forma que aparece en la figura 6.12 es la especiación **peripátrica**, donde una parte de los individuos de la especie (en general en el borde del rango geográfico) comienzan un proceso de diferenciación gradual a partir de una reducción en el intercambio de material genético con el resto. En principio se trata también de un fenómeno asociado a la distancia física entre individuos y, por lo tanto, es un mecanismo más esperable en especies con capacidad de dispersión reducida. La causa de la reducción del intercambio no tiene por qué ser geográfica ya que uno de los factores iniciales invocados es la adaptación a condiciones particulares del nicho. Lo esencial para nosotros, en el contexto de este capítulo, es la posibilidad de que las preferencias de apareamiento por individuos cercanos, sumado a la reducción sustantiva del intercambio con la población global, constituyen los elementos sustantivos para este mecanismo de especiación.

En la especiación **parapátrica** se produce sin una separación completa de las especies «original» y «derivada», que pueden compartir parte importante de su rango geográfico e incluso hibridar en determinadas locaciones. El mecanismo que se suele invocar es la adaptación local a condiciones diferentes dentro del rango y la reducción de la viabilidad de los híbridos respecto a las formas contrastantes, aunque se trata de un mecanismo usualmente difícil de diferenciar de la especiación **alopátrica** con contacto posterior a la especiación.

Finalmente, en la especiación **simpátrica** el proceso de diferenciación genética se da en el mismo rango geográfico, es decir sin separación de la especie «original» y la «derivada» (el concepto de original y derivada es a veces difícil de entender ya que ambas poblaciones derivan de una misma especie y acumulan cambios estocásticos ambas). El mecanismo usualmente invocado incluye la adaptación a nichos ecológicos bien diferentes dentro del rango geográfico o aún la creación de híbridos entre individuos de *demes* adaptados a dichos nichos. Un ejemplo a veces invocado de este tipo de especiación es la formación de una especie asexual a partir de una especie sexual. En general se invoca la «Selección disruptiva», que vimos en la sección «Diferentes formas de selección», como el proceso por detrás de este mecanismo de selección.

En resumen, para que exista especiación es necesaria la diferenciación genética de una parte de los individuos de una especie determinada. Para la existencia de esta diferenciación es necesario que exista una reducción sustancial de los migrantes entre *demes* por el tiempo suficiente hasta la aparición de las **barreras reproductivas** entre dichos *demes*. Las preferencias de apareamiento por individuos cercanos (es decir, el apareamiento no-aleatorio) constituye la base de la generación de diferenciación genética y ha sido el objeto de todo este capítulo.

PARA RECORDAR

- La base para el proceso de especiación es la diferenciación genética entre poblaciones que lleva a la aparición de **barreras reproductivas** que impiden la descendencia fértil.
- En la especiación **alopátrica**, en determinado momento en la historia evolutiva de una especie surge algún tipo de barrera física que separa una parte de los individuos de los otros. Ejemplos de ese tipo de barreras físicas pueden ser

el surgimiento de una cadena montañosa, la aparición de un cuerpo de agua en el medio del territorio, la aparición de islas como producto del ascenso del nivel del mar y cualquier otro evento que impida el intercambio por un tiempo suficiente de material genético.

- En la especiación **peripátrica** una parte de los individuos de la especie, en general en el borde del rango geográfico comienzan un proceso de diferenciación gradual a partir de una reducción en el intercambio de material genético con el resto. Se puede decir que en este caso también se trata de un aislamiento por distancia.
- En las especiaciones **parapátrica** y **simpátrica** el proceso de diferenciación genética se da en el mismo rango geográfico, es decir sin separación de la especie «original» y la «derivada» y los mecanismos involucrados tienen que ver con la especialización de nicho o la adaptación a condiciones locales.

6.10 Conclusión

En capítulos anteriores se trató cómo la formación de ensembles poblacionales lleva a la divergencia de las frecuencias alélicas entre subpoblaciones del ensemble. Este fenómeno puede ser meramente estocástico (caso cual hablamos de deriva genética), ligado a una ventaja selectiva diferencial en cada ensemble (donde el cambio en frecuencias será sistemático), o como el producto de una mezcla entre estas fuerzas. En términos más generales, una población puede estar estructurada en varios niveles jerárquicos. La existencia de estos niveles jerárquicos de estructuración implica la existencia de apareamientos no aleatorios entre los individuos de la población general; en última instancia, esto implica el apareamiento entre parientes de menor o mayor cercanía, lo cual implica un consiguiente aumento de la homocigotas. En este capítulo formalizamos este concepto (por ejemplo, definiendo la identidad por descendencia), a la vez que ahondamos en el manejo de coeficientes de fijación que nos permiten evaluar el grado de estructuración de una población. Por último, es importante remarcar que la migración constituye una fuerza que hace un contrapunto a la estructuración poblacional, pudiendo revertir su efecto aun cuando el número de migrantes por generación sea bajo. En el largo plazo, la síntesis entre estos procesos define el grado de divergencia genética entre los niveles jerárquicos de una población estructurada, pudiendo dar lugar a fenómenos tan importantes como la especiación en poblaciones naturales.

6.11 Actividades

6.11.1 Control de lectura

1. Durante este capítulo se introdujeron los conceptos de identidad por ascendencia (IBD) e identidad por ascendencia (IBS). A su vez, se introdujo el concepto de alocigosis y autocigosis. Responde:
    1. ¿Un par de alelos IBD puede ser autocigota, alocigota o ambos?
    2. ¿Qué sucede para el caso de un par de alelos IBS?
    3. ¿Puede un par de alelos ser homocigota y alocigota al mismo tiempo? ¿En dicho caso, serían IBD o IBS?
2. En el capítulo «Deriva genética» se introdujo el coeficiente de fijación F, definido como la probabilidad de que dos alelos heredados por un individuo fueran idénticos por ascendencia. La estructuración poblacional lleva a la definición de los estadísticos F_{IS}, F_{ST} y F_{IT}.
    1. Define cada uno de estos coeficientes, tanto conceptual como matemáticamente.
    2. ¿Cómo se puede describir matemáticamente el vínculo entre estos coeficientes?
    3. Explica, ¿cuál es su utilidad para el estudio de la estructuración poblacional?
3. Cuando existe un régimen de apartamiento no aleatorios en una población, se observa un déficit de heterocigotas respecto a lo esperado bajo equilibrio de Hardy-Weinberg. Este fenómeno es conocido como efecto Wahlund. Describe a qué se debe este fenómeno, teniendo en cuenta el rol que juega la deriva genética a nivel de frecuencias alélicas en dicha situación.
4. La estructuración poblacional lleva a un apareamiento no-aleatorio entre los individuos pertenecientes a cada subpoblación. La migración entre estas subpoblaciones, sin embargo, equivale a un flujo de alelos entre subpoblaciones al aumentar la cantidad de apareamientos entre individuos de diferentes subpoblaciones. ¿Qué tan efectiva es la migración en entre subpoblaciones en este aspecto, según el modelo de islas? ¿Se requiere de muchos migrantes por generación entre diferentes subpoblaciones para que la población total se comporte acorde a lo esperado bajo panmixia?
5. La existencia de apareamientos no-aleatorios en una población puede derivar en el largo plazo en el fenómeno de especiación. Según la forma de estructuración pueden distinguirse diferentes modos de especiación. Clasifica el modo de especiación en los siguientes ejemplos:
    1. La gramínea *A. capillaris* es una especie perteneciente a Eurasia. En regiones mineras existe diferente composición geoquímica respecto a suelos no explotados. Algunas variantes de *A. capillaris* se desarrollan mejor que en los suelos naturales (y viceversa; si bien existe hibridación entre las variantes, la misma posee desventaja selectiva. En el largo plazo, se genera especiación entre las poblaciones de regiones mineras y silvestres. (Ejemplo obtenido de Boggs 2001).
    2. Antes de la formación de las remotas Hébridas exteriores existía un único tipo de oveja fluorescente de la raza scottish blackface, las cuales se veían verdes bajo la luz crepuscular. Luego de la formación de las islas surgió en una de las subpoblaciones insulares un alelo que generó fluorescencia roja, fijándose en la en esta población (que además acumuló otros cambios morfológicos). Hoy en día las dos especies de ovejas ya no pueden generar descendencia fértil.

3. En Camerún existen pequeños lagos en cráteres. En un momento dado, existió una diversificación rápida de especies de peces cíclidos (familia Cichlidae); cada especie se adaptó a diferentes nichos dentro de estos pequeños lagos, en donde conviven juntas. (Ejemplo obtenido de Boggs 2001).
4. Se ha reportado que una población de mosquitos se adaptó al metro subterráneo de Londres, denominándose a la especie *Culex molestus*. La especie *Culex pipens*, de la que habría derivado *C. molestus* posee diferencias biológicas marcadas: a modo de ejemplo, la primera hiberna en invierno, mientras que la especie del subterráneo se reproduce todo el año. (Ejemplo inspirado en Fonseca et al. 2004).

6.11.2 ¿Verdadero o falso?

1. Si dos alelos son idénticos por ascendencia (IBD) deben ser idénticos en estado (IBS), ya que el estado de los alelos es, naturalmente, el mismo.
2. Si los individuos que conforman las subpoblaciones de un ensemble poblacional se aparean aleatoramente dentro de cada subpoblación, la probabilidad de que dos alelos sean IBD en un individuo de la población (es decir, F_{IT}) equivale a la proporción de la diversidad genética debida a las diferencias de frecuencia alélica entre poblaciones.
3. El aumento de apareamiento entre parientes que se da cuando se estructura una población lleva a una disminución del fitness medio de la población. Esta disminución es linealmente proporcional al coeficiente de fijación F.
4. La formación de un ensemble poblacional lleva con el tiempo a un exceso de individuos homocigotas (es decir, déficit de heterocigotas) en cada subpoblación respecto a lo esperado bajo equilibrio de Hardy-Weinberg. Cuando se considera el total de la población, sin embargo, se observa un déficit de heterocigotas; este fenómeno se conoce como efecto Wahlund.
5. La migración entre las subpoblaciones de un ensemble poblacional constituye un fenómeno que contrarresta los efectos de la estructuración poblacional (asumiendo que dentro de las subpoblaciones existe un régimen de panmixia). No obstante, es necesario de muchos migrantes por generación para que el aporte de alelos por migración se imponga al aumento de homocigosis que se espera debido a la estructuración poblacional.

Solución

1. **Falso**. El término «idéntico por estado» no debe ser tomado de forma literal. Está claro que dos alelos deberán ser idénticos en su estado si provienen de un mismo alelo ancestral identificado (es decir, si son «IBD»). No obstante, justamente se define a un par de alelos como «idénticos por estado» (IBS) cuando los alelos son idénticos pero no se puede identificar al ancestro remoto del cual derivan.
2. **Verdadero**. Se plantea un ensemble poblacional en el cual se asume panmixia para cada una de subpoblaciones. Esto implica $F_{IT} = 0$, y dado que $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \Leftrightarrow 1 - F_{IT} = 1 - F_{ST} \Rightarrow F_{IT} = F_{ST}$. Esto es de alguna forma intuitivo: si dentro de las poblaciones existe panmixia, entonces la probabilidad de que dos alelos sean IBD en un individuo debe relacionarse únicamente, en el esquema planteado, con el aumento de heterocigosidad debido a que las distintas subpoblaciones no se aparean de forma aleatoria entre sí.

3. **Verdadero.** En efecto, el fitness medio de una población estructurada decrecerá según a linealmente respecto al coeficiente de fijación F una cantidad $\Delta = -bF$, con $b = 2pqs[(\frac{1}{2}) - h]$. Resulta interesante remarcar que, según el modelo explayado en este libro, existirá una disminución en el fitness promedio de la población siempre y cuando $h < \frac{1}{2}$; en casos donde $h > \frac{1}{2}$ (modo de acción génica que contempla algunas dominancias incompletas y el caso de las sobredominancia) el modelo predice un aumento del fitness promedio de la población.
4. **Falso.** La estructuración poblacional sí lleva a un exceso de homocigotas en cada subpoblación. Esto se traduce, no obstante, como un déficit de individuos heterocigotas cuando se contempla el total de la población (lo cual se conoce como efecto Wahlund).
5. **Falso.** Al menos asumiendo un modelo sencillo, como el «modelo de islas» planteado por Sewall Wright, una tasa migratoria baja (es decir de uno o dos migrantes por generación entre subpoblaciones) es suficiente para que la varianza a nivel de frecuencias alélicas entre subpoblaciones se asemeje a la esperada bajo régimen de panmixia.

6.11.3 Ejercicios

Ejercicio 6.1

Una población de levaduras (*S. cerevisiae*) se encuentra reproduciéndose de forma asexual (es decir, bajo régimen de autocigosis). En la misma se segrega un locus bialélico A (con frecuencias alélicas $p = 0{,}56$ y $q = 0{,}44$ para los alelos A_1 y A_2, respectivamente). Se asume un tamaño poblacional a efectos prácticos infinito.

a. En un momento dado se realiza un genotipado, observándose que las siguientes frecuencias genotípicas

$$\begin{cases} f(A_1A_1) = 0{,}35 \\ f(A_1A_2) = 0{,}42 \\ f(A_2A_2) = 0{,}23 \end{cases} \tag{6.101}$$

¿Cuál es el valor del coeficiente de fijación F para la población en la generación genotipada?

b. Pasan 5 generaciones reproduciéndose bajo este régimen reproductivo. ¿Qué frecuencia de heterocigotas se espera observar? ¿A qué porcentaje del esperado bajo equilibrio de Hardy-Weinberg corresponde este valor?

Solución

a. Notemos que bajo equilibrio de Hardy-Weinberg la frecuencia esperada para individuos heterocigotas A_1A_2 es $f(A_1A_2)_{HW} = 2pq = 2 \times 0{,}56 \times 0{,}44 = 0{,}4928$. Se observa por lo tanto un déficit de heterocigotas, lo cual es esperable bajo régimen de autocigosis. Se esperan las siguientes desviaciones respecto a las frecuencias esperadas bajo equilibrio de Hardy-Weinberg

$$\begin{cases} f(A_1A_1) = p^2(1-F) + pF \\ f(A_1A_2) = 2pq(1-F) \\ f(A_2A_2) = q^2(1-F) + qF \end{cases} \tag{6.102}$$

Tenemos por lo tanto

$$f(A_1A_2) = 2pq \times (1-F) \tag{6.103}$$

$$f(A_1A_2) = f(A_1A_2)_{HW} \times (1-F) \Rightarrow \tag{6.104}$$

$$\frac{f(A_1A_2)}{f(A_1A_2)_{HW}} = 1 - F \Rightarrow \tag{6.105}$$

$$F = 1 - (\frac{f(A_1A_2)}{f(A_1A_2)_{HW}}) = \tag{6.106}$$

$$\frac{f(A_1A_2)_{HW} - f(A_1A_2)}{f(A_1A_2)_{HW}} \tag{6.107}$$

Se puede ver entonces el coeficiente de fijación como la desviación observada en heterocigosis respecto a la esperada, como se mencionó en el capítulo. Reemplazando con nuestros valores, tenemos

$F = 1 - (\frac{0{,}42}{2 \times 0{,}56 \times 0{,}44}) \approx 0{,}15$

Es decir, la desviación que se observa en la frecuencia de heterocigotas equivale a alrededor del 15 % del valor de heterocigotas esperado (un porcentaje considerable).

b. A medida que pasan las generaciones, se espera un descenso del número de heterocigotas (es decir, un aumento del coeficiente de fijación F). Este aumento sigue la ecuación

$$F_t = 1 - (\frac{1}{2})^t \times (1 - F_0) \tag{6.108}$$

donde t es el número de generaciones que transcurren a partir de un momento donde existe un coeficiente de fijación F_0 (en nuestro caso, el valor calculado allá arriba).

Por lo tanto, tendremos

$$F_5 = 1 - (\frac{1}{2})^5 \times [1 - (\frac{0{,}42}{2 \times 0{,}56 \times 0{,}44})] \tag{6.109}$$

$$F_5 \approx 0{,}995$$

Se observa un importante aumento del coeficiente de fijación tan solo cinco generaciones después. La frecuencia esperada de heterocigotas corresponde a

$$H = 2pq(1-F) = 2 \times 0{,}56 \times 0{,}44 \times (1-(1-(\tfrac{1}{2})^5 \times [1-(\tfrac{0{,}42}{2\times 0{,}56\times 0{,}44})])) = 2 \times 0{,}56 \times 0{,}44 \times ((\tfrac{1}{2})^5 \times [1-(\tfrac{0{,}42}{2\times 0{,}56\times 0{,}44})])$$

$$H = 2pq(1-F) \tag{6.110}$$

$$H = 2 \times 0{,}56 \times 0{,}44 \times (1-(1-(\frac{1}{2})^5 \times [1-(\frac{0{,}42}{2 \times 0{,}56 \times 0{,}44})])) = \tag{6.111}$$

$$H = 2 \times 0{,}56 \times 0{,}44 \times \tag{6.112}$$

$$((\frac{1}{2})^5 \times [1-(\frac{0{,}42}{2 \times 0{,}56 \times 0{,}44})]) \tag{6.113}$$

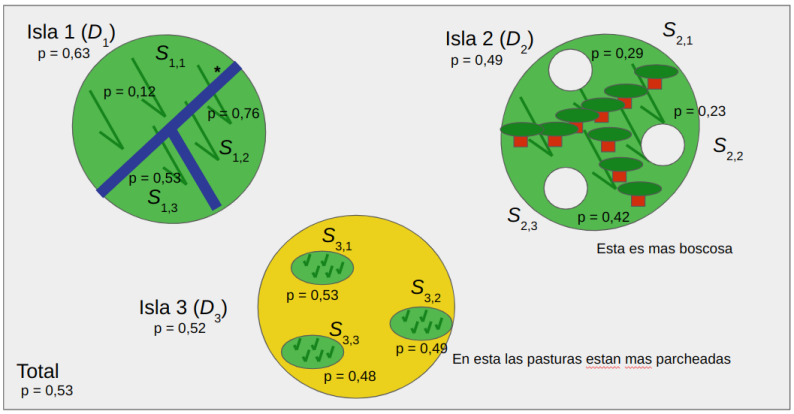
$$H = 0{,}002275 \tag{6.114}$$

Constatamos que la heterocigosidad esperada es, llegado este punto, extremadamente baja.

Ejercicio 6.2

(Inspirado en libro de Hartl 2020)

Un grupo de especialistas en ecología se dispone a estudiar la estructuración poblacional de las ya míticas ovejas que habitan las remotas Hébridas exteriores. Motiva su estudio la distribución heterogénea dentro de cada isla, debido a irregularidades en la geografía y distribución de recursos en las islas. Para ello se disponen a estimar la frecuencia de un locus marcador (frecuencia denotada con p) en tres islas diferentes, las cuales poseen distribuciones irregulares de ovejas (ver figura a continuación). El estudio fue hecho de tal forma de muestrear un número comparable de individuos tanto entre islas como dentro de ellas.



a. ¿Existe estructuración poblacional en alguna de las islas? Interpreta qué sugieren los resultados respecto a los regímenes de apareamiento dentro de cada isla.

b. Si la frecuencia de heterocigotas en la subpoblación $S_{1,1}$ es $H_{1,1} = 0{,}2$, ¿cuál es la probabilidad de que los alelos de un individuo muestreado de la misma sean idénticos por ascendencia?

Solución

a. Para determinar si existe estructuración poblacional dentro de cada isla, calcularemos el estadístico F en cada caso. Notemos que en este caso pueden existir varios niveles de estructuración: en primer lugar podría existir estructuración a nivel de islas (las cuales han sido marcadas con el símbolo D, por *deme*), mientras que dentro de las mismas existen posibles subpoblaciones (las cuales han sido marcadas con el símbolos $S_{i,j}$, para denotar a la j-ésima subpoblación de la i-ésima isla).

Tenemos para considerar la estructuración a la interna de la isla el siguiente coeficiente de fijación

$$F_{SD} = \frac{H_D - \bar{H}_S}{H_D} \tag{6.115}$$

donde H_D corresponde a la heterocigosidad esperada en cada isla y $\bar{H}_S$ es la heterocigosidad promedio observada entre subpoblaciones de la misma (la cual, por cómo se detalla se hizo el estudio, equivale al promedio de heterocigosidades observadas).

A continuación, se realiza el desglose de estos cálculos para las tres islas

	Heterocigosidad esperada (H_D)	Heterocigosidad observada	F_{SD}
Isla 1	$2 \times 0{,}63 \times (1 - 0{,}63) = 0{,}4662$	$\approx 0{,}31$	$\frac{0{,}4662 - 0{,}31}{0{,}4662} \approx 0{,}236$
Isla 2	$2 \times 0{,}49 \times (1 - 0{,}49) = 0{,}4998$	$\approx 0{,}4470$	$\approx \frac{0{,}4998 - 0{,}4470}{0{,}4998} \approx 0{,}106$
Isla 3	$2 \times 0{,}52 \times (1 - 0{,}52) = 0{,}4992$	$\approx 0{,}4991$	$\approx \frac{0{,}4992 - 0{,}4991}{0{,}4992} \approx 0{,}000200$

Basándonos en estos resultados y en los valores guía propuestos por Sewall Wright podemos decir que en la Isla 1 existe una diferenciación genética importante, en la Isla 2 una diferenciación genética moderada y en la Isla 3 una poca diferenciación genética. Esto podría reflejar diferencias respecto a en qué magnitud las condiciones geográficas impiden el apareamiento aleatorio entre las subpoblaciones de cada isla (es de esperar que un río sea una mayor limitante que una barrera de árboles, por ejemplo).

b. Consideramos que la probabilidad de que un individuo muestreado de cualquiera de las islas este conformado por alelos idénticos por descendencia debe tener en cuenta el efecto conjunto de todos los niveles de estructuración. Es decir, contemplar que sean puedan ser idénticos por ascendencia por pertenecer a una subpoblación de una isla en particular (en donde se conformaron las subpoblaciones), de una isla y del archipiélago en sí (en el que se conformaron las islas).

El coeficiente de fijación dentro de una subpoblación será

$$F_{IS} = \frac{\bar{H}_S - H_I}{\bar{H}_S} \tag{6.116}$$

donde H_I es la heterocigosidad observada y $\bar{H}_S$ es el promedio de heterocigosidad entre subpoblaciones. De modo análogo se pueden plantear los coeficientes

$$F_{SD} = \frac{\bar{H}_D - \bar{H}_S}{\bar{H}_D} \tag{6.117}$$

y

$$F_{DT} = \frac{H_T - \bar{H}_D}{H_T} \tag{6.118}$$

donde $\bar{H}_D$ es el promedio de heterocigosidad entre islas y H_T la heterocigosidad total esperada.

Contemplando todo esto tenemos

$$\begin{aligned}(1 - F_{IT}) = (1 - F_{IS})(1 - F_{SD})(1 - F_{DT}) \Rightarrow \\ F_{IT} = 1 - [(1 - F_{IS})(1 - F_{SD})(1 - F_{DT})]\end{aligned} \tag{6.119}$$

donde F_{IT} es la probabilidad de que los alelos de un individuo sean IBD. En este caso puntual, tendremos entonces

$$\bar{H}_S = [\frac{(2 \times 0{,}12 \times 0{,}88)}{9} + \frac{(2 \times 0{,}76 \times 0{,}24)}{9} + \frac{(2 \times 0{,}23 \times 0{,}77)}{9} + \tag{6.120}$$

$$\frac{(2 \times 0{,}29 \times 0{,}71)}{9} + \frac{(2 \times 0{,}23 \times 0{,}77)}{9} + \frac{(2 \times 0{,}42 \times 0{,}58)}{9} + \tag{6.121}$$

$$\frac{(2 \times 0{,}53 \times 0{,}47)}{9} + \frac{(2 \times 0{,}49 \times 0{,}51)}{9} + \frac{(2 \times 0{,}48 \times 0{,}52)}{9}] \tag{6.122}$$

$$\bar{H}_S \approx 0{,}409 \tag{6.123}$$

$$\begin{aligned}\bar{H}_D = \frac{(2 \times 0{,}63 \times 0{,}37) + (2 \times 0{,}49 \times 0{,}51) + (2 \times 0{,}52 \times 0{,}48)}{3} \\ \bar{H}_D = 0{,}4884\end{aligned} \tag{6.124}$$

$$\begin{aligned}H_T = 2 \times 0{,}53 \times 0{,}47 \\ H_T = 0{,}4982\end{aligned} \tag{6.125}$$

Por lo que tenemos

$$F_{IS} = \frac{0{,}409 - 0{,}2}{0{,}409} \tag{6.126}$$

$$F_{SD} = \frac{0{,}4884 - 0{,}4158}{0{,}4884} \tag{6.127}$$

$$F_{DT} = \frac{0{,}4982 - 0{,}4884}{0{,}4982} \tag{6.128}$$

$$F_{IT} = 1 - [(1 - (\frac{0{,}409 - 0{,}2}{0{,}409})) \times (1 - (\frac{0{,}4884 - 0{,}4158}{0{,}4884})) \times (1 - (\frac{0{,}4982 - 0{,}4884}{0{,}4982}))]$$
$$F_{IT} \approx 0{,}592 \tag{6.129}$$

Ejercicio 6.3

Una población de *Nicotiana tabacum* (tabaco) se encuentra segregando dos alelos en el locus N (por simpleza asumiremos que es diploide), el cual confiere resistencia frente al virus de mosaico del tabaco (TMV, por sus siglas en inglés). Uno de los alelos (N_A, con frecuencia promedio en la población $\bar{p} = 0{,}78$) logra inmunizar totalmente a las plantas frente al virus, mientras que portadores de la otra variante (N_B, frecuencia promedio en población $\bar{q} = 0{,}22$) son más susceptibles al patógeno. La población total se encuentra estructurada en subpoblaciones, debido a temas logísticos vinculados a su cultivo. (Nota: se asume que dentro de cada subpoblación existe un régimen de panmixia).

a. Luego de diez generaciones de establecido el cultivo, un genotipado arroja que la heterocigosidad promedio de la población es de $\bar{H} = 0{,}27$. ¿Causó la estructuración poblacional un aumento de la homocigosis respecto a lo esperado bajo régimen de panmixia?

b. Estima el tamaño poblacional efectivo de las subpoblaciones del ensemble.

c. ¿Cuál es la pérdida de fitness promedio en la población total, llegado este momento? Se asume un régimen de selección direccional, con coeficiente de selección $s = 0{,}02$ en contra de los individuos homocigotas N_BN_B y efecto de heterocigotas $h = 0{,}3$.

Solución

a. La heterocigosidad esperada bajo equilibrio de Hardy-Weinberg es de $\bar{H}_{HW} = 2\bar{p}\bar{q} = 2 \times 0{,}78 \times 0{,}22 = 0{,}3432$. Dado que la homocigosidad esperada corresponde a su complemento, tenemos $\bar{G}_{HW} = 1 - \bar{H}_{HW} = 1 - 0{,}3432 = 0{,}6568$. Ya que la homocigosidad observada es de $\bar{H} = 0{,}27 \Rightarrow \bar{G} = 1 - 0{,}27 = 0{,}73$. En efecto, la estructuración lleva a un aumento de la frecuencia de homocigotas respecto a lo esperado de $\Delta = \bar{G} - \bar{G}_{HW} = 0{,}73 - 0{,}6568 = 0{,}0732$.

b. Recordemos que la heterocigosidad esperada disminuye en un ensemble según (ver capítulo «Deriva genética»)

$$H_t = (1 - \frac{1}{2N_e})^t H_0 \tag{6.130}$$

Partiendo de la heterocigosidad H_0 esperada bajo panmixia, luego de diez generaciones observamos $H_{10} = 0{,}27$, de lo cual tenemos

$$0{,}27 = (1 - \frac{1}{2N_e})^{10} \times 0{,}3432 \tag{6.131}$$

$$\frac{0{,}27}{0{,}3432} = (1 - \frac{1}{2N_e})^{10} \tag{6.132}$$

$$\sqrt[10]{\frac{0{,}27}{0{,}3432}} = 1 - \frac{1}{2N_e} \tag{6.133}$$

$$\frac{1}{2N_e} = 1 - \sqrt[10]{\frac{0{,}27}{0{,}3432}} \tag{6.134}$$

$$2N_e = \frac{1}{1 - \sqrt[10]{\frac{0{,}27}{0{,}3432}}} \tag{6.135}$$

$$N_e = \frac{1}{2 \times (1 - \sqrt[10]{\frac{0{,}27}{0{,}3432}})} \tag{6.136}$$

$$N_e \approx 21 \tag{6.137}$$

c. La pérdida de fitness promedio en la población equivale a $\bar{w}_F - \bar{w} = -2pqs(\frac{1}{2} - h)F$. A su vez, la heterocigosidad observada equivale a $\bar{H} = 2pq(1 - F)$, de lo cual tenemos

$$\begin{aligned} 0{,}27 &= 2 \times 0{,}78 \times 0{,}22 \times (1 - F) \\ \frac{0{,}27}{2 \times 0{,}78 \times 0{,}22} &= 1 - F \\ F &= 1 - (\frac{0{,}27}{2 \times 0{,}78 \times 0{,}22}) \end{aligned} \tag{6.138}$$

Por lo tanto, reemplazando con los valores dados, tenemos una pérdida de fitness promedio frente al régimen selectivo establecido por la presencia del virus TMV de

$$\begin{aligned} \bar{w}_F - \bar{w} &= -2 \times 0{,}78 \times 0{,}22 \times 0{,}02 \times (\frac{1}{2} - 0{,}3) \times [1 - (\frac{0{,}27}{2 \times 0{,}78 \times 0{,}22})] \\ \bar{w}_F - \bar{w} &= -0{,}0002928 \end{aligned} \tag{6.139}$$

Ejercicio 6.4

En una isla compuesta por cien individuos diploides se fijó el alelo A_1, producto de la acción de la deriva genética. Cerca de la isla existe un continente donde en cambio se encuentra fijo el alelo A_2. A partir de un momento comienzan a arribar de forma anual a la isla, desde el continente. Debido al límite de recursos, cada migrante que llega implica una partida de algún habitante de la isla.

a. Luego de cinco años, ¿qué frecuencia se espera para el locus A_1 en la isla, si la tasa de migración es de 2, 10 o 30 migrantes por año?

b. Determina la tasa de migración mínima para que, luego de 20 años, la diferencia entre frecuencias alélicas A_1 entre isla y continente no sea mayor a 0,01.

Solución

a. Si aplicamos el modelo de continente-isla, y denotamos por p_i a la frecuencia del alelo A_1 en la isla a tiempo i, y con $p^\star$ a la frecuencia del alelo en el continente, la frecuencia luego de t eventos migratorios estará dada por

$$p_t = p^\star + (1-m)^t(p_0 - p^\star) \tag{6.140}$$

donde m es la proporción de migrantes desde el continente (siendo el número de migrantes por año equivalente a $2Nm$). De lo último, tenemos que las tasas migratorias en cada uno de los casos planteados equivalen a

$$\begin{cases} m_1 = \frac{2}{2\times 100} = 0{,}01 \\ m_2 = \frac{10}{2\times 100} = 0{,}05 \\ m_3 = \frac{30}{2\times 100} = 0{,}15 \end{cases} \tag{6.141}$$

Por lo tanto, se tiene luego de cinco años para cada uno de estos escenarios (y teniendo en cuenta que $p^\star = 0$ y $p_0 = 1$)

$$\begin{cases} p_5 = 0 + (1-0{,}01)^5 \times (1-0) \approx 0{,}95 \\ p_5 = 0 + (1-0{,}05)^5 \times (1-0) \approx 0{,}77 \\ p_5 = 0 + (1-0{,}15)^5 \times (1-0) \approx 0{,}44 \end{cases} \tag{6.142}$$

donde se observa claramente el decaimiento exponencial de la frecuencia p en la isla en función de la proporción de migrantes m que arriban desde el continente.

b. De la ecuación $p_t = p^\star + (1-m)^t(p_0 - p^\star)$ tenemos que la diferencia entre isla y continente equivale a $p_t - p^\star = (1-m)^t(p_0 - p^\star)$. Se plantea por lo tanto que, luego de veinte eventos migratorios, se observe

$$0{,}01 > p_t - p^\star = (1-m)^t(p_0 - p^\star) \tag{6.143}$$

$$0{,}01 > (1-m)^{20}(1-0) \tag{6.144}$$

$$0{,}01 > (1-m)^{20} \tag{6.145}$$

$$\sqrt[20]{0{,}01} > (1-m) \Rightarrow \tag{6.146}$$

$$1 - \sqrt[20]{0{,}01} < m \tag{6.147}$$

Esta proporción migratoria implica al menos $2Nm = 2 \times 100 \times (1 - \sqrt[20]{0{,}01}) \approx 41$ migrantes por año. A partir de esta cantidad, por lo tanto, luego de veinte años el alelo A_1 habrá prácticamente sido reemplazado por el alelo A_2 (solo tomando en cuenta el efecto migratorio, dejando de lado los posibles efectos de la deriva genética en el cambio de las frecuencias alélicas).

Ejercicio 6.5

En una población se encuentra segregando un locus bialélico A (con alelos A_1 y A_2 de frecuencias alélicas $p = 0{,}632$ y $q = 0{,}368$, respectivamente). En un momento dado se forma un ensemble poblacional, donde cada subpoblación posee un tamaño efectivo de alrededor de 40 individuos, los cuales se asume se reproducen bajo régimen de panmixia dentro de cada subpoblación.

a. Luego de 50 generaciones se realiza un genotipado y se observa que la varianza en la frecuencia del alelo A_1 entre las subpoblaciones del ensemble es $\sigma_p^2 = 0{,}00664$. ¿Concuerda esto con la variación esperada, asumiendo que la única fuerza evolutiva que actúa alterando las frecuencias alélicas entre las subpoblaciones es la deriva genética?

b. ¿Se encuentra la población total bajo régimen de panmixia?

c. Asumiendo que se puede aplicar el modelo de islas, determina el número promedio de migrantes por generación entre las poblaciones del ensemble. Discute los posibles efectos de la migración con relación a los valores calculados en las anteriores partes del ejercicio.

Solución

a. De acuerdo con lo visto en el capítulo «Deriva genética», la varianza esperada para la frecuencia de un alelo en un ensemble poblacional corresponde a

$$\sigma_t^2(p) = t \times \frac{pq}{2N_e} \tag{6.148}$$

En nuestro caso, luego de $t = 50$ generaciones tenemos una varianza esperada en el ensemble de $\sigma_{50}^2(p) = 50 \times \frac{0{,}632 \times 0{,}368}{2 \times 40} \approx 0{,}145$. La varianza alélica es por lo tanto bastante menor a la esperable ($\frac{0{,}00664}{0{,}145} \times 100 \approx 4{,}6\,\%$ del valor esperado) suponiendo que la única fuerza evolutiva operando sobre las frecuencias alélicas en el ensemble es la deriva genética.

b. La pregunta planteada equivale a evaluar si existe estructuración poblacional en el ensemble. Dado que cada subpoblación se reproduce bajo panmixia, tenemos

$$\begin{aligned} (1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \Rightarrow F_{IS} = 0 \Rightarrow \\ 1 - F_{IT} = 1 - F_{ST} \\ F_{IT} = F_{ST} \end{aligned} \tag{6.149}$$

es decir, el único grado de estructuración a considerar es a nivel de las subpoblaciones. Dado que $F_{ST} = \frac{\sigma_p^2}{\bar{p}\bar{q}}$ y que en un ensemble poblacional se espera que $\bar{p}$ permanezca constante, tenemos

$$F_{ST} = \frac{0{,}00664}{0{,}632 \times 0{,}368} \approx 0{,}0285 \tag{6.150}$$

un valor de $F_{ST} < 0{,}05$, lo cual sugiere que en efecto la población general se aproxima a lo esperado bajo régimen de panmixia.

c. Si se asume que aplica el modelo de islas, se tiene

$$\begin{aligned} \hat{F}_{ST} &= \frac{1}{1 + 4N_e m} \\ \frac{1}{\hat{F}_{ST}} &= 1 + 4N_e m \Rightarrow \\ 4N_e m &= \frac{1}{\hat{F}_{ST}} - 1 \\ N_e m &= (\frac{1}{4})(\frac{1}{\hat{F}_{ST}} - 1) \\ N_e m &= (\frac{1}{4})(\frac{1}{(\frac{0{,}00664}{0{,}632 \times 0{,}368})} - 1) \\ N_e m &\approx 9 \end{aligned} \tag{6.151}$$

Resulta claro que el número estimado de migrantes por generación es lo suficientemente sustancial como para que la estructuración geográfica de las subpoblaciones no corresponda a una estructuración genética: cada migrante transporta alelos de una subpoblación a la otra, lo cual lleva a que el régimen de apareamiento entre las mismas sea equivalente a un régimen de panmixia.

Capítulo 7

Dinámica y genética de poblaciones microbianas

Without question, the greatest invention in the history of mankind is beer. Oh, I grant you that the wheel was also a fine invention, but the wheel does not go nearly as well with pizza.[1]

Dave Barry

Chi è che dis ch'el vin el fa mal, l'è tutta gente, l'è tutta gente. Chi è che dis ch'el vin el fa mal, l'è tutta gente de l'ospedal. Io ne ho bevuto tanto, e non mi ha fatto male. L'acqua sí che fa male: il vino fa cantar.[2]

Anónimo popular

En general, todos tenemos una visión muy antropocéntrica de la vida y, por lo tanto, nos cuesta muchas veces tomar real magnitud de la variabilidad existente entre lo que vemos, mucho más en lo que no vemos. El conocimiento de la existencia de vida microscópica tiene varios siglos, pero no fue hasta la aparición del microscopio que comenzamos a apreciar su importancia y empezar a sospechar de su diversidad. Más aún, no fue hasta los trabajos de Louis Pasteur[3][4] que se apreció su relevancia, tanto en la enfermedad como en los procesos biotecnológicos usuales en nuestra alimentación. A partir de ese momento fue claro que la diversidad de organismos microscópicos era enorme y que su importancia para la vida del planeta también lo era. No es, sin embargo, hasta los trabajos de Carl Woese,[5] a inicios del último cuarto del siglo XX, que se aprecia que entre esos organismos la diversidad surgía a nivel basal en el árbol de la vida y que dos de los tres dominios de la vida (Archaea

[1] Frase atribuida a Dave Barry, humorista y escritor, Premio Pullitzer. Una traducción podría ser algo como «Sin lugar a dudas, el mejor invento en la historia de la humanidad es la cerveza. Bueno, reconozco que la rueda también es un buen invento, pero no va tan bien con la pizza.

[2] Canción popular de la Lombardía, región norte de Italia, popularizada por Nanni Svampa en los anõs sesenta del siglo XX a través de la canción *El Minestron*. Una traducción podría ser: «¿Quién dice que el vino hace mal? Todos son gente, todos son gente. Quién dice que el vino hace mal, son todos gente del hospital. Yo he bebido tanto y no me ha hecho mal, el agua sí que hace mal: el vino te hace cantar.

[3] Louis Pasteur (27 de diciembre de 1822 - 28 de setiembre de 1895) fue un físico-químico francés que contribuyó al establecimiento de la microbiología como rama de la ciencia. Además de notables contribuciones a la físico-química, por ejemplo, en la formación de cristales, realizó trabajos claves para refutar la teoría de la **generación espontánea**. Es además el fundador del Institut Pasteur, que surgió como instituto antirrábico y que ha realizado notables contribuciones a la biología y medicina, incluyendo desde 1908 diez premios Nobel en esa categoría.

[4] «Le vin peut-être à bon droit considéré comme la plus saine, la plus hygiénique des boissons». Louis Pasteur, *Études sur le Vin*, 1866. Si bien los procesos de fermentación no son producidos en general por bacterias y en el caso particular del vino se trata de levaduras, ambos son considerados microorganismos.

[5] Carl Richard Woese (15 de julio de 1928 – 30 de diciembre de 2012) fue un microbiólogo y biofísico conocido por sus trabajos en la filogenética de la vida y por haber definido en dominio Archaea a partir del análisis del ARNr 16S. Originalmente formado en matemáticas y física, su interés en descifrar el código genético lo fue llevando a trabajar en diferentes preguntas centrales de la biología, algunas relacionadas con el origen de la vida en nuestro planeta.

y Bacteria) eran de organismos microscópicos a los que juntos denominamos procariotas (porque no poseen un núcleo altamente organizado, como los eucariotas).

Estos organismos, los procariotas, poseen una diversidad de estilos de vida enorme y los podemos encontrar en los ambientes que nos imaginemos, desde las fuentes termales oceánicas con temperaturas cercanas a los 100 °C, hasta en la Antártida, desde nuestra piel hasta nuestros intestinos, en el rumen, en la rizósfera, en las aguas ácidas del drenaje de las minas, en el océano y en casi todos los ambientes conocidos en los que exista vida. Pequeños y con gran capacidad de multiplicación, comparten algunas características importantes entre ellos, pero se diferencian en otras importantes, por ejemplo, entre archaeas y bacterias a nivel de la maquinaria y procesos moleculares. Sin embargo, para el cometido del presente capítulo, alcanza con agruparlos en el término *procariotas* y ya los distinguiremos cuando sea necesario. Más aún, si bien en lo que hace a la parte genética estrictamente nos vamos a referir a los procariotas, en lo que hace a dinámica poblacional, los aportes de este capítulo bien podrían aplicarse a otros organismos usualmente unicelulares o aún a virus, lo que explica el título de este capítulo.

OBJETIVOS DEL CAPÍTULO

☐ Describir las particularidades genéticas de los procariotas, comparándolos con los eucariotas diploides que hemos utilizado como modelo en el desarrollo de modelos matemáticos y conceptuales en capítulos anteriores. Examinaremos modelos aleatorios alternativos al modelo de Wright-Fisher presentado en capítulos anteriores, interconectando los modelos obtenidos.

☐ Comprender la dinámica de las poblaciones microbianas y los factores que contribuyen a su crecimiento rápido, analizando los límites de este crecimiento a partir de modelos matemáticos explícitos.

☐ Explorar el papel de la selección en organismos monoploides y su influencia en la dinámica poblacional. Discutiremos también el rol del estudio de organismos procariotas en el debate entre las teorías seleccionistas y neutralistas de la evolución molecular.

☐ Discutir cómo los avances tecnológicos en genómica contribuyen a la comprensión de la dinámica poblacional en procariotas.

☐ Abordar el tema de la resistencia a los antimicrobianos en el contexto de la genética de poblaciones de microorganismos.

7.1 Genómica y mecanismos de herencia en procariotas

Hasta ahora nos hemos manejado con las ideas y conceptos fundamentales acerca de los mecanismos de herencia que se corresponden con organismos diploides y poliploides, esencialmente en eucariotas casi sin mención a los procariotas. Si bien en lo que respecta a los mecanismos conceptuales de duplicación de la información existen pocas diferencias que sean relevantes para nuestro enfoque del curso, las diferencias a otros niveles son sustantivas y tienen profundas implicancias en los distintos procesos evolutivos que ocurren en las poblaciones. Hasta hace relativamente poco se consideraba que los procariotas eran constitutivamente haploides, de alguna forma vinculándolo al hecho de un núcleo relativamente desorganizado (respecto a eucariotas), pero de un tiempo a esta parte se viene acumulando evidencia que apunta a que la poliploidía se trataría de un fenómeno bastante extendido también en procariotas (Soppa 2014). Ejemplos notables son la bacteria *Deinococcus radiodurans*, altamente tolerante a las radiaciones y que posee entre 4 y 10 copias de su genoma (dependiendo, entre otras cosas, de la velocidad de crecimiento), o la arquea *Haloferax volcanii* que es capaz de usar las copias extras de su genoma como fuente de almacenamiento de fosfato. Pese a los variados contraejemplos, la norma sigue siendo por el momento que la mayor parte de los procariotas son haploides (o monoploides, un término posiblemente más adecuado). Por lo tanto, en este capítulo le dedicaremos una atención importante a este tema y a cómo esta característica afecta a la dinámica evolutiva de las poblaciones procariotas.

Otra diferencia notable entre procariotas y eucariotas es la existencia de circularidad en los cromosomas de los primeros. Mientras que en los eucariotas los cromosomas son lineales (lo cual implica un gran problema respecto a la conservación de sus extremos), en el caso de los procariotas la mayor parte de las especies poseen cromosomas circulares, sin bordes libres. En eucariotas los riesgos y dificultades que implica la existencia de bordes libres en los cromosomas (por ejemplo, para la replicación del ADN) se han resuelto a través de la aparición de telómeros, y es generalmente aceptado que la longitud de estos se relaciona estrechamente con la edad celular. Por otro lado, la circularidad existente en cromosomas procariotas presenta también sus desafíos, en particular en lo que respecta a la separación de las dos copias resultantes de la replicación del genoma. De hecho, existen abundantes ejemplos de procariotas que poseen cromosomas lineales, y diversos eventos de linealidad-circularidad parecen haber ocurrido en diversos puntos de la filogenia de procariotas (Moldovan 2019).

Ambos puntos mencionados anteriormente, la existencia de genomas haploides en procariotas y la circularidad de los mismos, presentan un desafío importante al momento de aplicar los conceptos y modelos de recombinación y el desequilibrio de ligamiento tratados en capítulos anteriores. Mientras que la existencia de una sola copia del genoma en las células presenta la obvia restricción de que resulta imposible recombinar con la otra copia genómica (porque no existe), es importante tener en cuenta que los procariotas sin embargo acceden a otros mecanismos que posibilitan la recombinación genética dentro de la especie, o aún la incorporación de material genético de otras especies (fenómeno conocido como transferencia horizontal). Los mecanismos clásicos de intercambio genético de las bacterias incluyen la **transducción**, la **transformación** y la **conjugación**. En el primero de estos mecanismos un bacteriófago[6] se adsorbe sobre la superficie bacteriana para inyectar su material genético en la célula. Dependiendo de si se trata de un fago virulento o moderado, el **programa genético** puede derivar en un ciclo **lítico** donde se producirán múltiples copias del mismo

[6]Se denomina *bacteriófagos*, o simplemente *fagos,* a un grupo de virus que infectan bacterias.

y la concomitante lisis celular, o el fago puede integrar su ADN en el genoma bacteriano, transformándose en un **profago**, hasta que en algún momento se produzca la **inducción** del mismo (es decir, la activación del fago); a través de este proceso el fago puede arrastrar material genético adicional de unas bacterias a otras.

El segundo mecanismo es la transformación, donde luego de entrar en un estado particular conocido como **competencia**, las células bacterianas son capaces de captar ADN libre de doble cadena e incorporarlo en su genoma. En las bacterias **Gram-positivas** la bacteria no discrimina entre ADN homólogo y heterólogo, aunque la incorporación depende de la similaridad de secuencias, mientras que en el caso de las **Gram-negativas** la bacteria reconoce la diferencia y en general solo incorporará material de la misma especie o de especies estrechamente relacionadas. El tercer mecanismo, la **conjugación**, podríamos decir que es lo más cercano al intercambio de material genético vía sexo en bacterias. En este mecanismo existen bacterias (denominadas F^+) que poseen un plásmido conjugativo, que lleva a la formación de vios *pili*[7] en la superficie que le permiten hacer contacto con bacterias que carecen de dicho plásmido (F^-). Luego de ese contacto, la despolimerización en la base de los pelos lleva a que las bacterias se acerquen y exista un contacto estrecho, formándose un puente citoplasmático por donde se transfiere el material genético del plásmido. La transferencia de ADN desde una célula F^+ a una F^- es un proceso especial de replicación asimétrica por círculo rodante, donde una de las dos cadenas parentales del plásmido F pasa a la célula receptora, replicándose en ella, mientras que la otra cadena parental se queda en el donador, sirviendo a su vez como molde para la síntesis de una nueva cadena complementaria; este hecho explica además por qué las células F^+ siguen siendo F^+ después de la conjugación. De hecho, se ha observado que el proceso suele ocurrir en agregados de bacterias donde en lugar de formarse parejas específicas se dan intercambios entre varias a la vez (Achtman 1975). ¡No parece ser tan mala la vida de las bacterias, después de todo!

El otro problema con la forma en que vimos el desequilibrio de ligamiento es que a medida que nos alejábamos de una posición dentro de un mismo cromosoma (es decir, a medida que aumentaba la distancia física entre dos loci), se esperaba que el desequilibrio se fuese atenuando en los cromosomas lineales. Más aún, vimos a partir de algunas estimaciones que la distancia genética por cM es del orden de 1,25 $textMb$, o al revés $\approx$ 0,8 Mb/cM en bovinos, mientras que estimaciones más precisas para el genoma humano son del orden de 0,75 Mb/cM. En procariotas, si bien la variabilidad del tamaño del genoma es enorme, desde algunas centenas de kb hasta más de 10 Mb, la misma es mucho menor en órdenes de magnitud que la existente en eucariotas. Un genoma bacteriano promedio se encuentra en el orden de los 4 Mb (por ejemplo, *Escherichia coli*, el caballito de batalla de la microbiología). Con los números de recombinación en los genomas bovino y humano no existiría mucho espacio para el equilibrio de ligamiento en procariotas, ya que 4 Mb divididos entre 0,75 Mb/cM nos da aproximadamente 5,33 cM. Peor aún, eso sería entre loci ubicados en los extremos de un cromosoma lineal, pero como se aprecia en la figura 7.1, en un cromosoma circular la máxima distancia entre dos puntos para un cromosoma de largo L es igual a $L/2$: si nos pasamos de ese punto en el círculo, lo que aumentamos por un lado disminuye por el otro y como en lo que hace a distancia física lo relevante es la menor de las distancias consideradas, esta será siempre $\leqslant L/2$. Es decir, en un cromosoma de tamaño 4 Mb esto implicaría, con las tasas de recombinación de humano, un máximo de $\approx$ 2,7 cM entre los loci a mayor distancia.

[7] Un **pilus** -plural, *pili*- es una estructura similar a un pelo existente en las superficies de bacterias, a través del cual la célula ejerce la conjugación.

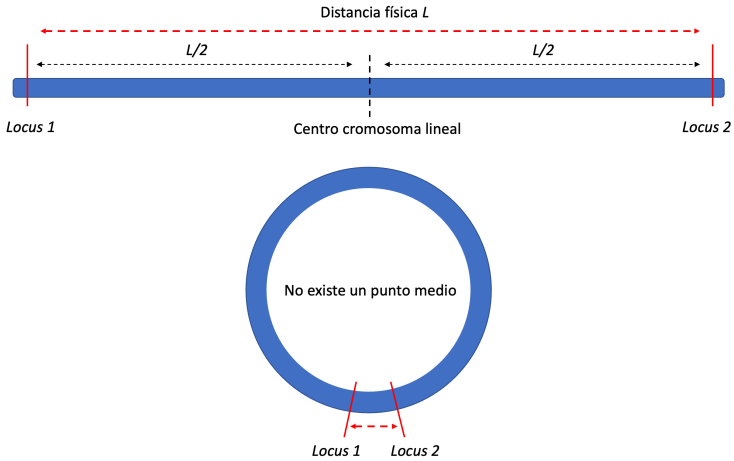


Figura 7.1. Distancia física y sus implicancias para la recombinación de cromosomas lineales y circulares. Mientras que en el cromosoma lineal la mayor distancia física ocurre entre loci que se encuentran en los extremos (y es equivalente, por lo tanto, al largo del cromosoma L), en los cromosomas circulares la distancia máxima es $L/2$, ya que en los mismos una distancia L implica una vuelta al punto de partida.

Obviamente que estos cálculos no aplican al caso de los procariotas. Por un lado, porque los mecanismos de recombinación son completamente diferentes. En procariotas un evento de recombinación requiere de dos puntos de intercambio, el inicio y fin, por lo que se asemeja al mecanismo de **conversión génica** que hemos discutido previamente. Por el otro, porque las tasas calculadas para bovinos y humanos en un evento de recombinación de material genético de la misma célula no son comparables con un proceso que requiere la llegada de material genético externo a la célula. Este material externo será en general producto de células cercanas y, dado que las bacterias se dividen por fisión, es de esperar que las mismas sean descendientes de la misma célula ancestral que originó a la receptora poco tiempo antes (ya veremos que la dinámica de las poblaciones bacterianas suele ser vertiginosa). Esto implica que el material con el que se va producir la recombinación es idéntico (o a lo sumo casi idéntico) que el del organismo receptor, por lo que será imposible distinguir los efectos de la recombinación. En conjunto estos fenómenos generan el efecto de casi-clonalidad que es posible apreciar en muchas poblaciones bacterianas, pudiendo llevar a subestimar la tasa de recombinación de las mismas. Dicho de otra forma, si nuestro muestreo de bacterias está formado por los integrantes de una colonia que desciende de una sola bacteria inicial, será casi imposible observar variabilidad genética. Esto se debe a que la misma depende de las mutaciones que hayan arribado, y solo aquellas que hayan ocurrido cerca de la fundación de la colonia producirán un número de representantes suficiente para ser observadas en el muestreo.

Una diferencia importante entre eucariotas y procariotas es lo que hace a la organización interna de los cromosomas y la arquitectura de los genes. Mientras que en eucariotas los genes están generalmente organizados en una estructura de exones separados por intrones, además de poseer secuencias reguladoras que pueden encontrarse a miles de pares de bases, en el caso de los procariotas la arquitectura de los genes es extremadamente sencilla, por lo general solo una secuencia codificante continua, con regulación en la región inmediata

upstream (antes del inicio de la codificante). Por un lado esto le da una ventaja evolutiva a los eucariotas en la posibilidad de reorganizar las proteínas para cumplir nuevas funciones a través del ***exon shuffling*** u otros mecanismos del tipo, además de ofrecer formas mucho más complejas de regulación.[8] Por el otro lado, implica un alto costo en secuencias no-codificantes que deben duplicarse, lo que es una importante desventaja también. De hecho, la escasa longitud del ADN en las secuencias codificantes de procariotas permite que la selección positiva opere para favorecer una organización de orden superior como lo son los operones, un conjunto de genes vinculados a través de una vía metabólica o de un proceso particular que se encuentran corregulados y normalmente muy cerca físicamente (lo que permite, además, que en caso de exportarse como ADN para recombinación vayan todas las piezas importantes juntas). Además, la presión por reducir el ADN que no es directamente codificante ha llevado a los procariotas a reducir las regiones intergénicas de forma sustancial. Mientras que en eucariotas como los mamíferos la proporción de ADN no-codificante es mayor al 95 %, en procariotas es usualmente del orden de $10 - 15$ % y a veces menos.

De alguna manera relacionado con todo lo anterior, si bien las posibilidades de recombinación aparecen como más complejas y con menos garantías intrínsecas, la estructura y organización de los genomas procariotas es muchísimo más flexible que la de eucariotas, al tiempo que los tamaños poblacionales suelen ser gigantescos. Esto permite que la evolución opere a una velocidad importante incluso bajo presiones selectivas bajas. La flexibilidad de la organización de los genomas procariotas se hace patente al observar la distribución del material genético en los mismos. La mayor parte de los procariotas (conocidos) poseen uno o dos cromosomas circulares que contienen casi la totalidad del material genético de las células y que definen la especie. Sin embargo, en procariotas en general y en algunos eucariotas unicelulares (por ejemplo, la levadura *Saccharomyces cerevisiae* (Chan et al. 2013)) existen elementos genéticos de carácter móvil denominados plásmidos. En general, los mismos son de pequeño tamaño comparado a los cromosomas estables del organismo correspondiente, y en los mismos la mayor parte del material genético no es indispensable para la vida en «condiciones normales» de la célula. Una parte importante de los mismos se transfiere mediante conjugación, un mecanismo altamente eficiente y, por lo tanto, se transforman en uno de los mecanismos de mayor impacto para el intercambio de material genético y base para la evolución rápida de adaptación en microorganismos (Aminov 2011).

PARA RECORDAR

- Una diferencia genética importante entre eucariotas y procariotas es que mientras que en los primeros gran parte de los genomas son diploides o poliploides, en los segundos la mayor parte son monoploides (haploides), aunque con excepciones importantes en ambos grupos.
- Mientras que en los eucariotas los cromosomas son lineales, lo que implica un gran problema respecto a la conservación de los extremos, en el caso de los procariotas la mayor parte de las especies poseen cromosomas circulares, sin bordes libres por lo tanto.

[8] Se denomina *exon shuffling* («barajado de exones», en inglés) al proceso mediante el cual se generan nuevos genes mediante el reordenamiento de la estructura intrón-exón de un gen. Otro mecanismo notable en lo referente a generación de diversidad molecular en eucariotas que tiene su base en la existencia de la estructura exón-intrón es el *splicing* alternativo («empalme» alternativo), donde a partir de un mismo gen se generan diferentes copias de ARN mensajero (y por tanto, posibles productos proteicos) mediante la selección de diferentes combinaciones de exones para la conformación del ARN mensajero que será exportado al citosol.

- Los mecanismos clásicos de intercambio genético de las bacterias incluyen la **transducción**, la **transformación** y la **conjugación**.
- En la **transducción** un bacteriófago se adsorbe sobre la superficie bacteriana para inyectar su material genético en la célula. Dependiendo de si se trata de un fago virulento o moderado, el **programa genético** puede derivar en un ciclo **lítico** donde se producirán múltiples copias del mismo y luego se producirá la lisis celular o el fago puede integrar su ADN en el genoma bacteriano, transformándose en un **profago**. En este proceso el fago puede arrastrar material genético adicional de unas bacterias a otras.
- En la **transformación**, luego de entrar en un estado particular conocido como **competencia**, las células bacterianas son capaces de captar ADN libre de doble cadena e incorporarlo en su genoma. En las bacterias **Gram-positivas** la bacteria no discrimina entre ADN homólogo y heterólogo, aunque la incorporación depende de la similaridad de secuencias, mientras que en el caso de las **Gram-negativas** la bacteria reconoce la diferencia y en general solo incorporará material de la misma especie o de especies estrechamente relacionadas.
- En la **conjugación**, existen bacterias (F^+) que poseen un plásmido conjugativo, que lleva a la formación de varios *pili* en la superficie que le permiten hacer contacto con bacterias con ausencia de ese plásmido (F^-). Luego de ese contacto, la despolimerización en la base de los pelos lleva a que las bacterias se acerquen y hagan contacto íntimo, formándose un puente citoplasmático por donde se transfiere el material genético del plásmido.
- Una diferencia importante entre eucariotas y procariotas es en lo que hace a la organización interna de los cromosomas y la arquitectura de los genes. Mientras que en eucariotas los genes están generalmente organizados en una estructura de exones separados por intrones, además de poseer secuencias reguladoras que pueden encontrarse a miles de pares de bases, en el caso de los procariotas la arquitectura de los genes es extremadamente sencilla, por lo general solo una secuencia codificante continua, con regulación en la región inmediata **upstream**.

7.2 Dinámica de las poblaciones bacterianas

Hasta ahora, en este libro nos hemos manejado con tamaños poblacionales constantes o, en limitados casos, mencionado qué consecuencias genera la variación en el tamaño poblacional en determinados contextos, pero sin dar mayor cuenta de la dinámica de las poblaciones. La justificación mayor a esto es que los supuestos asumidos cuadran razonablemente bien con la dinámica de muchas poblaciones eucariotas (incluidas las de interés económico), al menos a la escala considerada en procesos de mejoramiento genético. Sin embargo, las bacterias y otros organismos unicelulares poseen una capacidad de replicación muy alta, lo cual les permite tener dinámicas muy variables e incluso explosivas. Claramente, es mucho más sencillo duplicar una célula que un organismo multicelular, formado por muchísimas células y donde los eventos de multiplicación deben seguir un orden determinado. Existen bacterias de lenta duplicación y otras de duplicación rápida. Por ejemplo, el caballito de batalla de los laboratorios de microbiología experimental, la bacteria *Escherichia coli*, tiene capacidad de duplicarse cada 20 minutos en condiciones ideales, mientras que en otras, como

Mycobacterium tuberculosis, el tiempo de duplicación es de 12 a 16 horas. Algunas bacterias replican aún más rápido que *E. coli*. Por ejemplo, la bacteria halófila (con preferencia por medios salinos) *Vibrio natriegens* tiene capacidad de duplicarse en menos de 10 minutos. Estos números pueden no parecernos demasiado impresionantes, pero luego de unas breves cuentas estamos seguros de que se apreciará su magnitud. Si bien los modelos de dinámica poblacional que veremos en este capítulo no son exclusivos de microorganismos, en los mismos se puede apreciar en toda su magnitud el fenómeno y, por lo tanto, conforman un excelente caso para su estudio y aplicación.

El modelo exponencial

Una bacteria típica pesa en el orden de 1×10^{-12} g, o de otra forma, se precisa un billón (1×10^{12}, un millón de millones) de bacterias para que pesen un gramo. Supongamos que a tiempo cero tenemos una bacteria, por ejemplo *E. coli*, que se duplicará en condiciones óptimas y sin restricciones de nutrientes, espacio o cualquier otra limitante. Al cabo de 20 minutos tendremos dos bacterias, que al cabo de otros 20 minutos se duplicarán para dar lugar a 4 bacterias. Al llegar a la hora, es decir luego de 3 ciclos de duplicaciones, tendremos $2^3 = 8$ bacterias. Claramente, este número continúa siendo insignificante. En dos horas habrá 6 ciclos de duplicación, por lo que tendremos apenas $2^6 = 64$ bacterias, y al cabo de 3 horas (9 ciclos) apenas $2^9 = 512$. La primera impresión es que a este ritmo va a ser difícil llegar algún día no a un billón, sino incluso a un millón de bacterias. Para calcular cuánto tiempo (en generaciones, t) nos va a llevar llegar hasta **nuestro primer millón** podemos hacer una cuenta sencilla

$$2^t = 1 \times 10^6 \Leftrightarrow \ln(2^t) = \ln(1 \times 10^6) \Leftrightarrow t = \frac{\ln(1 \times 10^6)}{\ln(2)} \approx 19{,}94 \text{ generaciones} \quad (7.1)$$

Si tenemos 3 generaciones por hora, $19{,}94/3 \approx 6{,}65$ horas. Sin duda sorprendente, ya que al cabo de 3 horas teníamos apenas 512 bacterias y poco más de otras 3 horas llegamos al millón de bacterias. A esta altura no nos resultaría llamativo imaginarnos que no queda tan lejos llegar a un billón. Si hacemos la misma cuenta que antes, pero ahora con 1×10^{12} bacterias, llegamos a que son necesarias $t = \frac{\ln(1 \times 10^{12})}{\ln(2)} \approx 39{,}86$ generaciones, que es equivalente a $39{,}86/3 \approx 13{,}3$ horas. Es decir, en poco más de 13 horas llegamos a un gramo de bacterias, sin duda una proeza considerando que partimos de una bacteria, pero nada para asustarse. En función de esto, veamos si podemos calcular la masa de bacterias al cabo de un día. Un día tiene 24 horas y hay 3 duplicaciones por hora, por lo que al cabo de un día habrá 72 ciclos de duplicaciones. Por lo tanto, el número de bacterias será $2^{72} = 4{,}722 \times 10^{21}$. Si multiplicamos este número por la masa de una bacteria y lo dividimos entre los gramos que hay en una tonelada, tenemos que la masa al cabo de un día será ¡$\frac{(4{,}722 \times 10^{21} \text{ bacterias/día}) \times (1 \times 10^{-12} \text{ g/bacteria})}{1 \times 10^6 \text{ g/tonelada}} = 4722{,}4$ toneladas por día![9] Si tenemos en cuenta que un contenedor de 20 pies admite una carga de 21,92 toneladas, si fuésemos

[9] Los números exponenciales crecen de forma extremadamente rápida, y la intuición del cerebro humano suele ser especialmente mala en anticipar esta forma de crecimiento. Existen varios ejemplos referentes a esta subestimación del crecimiento exponencial (el ejemplo del crecimiento bacteriano mostrado en esta sección, sin ir más lejos). La persona interesada puede buscar la célebre leyenda de Sissa, referente al supuesto origen del ajedrez, que ilustra esta enseñanza.

a exportar la producción de bacterias de un día requeriría de $4722{,}4/21{,}92 \approx 215$ de estos contenedores. De hecho, el **HMM Algeciras**, el mayor barco portacontenedores del mundo carga 23.964 contenedores de 20 pies y alcanzaría con menos de 27 horas para completarlo. ¡Verdaderamente impresionante!

Los cálculos que realizamos previamente claramente no son razonables a la luz de nuestra experiencia o aun de extender las cuentas previas un poquito más: en 99 generaciones, 4 días y 3 horas, este proceso generaría una masa de bacterias que es mayor a la biomasa total existente en la tierra (del orden de 550×10^9 toneladas de carbono unido orgánicamente), un total sinsentido. ¿Dónde está el problema entonces? Obviamente, las condiciones que asumimos nosotros no son razonables en la vida real. La primera y más obvia es la necesidad de mantener las condiciones para que los nutrientes no sean una limitación al crecimiento y duplicación; a esto se suman otras varias limitaciones, como las relacionadas a la competencia e interacción entre los organismos (imaginemos el problema técnico de hacerles llegar el alimento a las primeras bacterias a medida que la masa y el volumen del cultivo crece), que sin duda llevan a quedar lejos del óptimo de crecimiento a medida que este se produce. De alguna manera, debemos buscar modelos más realistas, que nos permitan entender la dinámica de las poblaciones bacterianas.

En general, las poblaciones bacterianas como las de casi cualquier otro organismo, experimentan durante su vida un proceso que se repite una y otra vez, lo cual nos permite generar modelos para entender la dinámica de las mismas. La base de entender la dinámica de una población consiste en comprender cómo cambia el número de individuos a medida que pasa el tiempo. Normalmente, el estado de una población depende del número de individuos en estados previos de la misma, particularmente en el estado anterior inmediato. Para modelar esa dependencia existen dos alternativas respecto a la forma en que consideramos el tiempo: discreto y continuo.

En la figura 7.2 podemos ver una conceptualización sencilla del ciclo de vida simple de una población bacteriana en un modelo de tiempo discreto. En este caso decimos «simple» ya que no estamos considerando en el mismo otros mecanismos/estados (por ejemplo, **esporulación)**. Tiempo discreto se refiere a que manejamos el mismo como si fuese una variable discreta, por ejemplo, generaciones no-solapantes. En este caso, la mayor parte del problema suele concentrarse en entender cómo será el número de individuos en la generación siguiente, $n(t+1)$, si en la generación t existen $n(t)$ individuos. El círculo de la figura representa que se trata de un ciclo, el ciclo de vida, que se repetirá una y otra vez en la población.

Partiendo de la parte superior del círculo, el recorrido lo haremos en sentido horario e iremos agregando primas ($'$) para caracterizar la población luego de cada evento. Arrancamos con $n(t)$ individuos, y en la primera fase del círculo los individuos (bacterias, por ejemplo) se reproducirán a una tasa b (por *births*, nacimientos en inglés), agregando a la población $b\ n(t)$ individuos. Como antes había $n(t)$, si sumamos los recién nacidos tenemos ahora $n'(t) = n(t) + b\ n(t) = (1+b)\ n(t)$ individuos en la población. El siguiente evento será la muerte, que si ocurre a una tasa d (por *deaths*, muertes en inglés) entonces removerá $d\ n'(t) = d\ (1+b)\ n(t)$ individuos. Como antes teníamos $n'(t) = (1+b)\ n(t)$ individuos, si le restamos estos últimos nos quedaremos con $\dot{nz}(t) = n'(t) - d\ n'(t) = (1+b)\ n(t) - d\ (1+b)\ n(t) = (1-d)(1+b)\ n(t)$. Pero $\dot{nz}(t)$ marca el final del ciclo, por lo que se trata también del inicio del ciclo próximo (ya que no hay eventos entre fin e inicio), es decir $n(t+1)$. Por lo tanto, hemos arribado a la siguiente **relación de recurrencia** para el número de individuos en la población

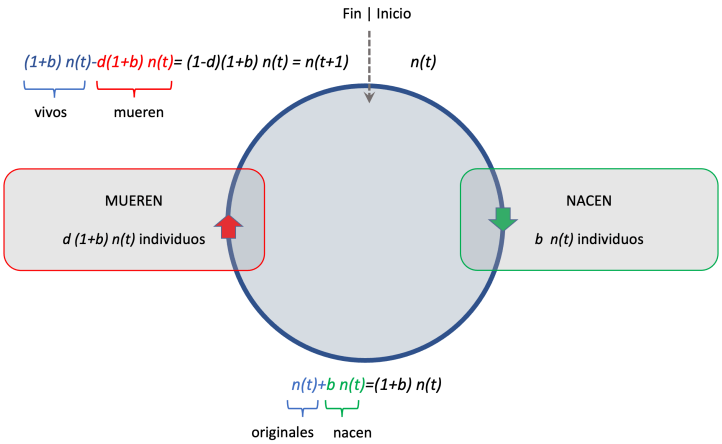


Figura 7.2. Modelo discreto del ciclo de vida simple de una bacteria. El ciclo gira en sentido horario, comenzando en la parte superior con $n(t)$ individuos. Luego de un evento (flecha verde) en que cada individuo deja un número b de individuos adicionales, el número de individuos que llega a la parte inferior del círculo es $n(t) + b \times n(t) = (1 + b) \times n(t)$. El siguiente evento (flecha roja) es la muerte, donde una proporción d de los individuos se muere, resultando en $b \times (1 - d) \times n(t)$ individuos muertos, que si los restamos de los vivos me permite llegar al final del ciclo con $n(t+1) = (1 + b) \times n(t) - d\,(1 + b) \times n(t) = (1 - d) \times (1 + b) \times n(t)$ individuos.

$$n(t + 1) = (1 - d) \times (1 + b) \times n(t) \tag{7.2}$$

Si definimos $R = (1 - d) \times (1 + b)$ como el **factor reproductivo**, entonces la ecuación de recursión (7.2) se simplifica a

$$n(t + 1) = R \times n(t) \tag{7.3}$$

El **factor reproductivo** es el número de individuos sobreviviente por **progenitor** (un abuso de lenguaje en nuestro caso, ya que las bacterias se reproducen por fisión). Haciendo la expansión de los componentes del **factor reproductivo** tenemos

$$R = (1 - d)(1 + b) = 1 + b - d - bd \tag{7.4}$$

La interpretación de la ecuación (7.4) es sencilla. El 1 es el factor que representa los organismos originales, ya que al multiplicar por este número nos da $1 \times n(t) = n(t)$. El b representa la fracción de organismos original que nacen en ese ciclo, mientras que d representa la fracción de los que mueren. Finalmente, el término bd representa la proporción de los que nacen en el ciclo que también mueren en ese ciclo.

Normalmente, para la mejor interpretación de la dinámica de la población nos interesa ver el cambio del número de individuos entre dos generaciones, es decir Δn. Para ello nos basta con restar $n(t)$ a $n(t+1)$, es decir, usando la ecuación (7.3), si le restamos $n(t)$ tenemos

$$\Delta n = n(t+1) \; - \; n(t) = R\, n(t) \; - \; n(t) = n(t)\,(R-1) \tag{7.5}$$

Al número $(R-1) = b - d - bd = r$ es el **cambio *per cápita* en el número de individuos de una generación a la siguiente,** y tiene una interpretación obvia en función de los componentes que lo definen, las tasas de nacimientos y muertes por generación.

Una forma alternativa de analizar la dinámica de las poblaciones es a partir de los modelos continuos. Ahora, a diferencia del caso de los modelos discretos, como el intervalo de tiempo considerado es infinitesimal, no tiene importancia el orden en los componentes del ciclo y además nunca van a ocurrir dos eventos en el mismo instante infinitesimal (en nuestro caso nacimiento y muerte, del que daba cuenta el término $-bd$). Para derivar la ecuación del modelo continuo podemos partir de la correspondiente ecuación del modelo discreto e ir reduciendo el intervalo de tiempo para pasar al límite. En particular, si consideramos un intervalo de tiempo arbitrariamente pequeño Δt (que podríamos ver como una fracción de la generación), entonces el número de individuos nacidos en ese intervalo de tiempo será $b \times \Delta t$, mientras que el número de muertes será $d \times \Delta t$. Por lo tanto, a partir de la definición de derivada tenemos

$$\frac{dn}{dt} = \lim_{\Delta t \to 0} \frac{[n(t+1) - n(t)]}{\Delta t} = \lim_{\Delta t \to 0} \frac{[(1 - d \times \Delta t)(1 + b \times \Delta t) \times n(t) - n(t)]}{\Delta t} \; \therefore \tag{7.6}$$

$$\frac{dn}{dt} = \lim_{\Delta t \to 0} \frac{[1 - d \times \Delta t + b \times \Delta t - bd(\Delta t)^2 - 1] \times n(t)}{\Delta t} = \lim_{\Delta t \to 0} [b - d - bd \times \Delta t]\; n(t) \tag{7.7}$$

En el límite, cuando $\Delta t \to 0$, $-bd \times \Delta t \to 0$, por lo que (dejando implícito que n es función de t) la ecuación (7.7) queda finalmente como

$$\frac{dn}{dt} = (b - d) \times n = r_c\, n \tag{7.8}$$

con $r_c = b - d$ la tasa neta de crecimiento para el modelo continuo (de ahí la c como subscrito).

Como la ecuación (7.8) involucra una variable n que es función de otra t, pero también aparecen derivadas de la función, entonces se trata de una ecuación diferencial, que como solo involucra la derivada primera es de primer orden y su solución es sencilla ya que es separable (es decir, podemos dejar agrupados en cada lado de la ecuación las variables y sus correspondientes diferenciales)

$$\frac{dn}{dt} = r_c n \Rightarrow \frac{dn}{n} = r_c \times dt \tag{7.9}$$

$$\int_{n(0)}^{n(t)} \frac{1}{n} \times dn = r_c \int_0^t dt \tag{7.10}$$

$$\ln n\big|_{n(0)}^{n(t)} = r_c \times t\big|_0^t \tag{7.11}$$

$$\ln(n(t)) - \ln(n(0)) = r_c \times t \tag{7.12}$$

$$e^{[\ln(n(t)) - \ln(n(0))]} = e^{r_c \times t} \tag{7.13}$$

$$\frac{n(t)}{n(0)} = e^{r_c\ t} \therefore \tag{7.14}$$

$$n(t) = n(0) \times e^{r_c\ t} \tag{7.15}$$

La ecuación (7.15) deja en claro que la dinámica de crecimiento poblacional bajo este modelo es exponencial. Si $r_c > 0$, es decir $b > d$ (mayor tasa de nacimientos que de muertes), entonces la población crecerá de forma exponencial. Si en cambio $r_c < 0$, es decir $b < d$ (menor tasa de nacimientos que de muertes), entonces, como es de esperar, la población decrecerá en tamaño y lo hará de forma exponencial.

En resumen, hemos modelado el crecimiento de las poblaciones bacterianas imaginando un ciclo de vida para tiempo discreto, donde los eventos siguen un orden determinado, que suele ser importante para el resultado de la dinámica de la población, así como otro modelo que considera el tiempo como una variable continua y que por lo tanto me permite describir la evolución a partir de una ecuación diferencial. El único parámetro relevante en este modelo es $r_c = b - d$, por lo que el comportamiento del mismo es poco flexible (la forma solo cambia cuando es positivo respecto a cuando es negativo) para describir situaciones reales.

Ejemplo 7.1

Una asistente de laboratorio decide, en un acto de subversión, dejar de lado la productividad y divertirse haciendo «carreras de crecimiento» entre poblaciones microbianas. Para ello genera en paralelo dos colonias bacterianas: una de ellas se encuentra (la de crecimiento más rápido) es fundada con tres células, mientras que la otra la inicia con cuatro para contar con cierta ventaja inicial. ¿Qué tan grande es la diferencia de número de células entre ambas luego de media hora de procrastinación en el laboratorio? Asuma un modelo continuo, con una tasa neta de crecimiento de $r_c = 3{,}632 \times 10^{-3}\text{min}^{-1}$ para la primera colonia y de

$r_c = 3{,}2630 \times 10^{-3}\text{min}^{-1}$ para la segunda. Se asume que no existen limitantes para el crecimiento bacteriano.

Utilizando la relación $n(t) = n(0) \times e^{r_c\ t}$ estimemos la relación entre ambas poblaciones celulares luego de media hora. Dividiremos el número de células en la segunda población por el de la primera, teniendo en cuenta que se espera existan más células en la colonia con mayor tamaño al momento de fundarse. Notemos que la constante r_c se encuentra en unidades de min^{-1}, por lo que tendremos en cuenta $t = 30$ min

$$\frac{n_2(30\ \text{min}^{-1})}{n_1(30\ \text{min}^{-1})} = \frac{3 \times e^{3{,}632 \times 30}}{4 \times e^{3{,}2630 \times 30}} \approx 48.161 \tag{7.16}$$

Vemos que una sutil diferencia en la tasa neta de crecimiento basta para que en media hora la población más veloz consiga ser cuarenta y ocho mil veces mayor que la otra, aun iniciando con una célula menos. Un gran día en el laboratorio, sin lugar a dudas.

El modelo logístico

Claramente, como vimos antes, el modelo exponencial de crecimiento nos lleva a resultados sorprendentes y que no se ajustan a nuestra experiencia o a nuestras observaciones empíricas (nuestro universo no está dominado por una sola especie de bacteria que sigue creciendo *ad infinitum*). Esto nos lleva a plantearnos la necesidad de establecer modelos más ajustados a la realidad. Una de las primeras limitaciones que se pueden notar en el modelo exponencial es la imposibilidad de modelar con este los cambios en la tasa de crecimiento de la población a medida que la misma evoluciona en el tiempo. Las implicancias de esto a nivel biológico son claras y difíciles de justificar: a medida que las poblaciones crecen también crece la demanda de recursos por parte de las mismas, pero el ambiente en el que viven tiene una capacidad limitada, que en algún punto se alcanzará y a partir del cual no será sustentable. Dicho en otras palabras, el sistema tiene una **capacidad de carga** determinada, que no se puede sobrepasar.

En términos de nuestros parámetros definidos previamente para el modelo exponencial, el que la tasa de crecimiento de la población dependa del tamaño de la misma, podemos explicitarlo utilizando una función $R(n)$ y asumiendo que se trata de una función que es decreciente en n (a medida que aumenta el tamaño de la población, menor va a ser R). Existen diversas funciones que cumplen con este requisito y que han sido utilizadas para modelar este fenómeno, pero una de las más usadas es la función lineal que nos lleva al **modelo logístico** de crecimiento. En este modelo se asume que la tasa de individuos que sobreviven por cada «progenitor» decrece de forma lineal a medida que aumenta el tamaño de la población.

En la figura 7.3 se puede apreciar el comportamiento del **factor reproductivo** $R(t) = (1 + r_d) - \frac{r_d}{K} n(t)$ (ordenadas) como función del tamaño poblacional (abscisas). Si llamamos K a la **capacidad de carga** del sistema. Cuando el tamaño poblacional es muy bajo en relación con la capacidad de carga (es decir, $\frac{n(t)}{K} \to 0$) entonces $\frac{r_d}{K} n(t) \to 0$ ya que r_d es una constante y, por lo tanto $R(n \to 0) = (1 + r_d)$, el intercepto de la recta. Dado que $\frac{n(t)}{K} \to 0$ acontece cuando no existe competencia por los recursos, a r_d le llamaremos la tasa intrínseca de crecimiento ($r_d > 0$) o decrecimiento ($r_d < 0$) de la población.

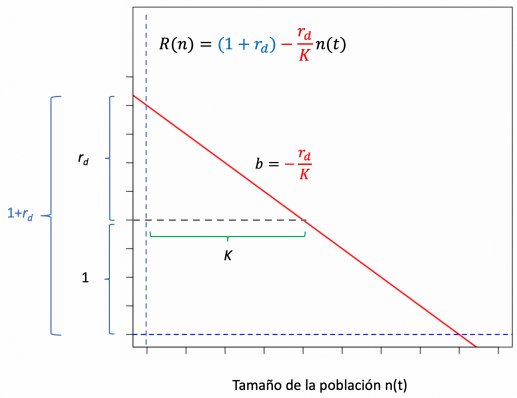


Figura 7.3. El factor reproductivo $R(n)$ como función lineal del tamaño poblacional. La función lineal $R(n) = (1 + r_d) - \frac{r_d}{K} n(t)$ podemos entenderla como la suma de intercepto igual a $(1 + r_d)$, es decir el factor reproductivo cuando no existen limitaciones impuestas por la población (ya que $n(t) = 0$) y un cambio en función de $n(t)$ con pendiente $b = -\frac{r_d}{K}$. Elaboración propia sobre idea de Otto y Day (2007).

Si expandimos y reordenamos los términos, tenemos que

$$R(n) = (1 + r_d) - \frac{r_d}{K} n(t) \Leftrightarrow R(n) = 1 + r_d[1 - \frac{n(t)}{K}] \tag{7.17}$$

$$1 + r_d[1 - \frac{n(t)}{K}] \tag{7.18}$$

y sustituyendo este resultado en la ecuación (7.3) tenemos

$$n(t+1) = R \times n(t) = n(t) \times [1 + r_d(1 - \frac{n(t)}{K})] = n(t) + n(t) \times r_d \times (1 - \frac{n(t)}{K}) \tag{7.19}$$

Más aún, el pasaje al cambio de frecuencia entre generaciones para este modelo es inmediato al restarle $n(t)$ a ambos lados de la ecuación (7.19)

$$\Delta n = n(t+1) - n(t) = n(t) + n(t) r_d(1 - \frac{n(t)}{K}) - n(t) = n(t) r_d(1 - \frac{n(t)}{K}) \tag{7.20}$$

De forma análoga, para el modelo continuo la tasa de cambio en la población queda determinada por

$$\frac{dn}{dt} = n(t)\, r_c(1 - \frac{n(t)}{K}) \tag{7.21}$$

En este punto ya nos debe resultar muy claro que la diferencia entre el modelo exponencial de crecimiento y el modelo logístico se encuentra en el factor $(1 - \frac{n(t)}{K})$ que aparece multiplicando en el modelo logístico. La interpretación de este factor es muy sencilla analizando los extremos. Cuando la población es muy pequeña $\frac{n(t)}{K} \to 0$ y por lo tanto $\lim_{n(t)\to 0}(1 - \frac{n(t)}{K}) = 1$, por lo que el modelo queda igual al modelo exponencial y el crecimiento seguirá ese patrón. Por otro lado, a medida que la población crece, llegará un punto en que su tamaño de acercará a la **capacidad de carga** del sistema y por lo tanto $\lim_{n(t)\to K}(1 - \frac{n(t)}{K}) = 0$ (ya que $\lim_{n(t)\to K} \frac{n(t)}{K} = 1$) y, por lo tanto, el crecimiento será nulo.

Al ser (7.21) una ecuación diferencial de primer orden y separable, relativamente sencilla, podemos obtener una solución para la misma al integrarla. Para facilitar la visualización llamaremos n_0 a $n(0)$ (número inicial de individuos a $t = 0$), y análogamente llamaremos $n_t = n(t)$ al número de individuos en tiempo t

$$\begin{aligned} \frac{dn}{dt} = n \times r_c \times (1 - \frac{n}{K}) \ \therefore \ \int_{n_0}^{n_t} \frac{1}{n(1 - \frac{n}{k})} \times dn = r_c \int_0^t dt \ \therefore \\ -K \int_{n_0}^{n_t} \frac{1}{n(n - K)} \times dn = r_c \, t \end{aligned} \tag{7.22}$$

Resolvamos primero la integral indefinida de la izquierda y luego le colocaremos los límites de integración. Sacando el factor común n^2 en el denominador del lado izquierdo de la ecuación (7.22) tenemos que

$$-K \int \frac{1}{(1 - \frac{K}{n})n^2} dn = r_c \, t \tag{7.23}$$

Si hacemos el cambio de variables $u = 1 - \frac{K}{n}$, entonces

$$\frac{du}{dn} = \frac{K}{n^2} \Leftrightarrow dn = \frac{n^2}{K} \times du \tag{7.24}$$

y sustituyendo en la ecuación (7.23), tenemos ahora

$$\frac{-K}{K} \int \frac{1}{u} \times du = r_c \, t \therefore -\ln(u) = r_c \, t \tag{7.25}$$

Haciendo el cambio de variable hacia atrás y poniendo los límites de integración ahora tenemos

$$
\begin{aligned}
&-\ln\left(1-\frac{K}{n}\right)\Big|_{n_0}^{n_t} = r_c t \Leftrightarrow \ln\left(1-\frac{K}{n_0}\right) - \ln\left(1-\frac{K}{n_t}\right) = r_c t \\
&\Leftrightarrow \ln\left[\frac{1-\frac{K}{n_0}}{1-\frac{K}{n_t}}\right] = r_c t \Leftrightarrow \ln\left[\frac{(n_0-K)n_t}{(n_t-K)n_0}\right] = r_c t \\
&\Leftrightarrow \frac{(n_0-K)n_t}{(n_t-K)n_0} = e^{r_c t} \Leftrightarrow n_t = \frac{(n_t-K)n_0 e^{r_c t}}{n_0-K} \\
&\Leftrightarrow n_t = \frac{n_t n_0 e^{r_c t} - K n_0 e^{r_c t}}{n_0-K} \Leftrightarrow n_t\left(1-\frac{n_0 e^{r_c t}}{n_0-K}\right) = -\frac{K n_0 e^{r_c t}}{n_0-K}
\end{aligned}
$$

$$
n_t \times [n_0 - K - n_0 \times e^{r_c\ t}] = -K n_0 \times e^{r_c\ t} \Leftrightarrow n_t = \frac{-K n_0 \times e^{r_c\ t}}{n_0 - K - n_0 \times e^{r_c\ t}} \tag{7.26}
$$

y finalmente, multiplicando y dividiendo por -1, tenemos la forma final

$$
n(t) = \frac{K \times n_0 \times e^{r_c\ t}}{K + n_0 \times (e^{r_c\ t} - 1)} \tag{7.27}
$$

Ejemplo 7.2

Calcular el tamaño poblacional para una colonia bacteriana que arranca con 5 individuos al cabo de un tiempo $t = 1$ hora y con una tasa neta de crecimiento para el modelo continuo de $r_c = 4{,}1667 \times 10^{-3} s^{-1}$ en el caso de que la capacidad de carga del sistema (K) sea de 5 millones de individuos.

Como $t = 1$ hora, expresado en segundos (ya que las unidades de r_c están en s^{-1}) sería $t = 3600s$. Usando el resultado de la ecuación (7.27), tenemos que $n(t) = \frac{K\ n_0\ e^{r_c\ t}}{K + n_0\ (e^{r_c\ t} - 1)}$ por lo que sustituyendo por los valores del problema, llegamos a $n(3600) = \frac{5\times10^6 \times 5 \times e^{4{,}1667\times10^{-3}\times 3.600}}{5\times10^6 + 5\times(e^{4{,}1667\times10^{-3}\times 3.600} - 1)} = 3.828.879$. Es decir, aproximadamente 3,8 millones de bacterias, aún relativamente lejos de la capacidad de carga del sistema.

En la figura 7.4 podemos apreciar el crecimiento de una población que arranca con un individuo y posee una capacidad de carga del sistema de $K = 10.000$ individuos, de acuerdo con el modelo logístico.

La **tasa intrínseca de crecimiento** es de $r_c = 0{,}25$ en la curva anaranjada, de $r_c = 0{,}20$ en la curva negra, mientras que es de $r_c = 0{,}15$ en la curva azul. Claramente, se aprecia en las tres curvas que al comienzo el comportamiento es aproximadamente exponencial, seguido por una fase **casi-lineal**, en el medio de la cual cambia la concavidad, para finalmente acercarse a una fase asintótica donde se aproxima lentamente a la capacidad de carga. A medida

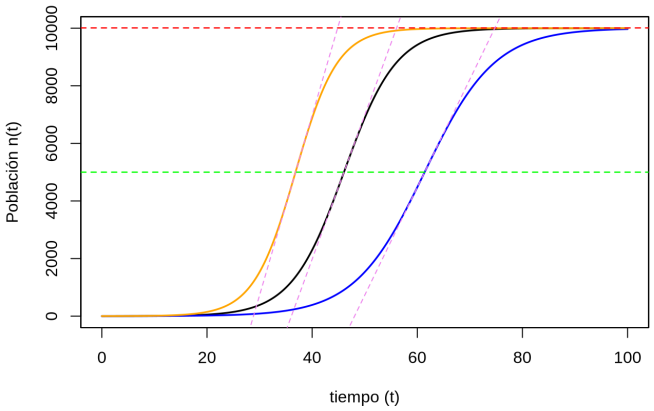


Figura 7.4. Evolución de la población de acuerdo con el modelo logístico, con parámetros $n(0) = 1$, $K = 10.000$ individuos y $r_c = 0{,}25$ (curva anaranjada), $r_c = 0{,}20$ (curva negra), $r_c = 0{,}15$ (curva azul). La línea roja indica la **capacidad de carga** del sistema, mientras que la línea verde se encuentra en $K/2$. Las líneas violetas segmentadas representan las pendientes en los correspondientes puntos de inflexión para las 3 curvas.

que la tasa intrínseca de crecimiento es mayor (azul, negra, anaranjada) el crecimiento es más explosivo en la fase exponencial, marcado a su vez por una pendiente mucho mayor en la fase casi-lineal y una entrada más abrupta en la fase asintótica. El modelo logístico es también conocido como **modelo de Verhulst** en honor del matemático belga que lo reportó por primera vez.[10]

Para determinar el tiempo que nos llevará llegar hasta la mitad de la capacidad de carga del sistema basta con sustituir $n_t = \frac{K}{2}$ en la ecuación (7.27) y despejar t,

$$\begin{aligned} n(t) = \frac{K}{2} = \frac{K \times n_0\, e^{r_c\, t}}{K + n_0 \times (e^{r_c\, t} - 1)} \Leftrightarrow 2n_0\, e^{r_c\, t} = K + n_0 \times (e^{r_c\, t} - 1) \Leftrightarrow \\ 2 \times n_0 \times e^{r_c\, t} - n_0 \times e^{r_c\, t} = K - n_0 \Leftrightarrow e^{r_c\, t} = \frac{K - n_0}{n_0} \Leftrightarrow \\ r_c \times t = \ln(\frac{K - n_0}{n_0}) \Leftrightarrow \\ t = \frac{\ln(\frac{K-n_0}{n_0})}{r_c} \end{aligned} \tag{7.28}$$

Aplicando el resultado de la ecuación (7.28) a los datos de las tres curvas de la figura 7.4, llegamos a que $t = \frac{\ln((10.000-1)/1)}{0{,}25} = \frac{9{,}21024}{0{,}25} = 36{,}84$ unidades de tiempo para la curva anaranjada, $t = \frac{9{,}21024}{0{,}20} = 46{,}05$ unidades para la curva negra, mientras que para la curva

[10] Pierre François Verhulst (28 de octubre de 1804 - 15 de febrero de 1849) fue un matemático belga, conocido como el primer proponente del modelo logístico para el crecimiento de las poblaciones como alternativa al modelo exponencial de Malthus.

azula harán falta $t = \frac{9{,}21024}{0{,}15} = 61{,}40$ unidades de tiempo para alcanzar la mitad de la capacidad de carga del sistema.

Las curvas de la figura 7.4 muestran que en algún punto de la fase casi-lineal tiene que existir un cambio en la convexidad-concavidad de la curva, es decir pasar de una tasa de crecimiento poblacional creciente en el tiempo a una fase en que **la tasa es decreciente**. Como recordarás de los cursos de cálculo, ese cambio (de existir) ocurre en uno de los ceros de la derivada segunda de la curva respecto al tiempo. Si bien la ecuación (7.21) ya nos da la derivada primera, debemos sustituir en la misma $n(t)$ por el resultado de la ecuación (7.27) y volver a derivar respecto al tiempo, lo cual no es difícil pero sí requiere cierto número de operaciones, lo cual vuelve al proceso tedioso.

Sin embargo, podemos llegar a este resultado de una manera más sencilla. En lugar de derivar dos veces la ecuación (7.27), podemos recurrir a la ecuación (7.21), donde ya está presente la derivada primera respecto al tiempo (es decir, la tasa de crecimiento de la población respecto al tiempo). En la figura 7.5 se observa que la tasa de crecimiento aumenta con el tamaño poblacional hasta determinado punto donde empieza a bajar, lo cual implica que existe un máximo de la derivada de dn/dt respecto a n (es decir, un máximo de $\frac{\partial^2 n}{\partial n \partial t}$); idealmente, este sería el punto de inflexión la gráfica de $n(t)$.

En función de esto, si volvemos a derivar la ecuación (7.21), pero ahora respecto a n, obtenemos

$$\frac{\partial^2 n}{\partial n \partial t} = \frac{\partial[r_c \times n \times (1 - \frac{n}{K})]}{\partial n} = \frac{\partial[r_c \times (n - \frac{n^2}{K})]}{\partial n} = r_c - \frac{2r_c}{K} n \tag{7.29}$$

Por lo tanto, igualando a cero, tenemos ahora

$$\frac{\partial^2 n}{\partial n \partial t} = r_c - \frac{2r_c}{K} n = 0 \Leftrightarrow n = \frac{K r_c}{2r_c} = \frac{K}{2} \tag{7.30}$$

Es decir, el punto de inflexión en la curva de crecimiento poblacional se alcanzará cuando el tamaño poblacional sea **igual a la mitad de la capacidad de carga del sistema** y, por lo tanto, el tiempo hasta ese punto estará dado por la ecuación (7.28). Si sustituimos ese valor en la ecuación (7.21) que define la derivada primera, es decir la tasa de crecimiento en el punto de inflexión, tendremos que la misma es igual a

$$\frac{dn}{dt}\Big|_{n=K/2} = r_c \times (\frac{K}{2}) \times (1 - \frac{K}{2K}) = \frac{K r_c}{4} \tag{7.31}$$

Por ejemplo, con los números de la figura 7.4, para la curva anaranjada la máxima tasa de crecimiento de la población será $\frac{dn}{dt}\big|_{n=K/2} = \frac{K r_c}{4} = \frac{10.000 \times 0{,}25}{4} = 625$ individuos por unidad de tiempo (la unidad en la que se expresa r_c). Para la curva negra tenemos $\frac{dn}{dt}\big|_{n=K/2} = \frac{K r_c}{4} = \frac{10.000 \times 0{,}20}{4} = 500$ individuos por unidad de tiempo. Finalmente, para la curva azul en cambio $\frac{dn}{dt}\big|_{n=K/2} = \frac{K r_c}{4} = \frac{10.000 \times 0{,}15}{4} = 375$ individuos por unidad de tiempo. Estas tres pendientes se pueden apreciar en color violeta en la figura 7.4.

El modelo logístico no es el único que presenta características razonables para describir el comportamiento de las poblaciones. De hecho, diversas funciones que comparten la característica de ser **sigmoidales** (con forma de S), son capaces de describir el comportamiento deseado. Por ejemplo, Zwietering y colaboradores comparan diversos modelos usados frecuentemente para describir la dinámica poblacional, incluyendo el logístico, la **función de Gompertz**, así como los modelos de Richards, Stannard y Schnute (todos estos de 4 parámetros), llegando a la conclusión de que en general el modelo logístico, con 3 parámetros ($n(0)$, r_c y K), es suficiente para describir adecuadamente las poblaciones (Zwietering et al. 1990).

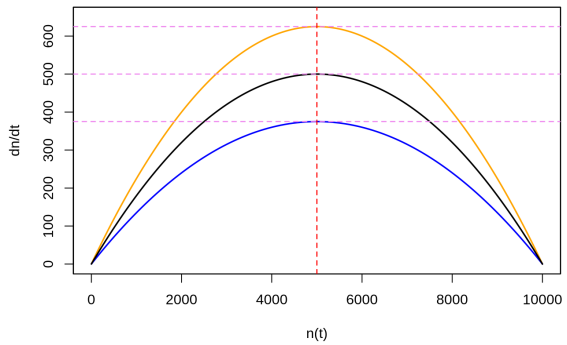


Figura 7.5. Tasas de crecimiento en función del tamaño de la población para las tres poblaciones de la figura 7.4. Se observa claramente que se trata de una curva parabólica con un máximo en el centro. En violeta los valores máximos, que coinciden con las respectivas pendientes en el punto de inflexión en la figura 7.4.

PARA RECORDAR

- En el caso del modelo exponencial de crecimiento poblacional $n(t+1) = (1-d)(1+b)\, n(t)$.
- Si definimos $R = (1-d)(1+b)$ como el **factor reproductivo** (el número de individuos sobreviviente por **padre**), entonces la ecuación de recursión se simplifica a $n(t+1) = R\, n(t)$.
- Normalmente, para la mejor interpretación de la dinámica de la población nos interesa ver el cambio del número de individuos entre dos generaciones, es decir Δn, que es $\Delta n = n(t+1) - n(t) = R\, n(t) - n(t) = n(t)\,(R-1)$ Además, $r = (R-1) = b - d - bd$ es **el cambio per cápita en el número de individuos** de una generación a la siguiente.
- En el caso del modelo exponencial en tiempo continuo, tenemos que $\frac{dn}{dt} = (b-d)\, n = r_c\, n$, con $r_c = b - d$ la tasa neta de crecimiento para el modelo continuo.
- La solución de la ecuación diferencial que representa la tasa de cambio en la población nos da el tamaño poblacional en función del tiempo t, la tasa de crecimiento r_c y el número inicial de individuos y tiene la forma $n(t) = n(0) \times e^{r_c\, t}$.

- El modelo exponencial no da cuenta adecuadamente de las condiciones de restricción a las que normalmente están sujetas las poblaciones. Una alternativa es el modelo logístico, donde $R(n)$ es una función lineal de la densidad de población $R(n) = 1 + r_d[1 - \frac{n(t)}{K}]$. Al parámetro K le llamamos **capacidad de carga** del sistema y representa el número máximo de individuos que puede acomodar el sistema.
- En forma análoga, para el modelo continuo la tasa de cambio en la población queda determinada por $\frac{dn}{dt} = n(t)\, r_c(1 - \frac{n(t)}{K})$.
- La solución a la ecuación diferencial anterior nos da el número de individuos en el tiempo t a partir del número inicial de individuos n_0, la **capacidad de carga** del sistema K y la tasa de crecimiento r_c y tiene la siguiente forma $n(t) = \frac{K\ n_0\ e^{r_c\ t}}{K + n_0\ (e^{r_c\ t} - 1)}$.

7.3 Modelos haploides de selección natural

En el capítulo «Selección natural» estudiamos el efecto de las diferencias en **fitness** entre los genotipos para un locus **diploide** con dos alelos. En el caso de los procariotas y algunos eucariotas unicelulares los genomas son haploides, o basculan entre fases haploides y diploides (o aún poliploides). Es momento entonces de analizar la dinámica del cambio en las frecuencias alélicas en un locus **haploide**. El libro de Otto y Day (2007) presenta un tratamiento simple y claro del problema, y es una excelente referencia para entender mejor los distintos modelos de dinámica de poblaciones y selección, así como sus derivaciones, por lo que seguiremos de cerca su razonamiento.

Selección haploide: modelo discreto

Supongamos, por ejemplo, una población bacteriana en crecimiento. Supongamos además que tenemos un locus con dos alelos, que llamaremos A_1 y A_2, como es nuestra costumbre. Si en una generación t determinada el número de células bacterianas que tienen el alelo A_1 es $n_1(t)$ y el número de las que tienen el alelo A_2 es $n_2(t)$, de acuerdo con el modelo exponencial en la siguiente generación $(t + 1)$ el número de células de cada uno estará determinado por

$$\begin{aligned} n_1(t+1) &= W_1 \times n_1(t) \\ n_2(t+1) &= W_2 \times n_2(t) \end{aligned} \tag{7.32}$$

siendo

$$\begin{aligned} W_1 &= (1d_1)(1 + b_1) \\ W_2 &= (1d_2)(1 + b_2) \end{aligned} \tag{7.33}$$

los factores reproductivos para cada uno de los dos alelos. En este caso d_1 y d_2 representan la tasa de muerte de las células bacterianas por generación, mientras que b_1 y b_2 representan

los nacimientos, por duplicaciones normalmente (las letras b y d vienen del inglés, *births* y *deaths*, nacimientos y muertes respectivamente). La lógica es clara: en cada generación muere una proporción d (sobrevive $1-d$), y de esos que sobreviven por cada célula habrá b nuevas células, que junto a las presentes constituyen $1+b$. Poniendo juntos los términos de muertes y nacimientos tenemos $W=(1-d)(1+b)$, como aparece en la ecuación (7.33) para cada uno de los dos alelos o genotipos.

Como ya hemos visto antes, para pasar a la frecuencia de cada uno de los alelos alcanza con dividir el número de células de cada genotipo (que es el de portadores del alelo correspondiente, ya que son **haploides**) entre el número total de células, que es la suma de las de los dos genotipos en (nuestro caso, n_1+n_2). Por lo tanto, si llamamos p a la frecuencia de A_1 y q a la de A_2, entonces

$$\begin{aligned} p &= \frac{n_1}{n_1 \ + \ n_2} \\ q &= \frac{n_2}{n_1 \ + \ n_2} \end{aligned} \tag{7.34}$$

con $p+q=\frac{n_1}{n_1+n_2}+\frac{n_2}{n_1+n_2}=\frac{n_1+n_2}{n_1+n_2}=1$

Para analizar la dinámica en el tiempo de las frecuencias debemos tener claro a qué generación nos referimos en cada momento, es decir, debe quedar claro en nuestra notación que p y q son funciones del tiempo (en generaciones), por lo que los notaremos $p(t)$ y $q(t)$. Puede parecer obvio, pero por si acaso: esto no quiere decir que se multiplique p por t, sino que como usual en la notación de funciones solo es una forma de expresar que p cambia en función del valor de t.

De acuerdo con la ecuación (7.34), para la generación $(t+1)$ las frecuencias de los genotipos estarán dadas por

$$\begin{aligned} p(t+1) &= \frac{n_1(t+1)}{n_1(t+1) \ + \ n_2(t+1)} \\ q(t+1) &= \frac{n_2(t+1)}{n_1(t+1) \ + \ n_2(t+1)} \end{aligned} \tag{7.35}$$

pero, usando la ecuación (7.32), que relaciona el número de individuos en dos generaciones sucesivas, en la ecuación (7.35), tenemos que

$$p(t+1)=\frac{W_1\times n_1}{W_1\times n_1+W_2\times n_2} \tag{7.36}$$

Notemos que, si bien n_1 y n_2 son funciones de t y, por lo tanto, deberíamos escribirlas como $n_1(t)$ y $n_2(t)$, para evitar que la ecuación (7.36) sea más difícil de leer sacamos el (t) de la misma. La ecuación (7.36) nos expresa la frecuencia del alelo A_1 en la generación $(t+1)$ en función del número de individuos y sus **fitness** en la generación previa. Para llegar a una expresión que sea función de la frecuencia en la generación anterior, una alternativa sencilla consiste en multiplicar y dividir por n_1+n_2, es decir

$$p(t+1)=\frac{(\frac{1}{n_1+n_2})}{(\frac{1}{n_1+n_2})}\times\frac{W_1\times n_1}{W_1\times n_1 \ + \ W_2\times n_2}=\frac{W_1\times(\frac{n_1}{n_1+n_2})}{W_1\times(\frac{n_1}{n_1+n_2})+W_2\times(\frac{n_2}{n_1+n_2})} \tag{7.37}$$

Pero $\frac{n_1}{n_1 + n_2} = p(t)$ y $\frac{n_2}{n_1 + n_2} = q(t)$, por lo que la ecuación (7.37) se transforma en

$$p(t+1) = \frac{W_1\, p(t)}{W_1 \times p(t) \;+\; W_2 \times q(t)} = \frac{W_1 \times p(t)}{W_1 \times p(t) \;+\; W_2 \times (1 - p(t))} \tag{7.38}$$

La ecuación (7.38) depende de dos parámetros (W_1 y W_2) para definir las frecuencias alélicas en la próxima generación. Si recordamos de antes, W_1 y W_2 eran los factores reproductivos de cada uno de los dos alelos, es decir la tasa neta de crecimiento de cada uno de los dos alelos, que no es otra cosa que el **fitness absoluto** de cada uno de ellos (de ahí la notación W), del que habíamos hablado al comienzo del capítulo «Selección natural». Como vimos también en ese capítulo, al dividir el **fitness absoluto** de los distintos genotipos entre el de alguno de ellos se obtiene el **fitness relativo** de los mismos, que nuevamente notaremos w.

Con esta lógica, si se define $w_1 = W_1/W_2$ como el **fitness relativo** del alelo A_1 respecto del A_2 (con $W_1 = w_1 \times W_2$), sustituyendo en la ecuación (7.38) se llega a

$$p(t+1) = \frac{w_1 \times W_2 \times p(t)}{w_1 \times W_2 \times p(t) \;+\; W_2 \times (1 - p(t))} = \frac{w_1 \times p(t)}{w_1 \times p(t) \;+\; (1 - p(t))} \tag{7.39}$$

La importancia de la ecuación (7.39) radica en que ahora se cuenta con una expresión de la frecuencia del alelo A_1 (y por lo tanto del A_2, ya que $p+q = 1$ en cualquier generación) que solo depende de un parámetro: el **fitness relativo** de los dos alelos. Dicho de otra manera, ahora las frecuencias en la siguiente generación no dependen de los factores reproductivos en términos absolutos, sino de la relación entre ellos. Este resultado a la vez intuitivo y muy importante, ya que deja en claro que el cambio en frecuencias estará dado por el éxito de un genotipo sobre el otro, aún en condiciones muy diferentes.

Para entender esto basta con imaginarse una situación donde el ambiente cambia en el tiempo, afectando el éxito reproductivo de ambos genotipos de manera similar. Ejemplos de esto pueden ser el cambio de temperatura, la abundancia o escasez de nutrientes, la existencia de un patógeno, entre otros escenarios de relevancia biológica. Supongamos que llamamos a esta **función del ambiente** $\sigma(t)$, para dejar claro que se trata de una función que varía con el tiempo. Ahora, asumiendo que $\sigma(t)$ **afecta de la misma manera a ambos genotipos** (esto es fundamental), el nuevo número de descendientes por ancestro será para cada uno de los genotipos $W_1' = \sigma(t)\, W_1$ y $W_2' = \sigma(t)\, W_2$. Usando la misma notación para el **fitness relativo** al incluir este efecto ambiental

$$w_1' = \frac{W_1'}{W_2'} = \frac{\sigma(t) \times W_1}{\sigma(t) \times W_2} = \frac{W_1}{W_2} = w_1 \tag{7.40}$$

Este resultado es fundamental ya que, en la medida que los factores ambientales afecten de la misma manera a ambos genotipos, desacopla los factores ambientales o ecológicos del estudio de la dinámica de los genéticos.

El objetivo que nos habíamos planteado es entender la dinámica del cambio en las frecuencias alélicas, y para eso debemos entender cómo cambia de generación a generación y obtener una expresión que represente el cambio entre dos generaciones sucesivas. El cambio de la frecuencia del alelo A_1 entre dos generaciones será la diferencia entre $p(t+1)$ y $p(t)$ para un t dado, por lo que haciendo uso de la ecuación (7.38) y restándole $p(t)$ se llega a la siguiente relación

$$\begin{aligned} \Delta_s p = p(t+1) - p(t) = \frac{W_1 \times p(t)}{W_1 \times p(t) \ + \ W_2 \times q(t)} - p(t) \ \therefore \\ \Delta_s p = \frac{W_1 \times p(t) - p(t) \times [W_1 \times p(t) \ + \ W_2 \times q(t)]}{W_1 \times p(t) \ + \ W_2 \times q(t)} \end{aligned} \tag{7.41}$$

Notemos que volvimos a usar la misma notación que en el capítulo «Selección natural», $\Delta_s p$, con el subscrito $_s$ para marcar que se trata del cambio debido a la selección (aunque ahora se trata del caso de un locus **haploide**). Expandiendo los términos de la derecha, tenemos

$$\Delta_s p = \frac{W_1 \times p(t) - W_1 \times p^2(t) - W_2 \times p(t) \times q(t)}{W_1 \times p(t) \ + \ W_2 \times q(t)} = \frac{W_1 \times [p(t) - p^2(t)] - W_2 \times p(t) \times q(t)}{W_1 \times p(t) + W_2 \times q(t)} \tag{7.42}$$

Pero $p(t) - p^2(t) = p(t) \times [1 - p(t)] = p(t) \times q(t)$, por lo que sustituyendo en la ecuación (7.43) llegamos finalmente a la siguiente expresión para el cambio de la frecuencia de A_1 en una generación:

$$\begin{aligned} \Delta_s p &= \frac{W_1 \times p(t) - W_1 \times p^2(t) - W_2 \times p(t) \times q(t)}{W_1 \times p(t) + W_2 \times q(t)} \\ &= \frac{W_1 \times [p(t) - p^2(t)] - W_2 \times p(t) \times q(t)}{W_1 \times p(t) + W_2 \times q(t)} \end{aligned} \tag{7.43}$$

Si, siguiendo la notación de Otto y Day (2007), definimos el coeficiente de selección en favor del alelo A_1 como $s_d = (W_1 - W_2)/W_2$ (el subscrito $_d$ denota que se trata de una derivación a partir de tiempo discreto), entonces $(W_1 - W_2) = s_d \times W_2$ y $W_1 = W_2 \ + \ s_d \times W_2 \Leftrightarrow W_1 = W_2 \times (1 + s_d)$. Sustituyendo estas dos últimas expresiones en (7.3), tenemos

$$\begin{aligned} \Delta_s p = \frac{s_d \times W_2 \times p(t) \times q(t)}{W_2 \times (1 + s_d) \times p(t) \ + \ W_2 \times q(t)} \ \therefore \\ \Delta_s p = \frac{s_d \times W_2 \times p(t) \times q(t)}{W_2 \times [(1 + s_d) \times p(t) \ + \ q(t)]} \ \therefore \\ \Delta_s p = \frac{s_d \times p(t) \times q(t)}{(1 + s_d) \times p(t) \ + \ q(t)} \end{aligned} \tag{7.44}$$

Finalmente, expandiendo los términos en el denominador de (7.44) y recordando que $p(t)+q(t)=1$, se tiene

$$\Delta_s p = \frac{s_d \times p(t) \times q(t)}{p(t) + s_d \times p(t) \ + \ q(t)} = \frac{s_d \times p(t) \times q(t)}{1 + s_d \times p(t)} \tag{7.45}$$

o, retirando la notación de la generación t, que a esta altura es evidente, el cambio en la frecuencia del alelo A_1 será

$$\Delta_s p = \frac{s_d \times pq}{1 + s_d \times p} \tag{7.46}$$

La forma del cambio de frecuencia entre generaciones, dependiendo de la frecuencia del alelo A_1 y del coeficiente de selección s_d, se puede apreciar en la figura 7.6.

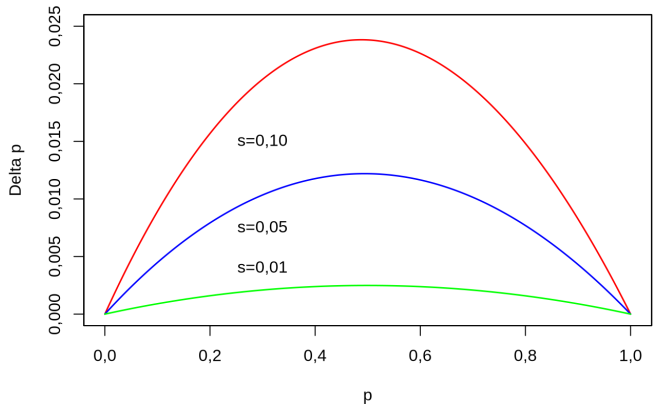


Figura 7.6. Cambio en la frecuencia del alelo A_1 ($\Delta_s p$) en función de la frecuencia del mismo (p) y del coeficiente de selección $s_d = (W_1 - W_2)/W_2$. En rojo $s_d = 0{,}10$, azul $s_d = 0{,}05$ y verde $s_d = 0{,}01$.

La forma luce como aparentemente parabólica aunque no lo sea, ya que el denominador aumenta a medida de aumenta p (pero a valores muy bajos de s_d, $1 + s_d \approx 1$). Esto nos hace recordar las formas de la selección estabilizadora. Sin embargo, esto se debe a que en la figura 7.6 solo manejamos valores de s_d positivos y relativamente bajos. Por ejemplo, si $s_d = 0{,}10 = (W_1 - W_2)/W_2 \Leftrightarrow W_1 = W_2 \times (1 + 0{,}10) \Leftrightarrow W_1 = 1{,}1 \times W_2$; es decir, el genotipo W_1 crecerá a una tasa un 10 % mayor que el W_2.

Sin embargo, para interpretar correctamente esta ecuación es importante notar la asimetría del coeficiente s_d respecto a los valores de W_1 y W_2. Un ejemplo claro consiste en intercambiar los valores de W_1 y W_2 y ver el resultado. Un caso extremo es cuando la ventaja de un genotipo respecto al otro es enorme, por ejemplo, $W_2 = 1$ y $W_1 = 0{,}01$. En esta situación,

$s_d = (W_1 - W_2)/W_2 = (0{,}01 - 1)/1 = -0{,}99$, pero si invertimos los valores se obtiene $W_1 = 1$ y $W_2 = 0{,}01$, por lo que $s_d = (W_1 - W_2)/W_2 = (1 - 0{,}01)/0{,}01 = 99$.

Una alternativa para la interpretación de $\Delta_s p$ es hacerlo a partir de la ecuación (7.44). El denominador es $W_1 \times p(t) + W_2 \times q(t) = \bar{W}(t)$, es decir el **fitness medio absoluto**, mientras que el numerador es $(W_1 - W_2) \times p(t) \times q(t)$, que se trata de la diferencia en los **fitness absolutos** de los dos genotipos, modulados por la mitad de la varianza genética: a frecuencias intermedias de p y q mayor será el efecto de la diferencia $(W_1 - W_2)$.

Ejemplo 7.3

En una especie bacteriana, un locus con dos alelos, A_1 y A_2 produce que los individuos que poseen los mismos presentan un **fitness absoluto** de $W_1 = 1{,}7$ y $W_2 = 1{,}2$. Si la frecuencia del alelo A_1 es $p = 0{,}45$, asumiendo que la división de las bacterias opera en tiempo discreto, determinar la frecuencia en la siguiente generación.

Una alternativa es usar la ecuación (7.41), es decir

$$\Delta_s p = p(t+1) - p(t) = \frac{W_1 p(t)}{W_1 p(t) + W_2 q(t)} - p(t) \tag{7.47}$$

por lo que, sustituyendo los valores tenemos

$$\Delta_s p = \frac{1{,}7 \times 0{,}45}{1{,}7 \times 0{,}45 + 1{,}2 \times (1 - 0{,}45} - 0{,}45 = 0{,}08684 \tag{7.48}$$

Por otro lado, si calculamos $s_d = (W_1 - W_2)/W_2 = (1{,}7 - 1{,}2)/1{,}2 = 0{,}41667$, entonces aplicando la ecuación (7.46) y sustituyendo los valores correspondientes tenemos

$$\Delta_s p = \frac{s_d\, pq}{1 + s_d\, p} = \frac{0{,}41667 \times 0{,}45 \times 0{,}55}{1 + 0{,}41667 \times 0{,}45} = 0{,}08684 \tag{7.49}$$

que es el mismo resultado obtenido previamente. En función de este resultado, la frecuencia $p(t+1) > p(t)$ y, por lo tanto, incrementará la frecuencia del alelo A_1, que tiene una ventaja en **fitness** respecto al A_2. Notemos que esta diferencia de **fitness** provoca un salto muy grande en las frecuencias alélicas en una sola generación (casi un 9 %).

Selección haploide: modelo continuo

Una forma alternativa de estudiar el cambio de frecuencias alélicas en un modelo de locus **haploide** con dos alelos es a partir del modelo de crecimiento continuo de la población. Como vimos antes, en este modelo las tasas de crecimiento instantáneo tienen la forma $r = b - d$, es decir la tasa de nacimientos menos la tasa de muertes. Supongamos entonces que los dos alelos (genotipos) A_1 y A_2 tienen tasas diferentes, las cuales expresaremos como

$$\begin{aligned} r_1 &= (b_1 - d_1) \\ r_2 &= (b_2 - d_2) \end{aligned} \tag{7.50}$$

Si, como antes, n_1 y n_2 representan los números de células (individuos) de cada uno de los dos genotipos, entonces el modelo de crecimiento exponencial nos permite describir el cambio en el número de células como las ecuaciones diferenciales

$$\begin{aligned} \frac{dn_1}{dt} &= r_1 \times n_1(t) \\ \frac{dn_2}{dt} &= r_2 \times n_2(t) \end{aligned} \tag{7.51}$$

Pero, de acuerdo con la ecuación (7.34), las frecuencias p y q (ambas funciones de t), son

$$\begin{aligned} p &= \frac{n_1}{n_1 + n_2} \\ q &= \frac{n_2}{n_1 + n_2} \end{aligned} \tag{7.52}$$

por lo que sustituyendo en la ecuación (7.51), tenemos que el cambio en la frecuencia del alelo A_1 será

$$\frac{dp}{dt} = \frac{d(\frac{n_1}{n_1 + n_2})}{dt} \tag{7.53}$$

Ahora, aplicando la regla del cociente para las derivadas (ver Apéndice B: Cálculo diferencial e integral), tenemos que

$$\frac{dp}{dt} = \frac{\frac{dn_1}{dt} \times (n_1 + n_2) - n_1 \times \frac{d(n_1 + n_2)}{dt}}{(n_1 + n_2)^2} \tag{7.54}$$

Pero la derivada de una suma (o resta) es la suma (o resta) de las derivadas, por lo que, expandiendo la ecuación (7.55), tenemos ahora

$$\frac{dp}{dt} = \frac{n_1 \times \frac{dn_1}{dt} + n_2 \times \frac{dn_1}{dt} - n_1 \times \frac{dn_1}{dt} - n_1 \times \frac{dn_2}{dt}}{(n_1 + n_2)^2} \therefore \quad \frac{dp}{dt} = \frac{n_2 \times \frac{dn_1}{dt} - n_1 \times \frac{dn_2}{dt}}{(n_1 + n_2)^2} \tag{7.55}$$

Pero como vimos antes, de acuerdo con la ecuación (7.51), $\frac{dn_1}{dt} = r_1 \times n_1$, mientras que $\frac{dn_2}{dt} = r_2 \times n_1$ (notando n_1 por $n_1(t)$ y n_2 por $n_2(t)$), por lo que sustituyendo en la ecuación (7.55), tenemos

$$\frac{dp}{dt} = \frac{n_2 \times (r_1 \times n_1) - n_1 \times (r_2 \times n_2)}{(n_1 + n_2)^2} = (r_1 - r_2) \times \frac{n_1 \times n_2}{(n_1 + n_2)^2} \tag{7.56}$$

Pero $\frac{n_1\, n_2}{(n_1 + n_2)^2} = \frac{n_1}{(n_1 + n_2)} \times \frac{n_2}{(n_1 + n_2)} = pq$, por lo que sustituyendo en la ecuación (7.56), obtenemos finalmente

$$\frac{dp}{dt} = (r_1 - r_2) \times p \times q \tag{7.57}$$

Considerando que la diferencia entre las tasas netas de crecimiento r_1 y r_2 representa el coeficiente de selección para el modelo continuo, es decir $s_c = (r_1 - r_2)$ (el equivalente de s_d en el modelo discreto), entonces la ecuación (7.57) la podemos escribir como

$$\frac{dp}{dt} = s_c \times p \times q \tag{7.58}$$

Es decir, la tasa de cambio en la frecuencia del alelo A_1 es igual al producto de la diferencia entre tasas de crecimiento entre los dos genotipos, multiplicado por pq, que como vimos en el capítulo «Variación y equilibrio de Hardy-Weinberg» es una función de la variabilidad genética existente (con máximo en $p = q = \frac{1}{2}$). Si bien las ecuaciones que describen el cambio de frecuencias en el modelo discreto ((7.46)) y en el modelo continuo ((7.58)) son diferentes, como discutimos antes, cuando $s_d \ll 1 \Rightarrow 1 + s_d \times p \approx 1$ y, por lo tanto,

$$\Delta_s p = \frac{s_d \times p \times q}{1 + s_d \times p} \approx s_d \times p \times q \sim s_c \times p \times q = \frac{dp}{dt} \tag{7.59}$$

por lo que bajo esas condiciones las soluciones para ambos modelos tienen formas funcionales similares (no son iguales, ya que $s_c \neq s_d$).

Ejemplo 7.4

Considerando que las tasas de crecimiento instantáneo que corresponden a los dos alelos, A_1 y A_2 en un locus bacteriano son $r_1 = 0{,}7$ y $r_2 = 0{,}2$ calcular la tasa de cambio en frecuencia, cuando la frecuencia del alelo A_1 es $p = 0{,}45$.

Utilizando la ecuación (7.58) y recordando que $s_c = r_1 - r_2 = 0{,}7 - 0{,}2 = 0{,}5$, entonces

$$\frac{dp}{dt} = s_c\, pq = 0{,}5 \times 0{,}45 \times (1 - 0{,}45) = 0{,}12375 \tag{7.60}$$

es decir que la tasa de cambio instantáneo de frecuencias hará que aumente la frecuencia de p.

PARA RECORDAR

- En el modelo discreto de selección haploide con dos alelos, los **fitness absolutos** están dados por

$$\begin{aligned} W_1 &= (1 - d_1)(1 + b_1) \\ W_2 &= (1 - d_2)(1 + b_2) \end{aligned} \tag{7.61}$$

con b_1, b_2 las tasas netas de nacimientos para los individuos de ambos alelos y d_1, d_2, las tasas de muertes de ambos.

- En el modelo discreto de selección, el cambio esperado en la frecuencia del alelo A_1 estará dado por

$$\Delta_s p = p(t+1) - p(t) = \frac{W_1\, p(t)}{W_1\, p(t) \;+\; W_2\, q(t)} - p(t) \tag{7.62}$$

- Si los **factores ambientales** varían en el tiempo de acuerdo con una función $\sigma(t)$, pero afectan en la misma proporción a los dos genotipos, es decir, si $W_1' = \sigma(t)\, W_1$, entonces

$$w_1' = \frac{W_1'}{W_2'} = \frac{\sigma(t)\, W_1}{\sigma(t)\, W_2} = \frac{W_1}{W_2} = w_1 \tag{7.63}$$

y, por lo tanto, el **fitness relativo** del genotipo 1 respecto al 2 es independiente de la forma de $\sigma(t)$.

- Por otro lado, si $s_d = (W_1 - W_2)/W_2$, entonces una forma alternativa de calcular el cambio de frecuencia del alelo A_1 en una generación estará dado por

$$\Delta_s p = \frac{s_d \times p \times q}{1 + s_d \times p} \tag{7.64}$$

- En el caso del modelo continuo, si $r_1 = b_1 - d_1$, $r_2 = b_2 - d_2$ son las tasas de crecimiento instantáneo para los individuos de los dos alelos y $s_c = r_1 - r_2$ es la diferencia entre ambas, entonces la tasa de cambio de frecuencia del alelo A_1 estará dada por

$$\frac{dp}{dt} = s_c \times p \times q \tag{7.65}$$

7.4 Modelos de Moran y de fisión versus Wright-Fisher

En el capítulo «Deriva genética» vimos por primera vez un modelo que daba cuenta de los procesos estocásticos que ocurren en la elección de los gametos que formarán la siguiente generación de individuos, es decir un **modelo reproductivo**. Ese modelo que vimos era el modelo Wright-Fisher, y consistía en el muestreo con reemplazo de los $2N$ gametos necesarios para formar una nueva población de N individuos diploides. Si bien se trata de un modelo sumamente útil y fácil de conceptualizar, no se trata del único modelo imaginable que represente el proceso de muestreo gamético. De hecho, existe un modelo también ampliamente utilizado que se conoce como el **modelo de Moran** en honor a Patrick Moran,[11] quien lo propuso por primera vez. Estrictamente el modelo de Moran se aplica a poblaciones de especies haploides, aunque con algunas modificaciones funciona para modelar especies diploides. En nuestro caso vamos a trabajar con una población de tamaño constante con N individuos de una especie **haploide**.

El algoritmo del modelo de Moran es muy sencillo: a cada instante en el tiempo un individuo es elegido para reproducirse y va a dejar una copia, mientras que un individuo es elegido para morir (puede ser el mismo individuo que acaba de dejar una copia, pero no la copia). Por lo tanto, la posible descendencia de un individuo (considerándose a sí mismo como descendencia) en cada instante es 0, 1 o 2 descendientes: 0 si es elegido exclusivamente para morir, 2 si es elegido exclusivamente para copiarse o 1 en el resto de los casos. Como en la población hay N individuos, la probabilidad de elegir un individuo para copiarse es $\frac{1}{N}$, mientras que la probabilidad de ser elegido para morir es también $\frac{1}{N}$. La probabilidad de dejar 0 descendientes es la que resulta de NO ser elegido para copiarse, es decir $(N-1)/N$, y luego ser elegido para morir $1/N$, por lo que multiplicando ambas tenemos la probabilidad $(\frac{(N-1)}{N}) \times (\frac{1}{N}) = \frac{(N-1)}{N^2}$. De igual manera, la probabilidad de que un individuo deje 2 descendientes es la probabilidad de ser elegido para copiarse $(1/N)$, multiplicada por la probabilidad de NO ser elegido para morir $((N-1)/N)$, por lo que también es $(\frac{1}{N}) \times (\frac{(N-1)}{N}) = \frac{(N-1)}{N^2}$. Como $p_1 = 1 - p_0 - p_2$, entonces

$$p_1 = 1 - \frac{(N-1)}{N^2} - \frac{(N-1)}{N^2} = \frac{N^2 - 2(N-1)}{N^2} = \frac{(N-1)^2 + 1}{N^2} \tag{7.66}$$

En resumen, para el modelo de Moran tenemos entonces

$$\begin{cases} p_0 = \frac{(N-1)}{N^2} \\ p_1 = \frac{(N-1)^2+1}{N^2} \\ p_2 = \frac{(N-1)}{N^2} \end{cases} \tag{7.67}$$

por lo que cuando $N \to \infty$, $p_1 \to 1$ y $p_0 = p_2 \to 0$, es decir la mayor parte de los individuos solo dejarán un descendiente en este modelo.

[11] Patrick Alfred Pierce Moran (14 de julio de 1917 - 19 de setiembre de 1988), FRS, fue un estadístico australiano que realizó importantes contribuciones a la teoría de la probabilidad y a su aplicación a la genética poblacional y evolutiva. En particular, fue el que propuso el modelo de muestreo de gametos como un proceso de nacimiento-muerte, un tipo de modelos con amplia aplicación en otras ramas de la ciencia.

Veamos que ocurre a nivel de la población. Supongamos que tenemos i individuos haploides que portan el alelo A_1, y que por lo tanto $N-i$ llevan el alelo A_2. Para pasar de pasar de i individuos en el instante de tiempo t a $i-1$ en el siguiente instante lo que debería ocurrir es que un individuo con alelo A_2 sea elegido para reproducirse y que uno con alelo A_1 sea elegido para morir. Es decir, en términos de probabilidades, tenemos

$$p_{i,i-1} = (\frac{(N-i)}{N}) \times (\frac{i}{N}) = \frac{(N-i) \times i}{N^2} \tag{7.68}$$

donde el término $\frac{(N-i)}{N}$ es la probabilidad de elegir un A_2 (para que se copie) y el término $\frac{i}{N}$ es la probabilidad de elegir un A_1 (para que se muera); como ambos son eventos independientes la probabilidad conjunta es el producto de sus probabilidades. De la misma manera, para pasar en el instante t de i copias del alelo A_1 a $i+1$ copias en el siguiente, debemos elegir un alelo A_1 para que se copie y un alelo A_2 para que muera. Esto ocurrirá con probabilidad

$$p_{i,i+1} = (\frac{i}{N}) \times (\frac{(N-i)}{N}) = \frac{i \times (N-i)}{N^2} \tag{7.69}$$

donde, nuevamente, $\frac{i}{N}$ es la probabilidad de elegir un alelo A_1 (esta vez para que se copie) y $\frac{(N-i)}{N}$ es la probabilidad de elegir un alelo A_2 (para que se muera, ahora). Claramente, $p_{i,i-1} = p_{i,i+1}$. Por otro lado, la probabilidad de mantener el mismo número de copias de A_1 (y por lo tanto de A_2, ya que la población es de tamaño constante) en el siguiente instante está dada por $1 - p_{i,i-1} - p_{i,i+1}$, o lo que es equivalente

$$p_{i,i} = \frac{i^2}{N^2} + \frac{(N-i)^2}{N^2} = \frac{i^2 + (N-i)^2}{N^2} \tag{7.70}$$

La explicación de la ecuación (7.70) es muy sencilla; para mantener el mismo número de copias hay dos posibilidades: *a)* copiar un alelo A_1 y que muera un alelo A_1, o *b)* copiar un alelo A_2 y que muera un alelo A_2. El primer escenario ocurre con probabilidad $(\frac{i}{N}) \times (\frac{i}{N}) = \frac{i^2}{N^2}$, mientras que el segundo ocurre con probabilidad $(\frac{(N-i)}{N}) \times (\frac{(N-i)}{N}) = \frac{(N-i)^2}{N^2}$. Como ambos son eventos mutuamente excluyentes debemos sumar sus probabilidades, lo que nos da el resultado de la ecuación (7.70).

En resumen, tenemos entonces que las probabilidades de transición en el modelo de Moran están dadas por

$$\begin{cases} p_{i,i-1} = \frac{(N-i) \times i}{N^2} \\ p_{i,i} = \frac{i^2 + (N-i)^2}{N^2} \\ p_{i,i+1} = \frac{i \times (N-i)}{N^2} \end{cases} \tag{7.71}$$

A diferencia del modelo de Wright-Fisher en donde excepto desde los estados absorbentes siempre existía una probabilidad de pasar a cualquiera de los estados (ver la sección «Cadenas de Markov»), en este caso nuestra **matriz de transición será *tridiagonal***, una matriz cuyos elementos son solo distintos de cero en la diagonal principal y las diagonales adyacentes por encima y por debajo de esta. Ambos modelos, en su forma básica, dependen de una

serie de supuestos que simplifican el tratamiento, esencialmente que se trata de un locus selectivamente neutro, con todos los alelos de idéntico **fitness** y que no existe estructura demográfica (es decir, el apareamiento es al azar).

Si recuerdas de la sección «El modelo coalescente», a partir de la distribución del número de alelos en una muestra podíamos deducir varias cosas. En particular, la probabilidad de que ninguno de los k alelos converja en las pasadas t generaciones es igual a

$$P(T_k^N > t) = (1 - \frac{k(k-1)}{2N})^t \tag{7.72}$$

Notemos que a diferencia de la sección «El modelo coalescente» en la que trabajamos con individuos **diploides**, ahora estamos trabajando con especies **haploides** y por eso el factor $2N$ en el denominador en lugar del $4N$ que habíamos visto antes.

De hecho, vimos que era posible escalar el modelo para que no dependiese del tamaño poblacional. Si aplicamos la transformación $T_k = T_k^N/N$, entonces la siguiente aproximación es válida

$$P(T_k > v) \approx e^{-\frac{k(k-1)}{2}v} \tag{7.73}$$

Claramente, el cambio en el número de alelos en una generación es muy diferente entre el modelo de Wright-Fisher y el de Moran, y eso afectará el número de generaciones necesarias para cada nuevo evento de coalescencia. En una aproximación muy cruda podemos ver el modelo de Wright-Fisher como requiriendo N generaciones del modelo de Moran y, por lo tanto, para poder comparar las expectativas bajos diferentes modelos reproductivos debemos escalar los procesos de manera consistente. Kingman muestra que es posible usar un escalado tal que permita que los tiempos de coalescencia sigan una distribución exponencial con tasa $k(k-1)/2$ (Kingman 1982a). Ese factor de escala es nuestro conocido **tamaño efectivo poblacional** (N_e), pero ahora partiendo de la definición que se refiere a la varianza en el número de descendientes por progenitor. Para tamaños poblacionales grandes $N_e^{*WF} = N$, mientras que para el modelo de Moran $N_e^{*M} = N^2/2$.

Un modelo alternativo es el **modelo de fisión**, en el cual, en el instante t cada individuo deja 0, 1 o 2 descendientes con probabilidades p_0, p_1 y p_2. En principio este modelo no asume tamaño constante, pero para que se cumpla la restricción de tamaño constante (y que sea comparable con lo que hemos asumido en otros modelos) es necesario que $p_0 = p_2 \leqslant \frac{1}{2}$. Claramente este modelo es intuitivo para modelar la dinámica de poblaciones bacterianas ya que representa el proceso de fisión por el cual se reproducen: a cada instante la bacteria puede morir (0 «descendientes», incluyéndose), puede no dividirse (1 «descendiente», ella misma) o puede dividirse y, por lo tanto, dejar 2 bacterias para el siguiente instante del tiempo. Para este modelo el factor de escala es $N_e^{*F} = N/(2p_2)$. En el caso de $p_2 = 1/N$, como sería en el modelo de Moran, entonces $N_e^{*F} = N/(2p_2) = N/[2(1/N)] = N^2/2 = N_e^{*M}$ que es el mismo factor de escala que habíamos visto para Moran. En el caso de que $p_2 = \frac{1}{2}$, entonces $N_e^{*F} = N/(2p_2) = N/[2(1/2)] = N = N_e^{WF}$, que es el factor de escalado para el modelo de Wright-Fisher. Un tratamiento más exhaustivo y claro de este tema, así como las implicancias para la inferencia demográfica a partir de las frecuencias alélicas puede encontrarse en Robinson, Falush y Feil (2010).

PARA RECORDAR

- El algoritmo del modelo de Moran es muy sencillo: a cada instante en el tiempo un individuo es elegido para reproducirse y va a dejar una copia, mientras que un individuo es elegido para morir (puede ser el mismo individuo que acaba de dejar una copia, pero no la copia). Por lo tanto, la posible descendencia de un individuo (considerándose a sí mismo como descendencia) en cada instante es 0, 1 o 2 descendientes.
- La probabilidad de dejar 0, 1 o 2 descendientes (p_0, p_1, p_2) bajo el modelo de Moran es

$$\begin{cases} p_0 = \frac{(N-1)}{N^2} \\ p_1 = \frac{(N-1)^2+1}{N^2} \\ p_2 = \frac{(N-1)}{N^2} \end{cases} \tag{7.74}$$

  por lo que cuando $N \to \infty$, $p_1 \to 1$ y $p_0 = p_2 \to 0$; es decir la mayor parte de los individuos solo dejarán un descendiente según este modelo.
- En general, las probabilidades de transición en el modelo de Moran están dadas por

$$\begin{cases} p_{i,i-1} = \frac{(N-i)\times i}{N^2} \\ p_{i,i} = \frac{i^2+(N-i)^2}{N^2} \\ p_{i,i+1} = \frac{i\times(N-i)}{N^2} \end{cases} \tag{7.75}$$

- Vimos que es posible escalar el modelo coalescente para que no dependa del tamaño poblacional. Si aplicamos la transformación $T_k = T_k^N/N$, entonces la siguiente aproximación es válida

$$P(T_k > v) \approx e^{-\frac{k(k-1)}{2}v} \tag{7.76}$$

- El factor de escalado que permite que los tiempos de coalescencia sigan una distribución exponencial con tasa $k(k-1)/2$ es el **tamaño efectivo poblacional** correspondiente a cada uno de los modelos. Para tamaños poblacionales grandes $N_e^{*WF} = N$, mientras que para el modelo de Moran $N_e^{*M} = N^2/2$.
- Un modelo alternativo es el **modelo de fisión**, en el cual en el instante t cada individuo deja 0, 1 o 2 descendientes con probabilidades p_0, p_1 y p_2. En principio este modelo no asume tamaño constante, pero para que se cumpla la restricción de tamaño constante (y que sea comparable con lo que hemos asumido en otros modelos) es necesario que $p_0 = p_2 \leqslant \frac{1}{2}$.
- Para el modelo de fisión el factor de escala es $N_e^{*F} = N/(2p_2)$. En el caso de $p_2 = 1/N$, como sería en el modelo de Moran, entonces $N_e^{*F} = N/(2p_2) = N/[2(1/N)] = N^2/2 = N_e^{*M}$ que es el mismo factor de escala que vimos para Moran, mientras que en el caso de que $p_2 = \frac{1}{2}$, entonces $N_e^{*F} = N/(2p_2) = N/[2(1/2)] = N = N_e^{WF}$, que es el factor de escalado para el modelo de Wright-Fisher.

7.5 El rol de la transferencia horizontal

La forma general de herencia del material genético es que los ancestros traspasen el mismo a sus descendientes. Es decir, el pasaje del material genético es de una generación a la siguiente (por lo tanto, entre individuos de diferentes generaciones) y por eso se conoce a esta forma de herencia como **transferencia vertical** de genes. Este es el mecanismo clásico y el que asegura que los descendientes sean, ignorando la mutación, una mezcla del material genético de sus progenitores. Como vimos anteriormente, pese a que en los procariotas no existen los mecanismos de recombinación clásicos de los eucariotas, existen diferentes formas alternativas que permiten el intercambio de material genético entre los mismos. Vimos así que es usual que bacterias y arqueas intercambien material genético con otros individuos de su especie y aún con individuos de otras especies. Cuando la recombinación es homóloga, en particular con otros individuos de la especie, podemos (abusando de la terminología) considerar a la misma como una forma de recombinación, ya que básicamente existe una conservación del orden y de las secuencias (más allá de las pequeñas diferencias existentes). Sin embargo, cuando se trata de un evento de intercambio no homólogo, o aún con individuos de otras especies, entonces ya no podemos seguir considerando esto como un evento de recombinación y debemos buscar una nueva forma de denominar al fenómeno y de comprender sus consecuencias.

Dado que la transferencia de material genético entre especies no es la forma **vertical** de transmitirlo, podemos llamar a esta forma **horizontal** o **lateral** y, por lo tanto, al proceso **Transferencia Horizontal de Genes** (HGT, por sus siglas en inglés) o **Transferencia Lateral de Genes** (LGT, ídem). Para comprender mejor alguna de sus consecuencias definamos antes algunos términos evolutivos. En la figura 7.7 podemos observar la evolución de un gen, la histona H1, en el ser humano y en el chimpancé.

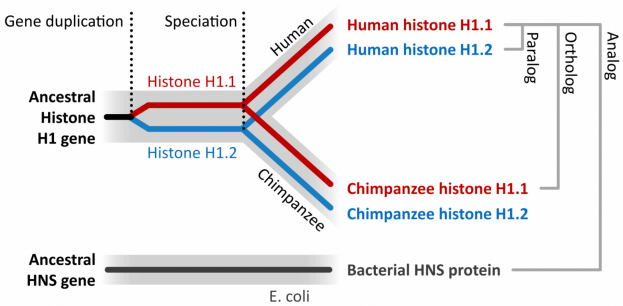


Figura 7.7. Distinción entre diferentes formas de homología entre secuencias. Un evento de duplicación de una histona en el genoma de una especie ancestral al humano y el chimpancé provoca que en la actualidad cada una de las especies también posean un par copias génicas codificantes para histonas. Desde su duplicación inicial las dos copias han evolucionado de forma diferente y, por lo tanto, la histona H1.1 es diferente de la histona H1.2. Más aún, desde el evento de especiación que llevó a que seamos una especie diferente al chimpancé cada una de las histonas ha continuado evolucionando y, por lo tanto, ya no serán idénticas las histonas H1.1 de estas especies, y tampoco las copias codificantes para histona H1.2. Se denomina *ortólogos* a los genes que se relacionan entre sí por un evento de especiación, mientras que se define *parálogos* a los que se relacionan a través de un evento de duplicación ancestral. Figura de Wikipedia, autor: Thomas Shafee (CC BY 4.0).

En algún ancestro de ambas especies, mucho tiempo atrás, se produjo un evento que duplicó el gen de la histona H1 produciendo dos copias del mismo, las cuales llamaremos H1.1 y H1.2. Es fundamental entender que a partir de dicho punto existirán dos copias del mismo gen en el genoma haploide de la especie, ubicadas en diferentes lugares del genoma y no necesariamente en el mismo cromosoma (aunque suele ocurrir que sí).[12] Cada una de estas copias del gen ancestral de la histona H1 seguirá un camino evolutivo relativamente independiente de la otra, y a lo largo del tiempo irá acumulando diferentes mutaciones y cambios. Por lo tanto, a medida que el tiempo transcurre las secuencias de las copias génicas H1.1 y H1.2 ya no serán idénticas. Lo que es más, cuando se produce un evento de especiación los caminos evolutivos pueden divergir también en este nivel. Este concepto se ve también ilustrado en el ejemplo de la figura 7.7. En los primates un evento de especiación dio lugar a dos ramas que divergieron evolutivamente: una de estas ramas lleva a los humanos, mientras que la otra lleva al chimpancé (*Pan troglodytes*, la especie viva más cercana al humano). ¿Cómo afecta esto a la divergencia evolutiva entre copias génicas de histona? Una vez se establece el proceso de especiación, las copias de H1.1 en cada una de las ramas seguirá un camino evolutivo independiente; lo mismo ocurrirá para las copias de H1.2. Esto establece diferentes relaciones evolutivas entre las 4 copias (en los genomas haploides) que encontramos en chimpancé y humano. Por ejemplo, en el caso de H1.1 de chimpancé y H1.1 de humanos las copias eran idénticas en el último ancestro de ambas especies, una relación que se define como ***ortología***. Decimos por lo tanto que los genes H1.1 de chimpancé y humano son genes ortólogos. Si en cambio analizamos la situación existente entre copias génicas codificadas en humano, notaremos que las dos copias se relacionan entre sí por el evento de duplicación ancestral mencionado anteriormente. Este tipo de relación se define como ***paralogía***. Decimos por lo tanto que los genes codificantes para histona H1.1 y H1.2 de humano son **parálogos**.

Hasta este momento se analizaron las relaciones existentes entre genes que son el producto de eventos de duplicación intragenómica y de especiación. Sin embargo, cuando ocurren eventos de **transferencia horizontal** de material entre dos especies diferentes puede ocurrir que en el genoma receptor ya estén codificados genes homólogos a parte (o todo) el material transferido. Esto lleva a la duplicación del número de «copias» del material genético en el genoma receptor. Pero el material duplicado es homólogo, y si bien puede ser de alguna manera similar, no es para nada necesario que se trate de copias idénticas. Esto lleva a definir una nueva forma de relación evolutiva entre genes, a los cuales llamaremos genes ***xenólogos***. En la figura 7.8 podemos apreciar, de forma esquemática, el surgimiento de un par de genes xenólogos producto de un evento de transferencia horizontal.

Los genes A_1 y A_2 son genes homólogos (es decir, comparten alguna relación de similaridad evolutiva), el primero en la especie 1 y el segundo en la especie 2. Cuando ocurre un evento de transferencia horizontal de la especie 2 a la especie 1, llevando una copia de A_2 hacia el genoma de la especie 1, entonces todos los descendientes de ese individuo de la especie 1 pasarán a tener dos «copias» (diferentes entre sí) del gen A. Estas dos copias serán xenólogas, ya que la relación de homología entre ellas proviene de un evento de transferencia horizontal entre especies diferentes.

Una de las consecuencias de los eventos de transferencia horizontal, posiblemente la menos relevante, es la dificultad para entender las relaciones evolutivas entre especies a partir de unas pocas secuencias nucleotídicas, ya que los árboles filogenéticos (una de las vías usuales

[12] El mecanismo específico mediante el cual se produce la duplicación génica será el que determinará si las copias génicas se encuentran o no ubicadas en un mismo cromosoma.

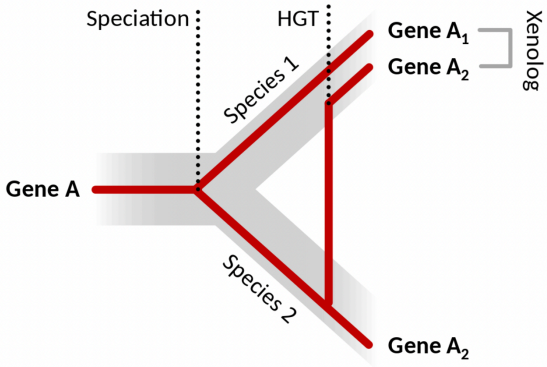


Figura 7.8. Aparición de una nueva copia de una secuencia proveniente de un homólogo en otra especie a partir de un evento de **transferencia horizontal de genes**. Los genes A_1 y A_2 son homólogos entre sí en diferentes especies, pero cuando el A_2 se transfiere de la especie 2 a la especie 1, entonces en la especie 1 hay ahora dos copias, que son ***xenólogas*** entre sí. Figura de Wikipedia, autor: Thomas Shafee (CC BY 4.0).

de reconstrucción de las relaciones) asumen que la homología y la proporción de similitud entre secuencias es un fiel reflejo de las relaciones evolutivas existentes entre especies, una asunción que es razonable, hasta cierto punto, si se asume la transferencia vertical de información genética durante el proceso evolutivo. A partir de secuencias que han experimentado eventos de transferencia horizontal esto último deja de ser cierto, ya que en dicho caso las trayectorias evolutivas de secuencias y especies serán diferentes. Alcanza con ver nuevamente la figura 7.8 para comprender este fenómeno. Si analizamos las secuencias codificantes para A_2 en la especie 1 y en la especie 2 inmediatamente después del evento de HGT, las mismas serán idénticas, o lo que es lo mismo coincidirán 100 % en sus secuencias nucleotídicas. Supongamos ahora que en la rama que lleva desde el ancestro común de la especie 1 y 2 a la especie 2 hubiese aparecido una tercera especie, por ejemplo, a nivel relativamente basal, la cual llamaremos especie 3. Las copias de A_2 en la especie 3 acumularían diferencias con la copia de A_2 codificada en la especie 2. Al reconstruir un árbol filogenético basado en las secuencias de A_2, al ser idénticas las secuencias de A_2 en la especie 1 y la especie 2 (dada la existencia de un evento de HGT), las mismas quedarían agrupadas muy juntas, mientras que la especie 3 quedaría por fuera de estas dos, más lejana, pese a que en realidad la especie 2 y la especie 3 deberían estar más estrechamente emparentadas respecto a la especie 1 (que se separó previamente de ellas) en un árbol filogenético que refleje el proceso evolutivo que dio lugar a estas especies. Si bien esto es una dificultad que ocurre a menudo, existen diferentes alternativas para modelar este fenómeno e incorporarlo a la reconstrucción filogenética, o al menos para identificar que las secuencias analizadas son xenólogas y no basar en ellas nuestras conclusiones evolutivas respecto al proceso evolutivo entre especies.

Una forma usual de identificar **islas de transferencia horizontal** (regiones relativamente largas del genoma que provienen de otra especie) es basarse en la composición de las mismas. Como hemos visto antes (capítulo «Introducción a la genómica»), las diferentes especies poseen diferentes estilos de vida y de funcionamiento de la maquinaria celular, con diferentes sesgos mutacionales, tasas de mutación, de aceptación de material externo, así como de requerimientos. Todos estos factores usualmente afectan la composición de los genomas,

tanto a nivel de los aminoácidos que constituyen las proteínas, como de los nucleótidos, ya sea en el caso que estos correspondan a secuencias codificantes como no-codificantes para proteínas u ARN funcionales. En procariotas, donde la uniformidad composicional a nivel nucleotídico es importante, un primer indicador de material procedente del exterior de la especie son las regiones de un genoma que poseen un contenido **GC** marcadamente diferente al del resto del genoma. Un segundo nivel, algo más complejo pero que puede llegar a tener una mayor capacidad de discriminación respecto del origen de material genético, es el uso de codones de estas regiones (cuando se consideran secuencias codificantes para proteínas, obviamente). En el caso de procariotas, muchas especies suelen tener una preferencia marcada por el uso de algunos codones, en particular dentro de cada aminoácido (para aquellos aminoácidos que tienen codones sinónimos), por lo que aquellas regiones con una composición inusual de codones para la especie pueden ser candidatas interesantes a provenir de eventos de HGT relativamente recientes. Es importante remarcar que la evolución continúa luego de los eventos de HGT, por lo que aún las secuencias que provienen de ellos perderán sus características propias con el paso del tiempo y se irán pareciendo en los aspectos composicionales a las secuencias del genoma receptor donde se instalaron.

Los eventos de transferencia horizontal tienen y han tenido una relevancia extrema en la evolución de la vida en nuestro planeta. Por un lado, tanto las mitocondrias como los plástidos (los cloroplastos, por ejemplo) provienen de eventos de transferencia horizontal desde procariotas a eucariotas. En ambos casos se trataría no solo del material genético sino de la célula en su conjunto, pero desde el punto de vista genético se trata de eventos de HGT (que afortunadamente podemos identificar fácilmente porque no se han integrado totalmente al genoma nuclear, además). Por otra parte, la transferencia de plásmidos entre bacterias constituye también un tipo particular de eventos de transferencia horizontal, eventos mucho más plásticos en términos evolutivos ya que pueden ser fácilmente revertibles: tanto como una «especie» puede ganar un plásmido en un contexto determinado, lo puede perder cuando este no representa una ventaja evolutiva si las presiones selectivas cambian. Por ejemplo, este mecanismo de HGT es usual para adquirir genes de resistencia a compuestos antimicrobianos, que muchas veces se encuentran en los plásmidos por esta razón. Otros eventos de transferencia horizontal en procariotas pueden conferir resistencia a condiciones ambientales desafiantes como la existencia de metales tóxicos, la hipersalinidad o temperaturas extremas, entre otros. Por lo tanto, su importancia va más allá de los fenómenos evolutivos, ya que suelen afectar las dinámicas de las diferentes poblaciones microbianas. Este es sin duda un mecanismo importante de compartir novedades evolutivas dentro y entre especies.

PARA RECORDAR

- Los eventos de **transferencia horizontal de genes** involucran el pasaje de material genético entre individuos (generalmente) de diferentes especies.
- En el caso de que en el genoma receptor existan secuencias homólogas a las transferidas, se genera una nueva relación evolutiva entre las mismas que se llama **xenología** (porque su origen es de diferentes especies).
- Los eventos de **transferencia horizontal de genes** involucran el pasaje de material genético entre individuos (generalmente) de diferentes especies.
- En el caso de que en el genoma receptor existan secuencias homólogas a las transferidas, se genera una nueva relación evolutiva entre las mismas que se llama **xenología** (porque su origen es de diferentes especies).

- En procariotas es relativamente sencillo identificar eventos **recientes** de HGT en los casos en que las diferencias composicionales (nucleotídicas, uso de codones, etcétera) entre las dos especies sean relativamente importantes.
- Los eventos de HGT son de gran importancia en la evolución, pero también en la genética de poblaciones, por ejemplo, a través de la transferencia de genes de resistencia a antimicrobianos, o de tolerancia a concentraciones elevadas de citotóxicos.

7.6 Seleccionismo versus neutralismo: los procariotas en el debate

Uno de los puntos de fuerte controversia entre las teorías neutralista y seleccionista de la evolución molecular se encuentra en la explicación de las importantes diferencias observadas en el contenido GC de los genomas, tanto de eucariotas como de procariotas, ya sea entre genomas o dentro de ellos. En vertebrados homeotermos, Bernardi y colaboradores descubrieron hace mucho tiempo que el genoma se encuentra estructurado en largas regiones ($>$ 200 kb) de contenido GC relativamente homogéneo en comparación a las diferencias observadas entre regiones (Bernardi et al. 1985). A lo largo de los años se han encontrado diversas asociaciones entre estas estructuras y propiedades genómicas de distinta importancia: diferencias en el tiempo de replicación (Eyre-Walker 1992), longitud de los genes (Duret, Mouchiroud, and Gautier 1995), bandeo cromosómico (Saccone et al. 1993), islas CpG[13] (Aïssani and Bernardi 1991), patrones de metilación (Cacciò et al. 1997), así como importantes diferencias en la estructura entre vertebrados homeotermos y poiquilotermos (Belle, Smith, and Eyre-Walker 2002). Existe aún un fuerte debate respecto al origen y evolución de esta estructura en mosaico (Eyre-Walker and Hurst 2001), aunque la disponibilidad de gran cantidad de nuevos genomas ha disminuido el margen de controversia considerablemente (Costantini, Cammarano, and Bernardi 2009).

En procariotas el debate acerca del papel de la selección natural en la evolución del contenido GC ha sido considerablemente más fuerte, en parte debido a la estructura mucho más simple de sus genomas, a los importantes tamaños poblacionales y a las enormes diferencias de GC entre organismos (ya en la década de los sesenta se observaron valores en el rango del 25 % al 75 %) (Sueoka 1962). Esto fue tomado en un principio como soporte a la teoría neutral de la evolución, propuesta por Motō Kimura.[14] En razón de que a esta altura resulta muy claro que el sesgo mutacional en procariotas es hacia los pares de base $A = T$, por lo que la existencia de organismos con un contenido **GC** muy alto (o lo que es igual, bajo contenido AT) no son esperables, es necesario recurrir a otras explicaciones.

[13] Las «islas CpG» son regiones en el genoma que se caracterizan por la repetición de dinucleótidos de citosina (C) seguidos de guanina (G). La *p* en este término hace referencia al enlace fosfodiéster entre los nucleótidos. Es relevante destacar que en la hebra complementaria también se encuentra el mismo dinucleótido, lo que lleva a la analogía con una 'isla'.

[14] Motō Kimura (13 de noviembre de 1924 – 13 de noviembre de 1994) fue un biólogo evolutivo japonés y uno de los genetistas de poblaciones más importantes del siglo XX. Originalmente formado como botánico, pero con gran interés en las matemáticas, después de la segunda guerra mundial comenzó a publicar en genética teórica. Propuso la teoría neutral de la evolución que pese a no ajustarse al conocimiento actual y a los datos moleculares disponibles, aún hoy en día sigue constituyendo el patrón de comparación para otras teorías alternativas.

Una opción natural en esta búsqueda de explicaciones alternativas a la teoría neutral es la teoría de la selección natural, que ya hemos visto, operando a nivel molecular. Para que la selección sea operativa es necesario encontrar no solo la característica seleccionada, sino intentar determinar cuál es la ventaja selectiva (evolutiva) que confiere. En nuestro caso se trata del incremento en el contenido **GC** la característica que estaría siendo seleccionada de forma favorable, aunque nos faltaría encontrar una explicación consistente de la razón para esta ventaja. Entre los diferentes factores ecofisiológicos propuestos para la ventaja de un contenido **GC** más elevado se encuentran la temperatura óptima de crecimiento (Kagawa et al. 1984; Musto et al. 2004), la fijación de nitrógeno (McEwan, Gatherer, and McEwan 1998), oxígeno (Naya et al. 2002; Romero et al. 2009) y parasitismo (Rocha and Danchin 2002), entre otras tantas. Mientras que algunas de estas explicaciones no han soportado demasiado bien el paso del tiempo, viéndose restringidas a casos particulares, otras aún parecen plausibles a la luz de nueva evidencia y explican mejor los datos que las hipótesis neutralistas (excelentes revisiones recientes se encuentran en Rocha and Feil 2010; Rocha 2018). Alternativamente, el mecanismo de ***biased gene conversion***, un mecanismo de recombinación homóloga, ha sido propuesto como otra posible explicación del incremento del contenido **GC**. Como el mismo es dependiente de la recombinación, mientras que la selección se ve favorecida por la recombinación (como vimos en el capítulo «Dinámica de dos loci»), es bastante difícil conseguir resolver cuál de estas dos explicaciones alternativas es más verosímil (Rocha 2018).

En resumen, mientras que el debate acerca de las causas subyacentes continúa, resulta claro que existe un papel fundamental del par GC en la evolución a nivel molecular.

PARA RECORDAR

- El debate entre las teorías neutralista y seleccionista de la evolución molecular persiste aún hoy en día, aunque la teoría neutral en su forma original no es capaz de explicar la distribución de la variabilidad en la mayor parte de los casos conocidos.
- Una arena interesante donde batirse ha resultado ser el contenido GC de los procariotas ya que el mismo muestra una enorme variabilidad entre especies y porque en procariotas la mayor parte de los sitios está posiblemente sujeta a selección, directamente o por ligamiento con regiones bajo selección.
- En procariotas parece bastante claro que el sesgo mutacional-sustitucional es hacia los nucleótidos AT, por lo que, para explicar la distribución, con muchos genomas hacia GC alto, requiere de otras explicaciones que el sesgo mutacional.
- La teoría neutral de la evolución no alcanza a explicar esta enorme variabilidad en contenido GC pero el ***biased gene conversion***, un mecanismo de recombinación homóloga independiente de la selección, ha sido propuesto como otra posible explicación del incremento del contenido **GC**.
- Diferentes factores ambientales y evolutivos han sido propuestos para explicar la distribución de contenido GC en procariotas, entre ellos la temperatura óptima de crecimiento, la fijación de nitrógeno, el parasitismo y la aerobiosis. Algunos han soportado bien el paso del tiempo y son bastante generales, mientras que otros apenas pueden explicar algo en un entorno filogenético muy restringido o han sido descartados por evidencia más reciente.

- Como el mecanismo de *biased gene conversion* es dependiente de la recombinación, mientras que la selección se ve favorecida por la recombinación, es bastante difícil conseguir resolver cuál de estas dos explicaciones alternativas funciona mejor.

7.7 Genómica poblacional

El primer genoma procariota completamente secuenciado fue el de la bacteria *Haemophilus influenzae Rd* (Fleischmann et al. 1995) y fue el debut en organismos vivientes del método de secuenciación conocido como ***shotgun sequencing***, en el que la secuenciación de fragmentos es aleatoria y, por lo tanto, el ensamblado (reconstrucción) del genoma depende de secuenciar muchas veces cada base, en general en fragmentos solapantes. El genoma de *H. influenzae Rd* en el ensamblado original fue de 1.830.137 pares de bases (un genoma relativamente pequeño para una bacteria), pero significó un salto enorme en comparación con los logros previos como la secuenciación del bacteriófago ΦX174 (5.386 pb), el bacteriófago λ (48.502 pb), o aún el genoma mitocondrial (187 kb) y del cloroplasto (121 kb) de la planta *Marchantia polymorpha*. Todas estas secuencias completas fueron obtenidas mediante el método de Sanger,[15] que originalmente implicaba un enorme trabajo manual para cada secuencia. Con el correr de los años el método se fue automatizando, permitiendo su uso masivo e inclusive llegando a ser utilizado en los proyectos originales para secuenciar el genoma humano.

En los años que siguieron a la secuenciación completa de *H. influenzae Rd* se resolvieron las secuencias de decenas de otros procariotas, y en el año 2001 se finalizó el primer borrador del genoma humano, un evento histórico si se tiene en cuenta que su genoma haploide es de $\approx$ 3,3 Gb (es decir, el equivalente a 711 genomas de la bacteria *Escherichia coli K-12* o 1803 genomas de la bacteria *H. influenzae Rd*). La forma de resolver el genoma humano fue una combinación de dos grandes proyectos, uno público y otro privado, dos estrategias de secuenciación (una de ellas ***shotgun***), cientos de centros que colaboraron en la generación de secuencias y un esfuerzo computacional masivo para la época. Una de las consecuencias de la «finalización» de este proyecto fue la disponibilidad de un gran número de secuenciadores, en algunos casos dispuestos en ***factories*** (fábricas) a efectos de paralelizar al máximo los procesos de secuenciado. Esta disponibilidad de capacidad **ociosa** permitió que ideas relativamente nuevas tuvieran lugar para desarrollarse y así surgió la posibilidad de pasar de una visión de la genómica basada en individuos a una centrada en poblaciones. Un artículo histórico, en este sentido, es el de Stein y colaboradores que plantea la secuenciación de un fragmento de genoma de 40 kb de una arquea planctónica marina (Stein et al. 1996), es decir, la secuencia de un organismo **no-cultivable** con los estándares de aquel momento. La relevancia de esto es extrema, aunque pueda no parecernos, ya que nos permitió pasar del estudio únicamente de organismos cultivables a todos los organismos que podamos de alguna manera muestrear. Si bien en dicho artículo no se menciona el término ***metagenómica***,

[15] Frederick Sanger (13 de agosto de 1918 - 19 de noviembre de 2013) fue un bioquímico británico que obtuvo dos veces el Premio Nobel de Química, una de las dos únicas personas en haberlo obtenido dos veces en la misma categoría (John Bardeen es el otro, en Física) y la cuarta persona del mundo en recibir dos premios Nobel. En 1958 obtuvo su primer Premio Nobel por los trabajos que llevaron a la comprensión de la estructura de las proteínas, de la insulina en particular, mientras que el segundo lo obtuvo de forma compartida con Walter Gilbert en el año 1980 por sus contribuciones a la determinación de la secuencia de bases del ADN.

que recién fue acuñado en el artículo de Handelsman y colaboradores (Handelsman et al. 1998), se trata de alguna manera del inicio de una era en la forma de comprender la diversidad microbiana, así como las dinámicas a las que se encuentran sujetas las poblaciones bajo diferentes condiciones y presiones selectivas.

En la **metagenómica** el objetivo suele ser comprender la diversidad de organismos presentes, así como sus proporciones en diferentes muestras, a partir de la identificación y cuantificación de ácidos nucleicos, ADN o ARN. A partir de los trabajos de Woese,[16] es conocido que el ARN ribosomal 16S es un excelente marcador filogenético cuando consideramos la profundidad de los dos dominios procariotas (es decir, Archaea y Bacteria), en gran parte debido al alto nivel de conservación de esta molécula. Tener una molécula o, mejor aún, una región de la misma altamente conservada, abre las puertas no solo al estudio de las relaciones evolutivas entre organismos, sino también al estudio de la distribución y variabilidad de organismos en diferentes muestras y ambientes (ver el trabajo de Lane y colaboradores para usos filogenéticos (Lane et al. 1985)). En una de las formas clásicas, los estudios metagenómicos se realizan secuenciando en cada muestra todo el ARNr 16S detectable, usualmente a través de protocolos que permiten amplificar el material original y luego realizar una transcripción reversa para convertirlo en ADN, que es mucho más estable. Posteriormente, se identifica cada fragmento secuenciado para finalmente cuantificar el número de fragmentos secuenciados (*reads*) como un ***proxy*** de la abundancia de los distintos organismos.

Si bien la **metagenómica** nació dentro de la era de los secuenciadores Sanger (por el método de secuenciación usado), su alcance fue bastante reducido debido a la poca capacidad de generar secuencias de este método. Un cambio fundamental se produjo a comienzos de la primera década del dos mil, cuando rápidamente aparecieron varios métodos nuevos de secuenciación que fueron conocidos como ***Next Generation Sequencing*** o secuenciadores de segunda generación. De todas las ideas originales, tres métodos en particular terminaron en secuenciadores de gran capacidad que influyeron enormemente en los años siguientes: el **454** de Roche (pirosecuenciado), el secuenciador de **Illumina** (secuenciación por síntesis) y el **SOLiD** (secuenciación por ligación). Si bien los tres tenían en común una enorme capacidad de generar secuencias a bajo costo, cada uno tenía una *performance* distinta, así como diferentes debilidades. El **454** producía secuencias relativamente largas (del orden de los 600 pb, cercano al largo obtenible por **Sanger**), pero el número de fragmentos producidos era mucho menor al de las otras tecnologías y era muy sensible (poseía una mayor tasa de error) a las secuencias de homopolímeros (es decir, cuando se repetía una base). El secuenciador **SOLiD**, por otra parte, producía una enorme cantidad de fragmentos (órdenes de magnitud superior al **454**), pero los mismos eran extremadamente cortos (del orden de las 50 pb), y el resultado primario que se obtenía no era la secuencia de bases sino una secuencia en **espacio color** que luego debía ser traducida al **espacio de bases**. Finalmente, el secuenciador de **Illumina**, basado en secuenciación por síntesis, es el único de los tres que sobrevivió al paso del tiempo y hoy en día convive con los **secuenciadores de tercera generación**. El mismo produce fragmentos de largo fijo, y si bien en los comienzos estos eran cortos en la actualidad son de 150 o de 300 pb, pudiéndose complementar con otras técnicas (por ejemplo, secuenciación ***paired-end***) que permite cubrir grandes regiones con precisión. Más aún, es de las pocas tecnologías que ha sido escalada en los equipos, con

[16] Carl Richard Woese (15 de julio de 1928 – 30 de diciembre de 2012) fue un microbiólogo y biofísico conocido por sus trabajos en la filogenética de la vida y por haber definido en dominio Archaea a partir del análisis del ARNr 16S. Originalmente formado en matemáticas y física, su interés en descifrar el código genético lo fue llevando a trabajar en diferentes preguntas centrales de la biología, algunas relacionadas con el origen de la vida en nuestro planeta.

algunos como el **MiniSeq** con una producción de 1,65 – 7,5 Gb por corrida (es decir, entre 7 – 25 millones de fragmentos cada vez), hasta equipos como el **HiSeq X** con la capacidad de secuenciar hasta 18 mil genomas humanos por año, siendo el primer secuenciador que logró quebrar la barrera de obtener un genoma humano a la profundidad de 30x (cada base cubierta en promedio por 30 fragmentos) por menos de U$S 1000. En la figura 7.9 se observa el resultado de la secuenciación de una región del exoma humano usando el equipo **HiSeq X.** En la misma se aprecia claramente que el inicio y fin de los fragmentos no coincide (cada fragmento es una barra gris horizontal), algo esperable ya que se trata de una fragmentación al azar y posterior secuenciación (técnica de ***shotgun***). Además, en la región que se muestra en la figura no se observan posiciones heterocigotas, ya que los pocos puntos de no coincidencia con el genoma de referencia (segmentos de color en las barras grises) no se acercan a la frecuencia esperada para un organismo diploide (50 % o 100 %) y se trata por lo tanto de errores de secuenciación.

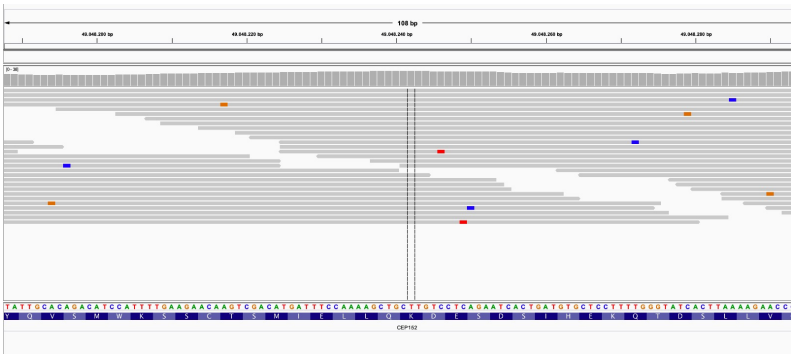


Figura 7.9. Solapamiento de fragmentos para reconstrucción de una secuencia en el método de ***shotgun***. En este caso los fragmentos, luego de secuenciados se alinean contra un genoma de referencia (el del organismo de donde provienen) que fue resuelto previamente (*Homo sapiens*, en este caso). Se aprecia claramente que al ser fragmentos producido de forma aleatoria (cada barra horizontal de color gris) su inicio y fin difieren, aunque pueden coincidir en caso de que la profundidad de secuenciación sea muy alta. En la parte inferior de la figura se observa la secuencia de nucleótidos y de aminoácidos en el genoma de referencia. Los segmentos cortos de color sobre las barras grises representan nucleótidos diferentes del correspondiente en el genoma de referencia, debido a variantes o a errores de secuenciación. Gentileza de Lucía Spangenberg, Unidad de Bioinformática, Institut Pasteur de Montevideo.

Estos secuenciadores permitieron la explosión de estudios metagenómicos al reducir los costos de manera sustantiva, así como los tiempos requeridos para procesar las muestras. El resultado típico de estos estudios es una tabla con varias muestras por condición en las filas, las diferentes **OTU** (*Operational Taxonomic Unit*, en general un equivalente a **especie**) identificadas en las muestras y el número de **OTU** observado en cada celda. Es decir, se trata de una tabla de abundancias en la escala de lo observado. Como en general no es posible estandarizar las condiciones para secuenciar exactamente el mismo número de fragmentos en cada muestra, usualmente se realiza algún tipo de transformación de estos datos que posibilite comparar de forma «justa» las diferentes muestras y condiciones. Más aún, como el número de **OTU** detectadas en cada muestra es una función del número de fragmentos secuenciados, cuando la cantidad de fragmentos obtenidos por muestra no es enorme, es usual realizar algún tipo de procedimiento para comprobar que el esfuerzo de muestreo es suficiente para no perder una parte importante de la diversidad a medida que el número de

fragmentos en las muestras disminuye (***rarefactions analysis***, en inglés); una alternativa para equilibrar las muestras es realizar un submuestreo aleatorio de todas las muestras al tamaño de la más pequeña. Luego de obtenida una tabla de abundancias estandarizada de alguna manera, el procedimiento suele involucrar algún tipo de análisis exploratorio multidimensional, como el análisis de componentes principales (PCA) o escalado multidimensional (MDS) que nos permita reducir la dimensionalidad del espacio original a espacios donde la variabilidad se concentra en pocas dimensiones y que por lo tanto pueden ser graficadas para su interpretación. Como se vio en el capítulo «Introducción a la genómica», en estas gráficas suele ser deseable observar un patrón de agrupamiento espacial por condición de la muestra (por ejemplo, por ambientes de donde fueron extraídas), lo que indicaría que la variabilidad entre muestras se corresponde mayormente a las condiciones biológicas consideradas en el experimento. Finalmente, se suele realizar algún tipo de comparación estadística entre la abundancia de los OTU más discriminantes, a fin de ayudar a identificar las fuentes más importantes de variación y poder generar hipótesis razonables que expliquen esta variabilidad.

En la actualidad los **secuenciadores de tercera generación** vienen reemplazando a los de segunda (excepto **Illumina**) ya que poseen algunas ventajas importantes. El secuenciador de **Ion Torrent** se basa un **chip** que posee microscópicos medidores de pH en unos pocillos creados en la superficie; los fragmentos de ADN se amplifican en microesferas que se instalarán en esos pocillos. A intervalos regulares se hace pasar una solución de cada uno de los nucleótidos (A, C, G, T) que se incorporarán solamente en aquellos en que la siguiente base del fragmento sea la complementaria. Cuando se incorpora un nucleótido se libera el ion H^+, lo que indicará el correspondiente pHímetro como la secuencia de esa base. La simplicidad de concepción hace de este instrumento una alternativa económica a los secuenciadores de segunda generación, pese a que comparte problemas con el **454** (problema con los homopolímeros y largo de fragmentos algo reducidos para los estándares actuales). Sin embargo, tanto en la tecnología de **Oxford Nanopore**, como en la de **Pacific Biosciences**, ambos instrumentos de tercera generación, el largo de los fragmentos no está limitado *a priori* y, por lo tanto, se obtienen fragmentos de miles de pares de bases de longitud. Esto permite otros abordajes al problema de la diversidad microbiana, además de ayudar a la comprensión de la base molecular de la ecología de los distintos ambientes. La tecnología de **Oxford Nanopore** (ONT) se basa en la creación de unos poros biológicos en una superficie a cuyos lados se aplica un campo eléctrico y que poseen sensores de corriente (cada poro). Cada molécula que atraviesa un poro, por ejemplo, una molécula de ADN, al hacerlo irá variando la corriente a través del poro ya que variará la superficie disponible para que los iones lo atraviesen. En principio, cada nucleótido tiene su «firma» al pasar, por lo que leyendo la secuencia de cambios en la corriente es posible reconstruir la secuencia de bases correspondiente (en realidad cada secuencia corta de nucleótidos tiene una firma ya que los nucleótidos que acaban de pasar y los que van a pasar también obstruyen parte del poro). Lo fantástico de esta tecnología es la posibilidad de escalar los equipamientos, desde grandes equipos que permiten la secuenciación de centenares de muestras al mismo tiempo hasta equipos del tamaño de un estuche de lentes, como el **MinION**, que conecta en un puerto USB de la computadora, figura 7.10. En estos equipos las secuencias comienzan a verse en tiempo real, a medida que son generadas y permiten una enorme flexibilidad, ya que puede interrumpirse la corrida, lavar la celda y cambiar de muestras a procesar cuando se alcance algún objetivo con las secuencias (en teoría permiten un par de lavados y un tiempo máximo de corrida de 48 horas).

Figura 7.10. Secuenciador **MinION** de Oxford Nanopore Technologies, a la izquierda y secuenciador **MinION Mkc1**, a la derecha, ambos con sus tapas abiertas, al lado de un lápiz mecánico para referencia de su tamaño. El equipo de la izquierda no es autónomo y debe conectarse al puerto USB de una computadora para funcionar, mientras que el equipo de la derecha posee capacidad propia de almacenamiento y procesamiento de los datos (foto Hugo Naya).

El extremo opuesto de la tecnología de ONT lo representa la tecnología de Pacific Biosciences, con su equipo **PacBio Sequel IIe**, con hasta 8 millones de pocillos (≈ 5 millones de fragmentos) capaz de secuenciar 150 Gb por celda. Se trata de una tecnología basada en unos pocillos trabajados en una superficie y donde se encuentra una polimerasa que irá procesando el ADN. Los 4 nucleótidos son marcados con colores diferentes y están mezclados en una solución que contiene a los 4, los que por difusión se incorporan de forma aleatoria. Cada pocillo es iluminado desde abajo con un láser y cuando un nucleótido es incorporado se mantiene por más tiempo en la región en que puede generar una señal luminosa, lo que será registrado por una cámara que detecta esa señal y la traduce a la base correspondiente al color retenido. Con esta tecnología es posible obtener fragmentos de decenas de miles de pares de bases de largo y con gran fidelidad, lo que permite además de la identificación de la secuencia de nucleótidos entender las variaciones estructurales en las secuencias, cosa que usualmente escapa a las tecnologías de fragmentos cortos.

A partir de las tecnologías de segunda y tercera generación los estudios posibles de genética y dinámica de poblaciones microbianas se han ampliado notablemente. Por un lado, es posible realizar estudios poblacionales no basados en el estudio de **amplicones** (como era el de 16S), lo que elimina el sesgo de amplificación asociado a esas estrategias (las distintas secuencias pueden amplificar diferente por PCR, distorsionando los datos originales). En estos estudios se secuencia todo el ADN o el ARN (en el caso de **metatranscriptómica**) que existe en una muestra, para luego ensamblar los fragmentos y recién después identificar los organismos que produjeron esas secuencias. Entre otras cosas, este enfoque permite reconstruir vías metabólicas más comunes en unas condiciones que en otras, una identificación más precisa de organismos y en algunos casos determinar la variabilidad entre individuos de la misma especie. En el área agraria son abundantes los estudios de suelos y aguas a través de la metagenómica, estudios en la rizósfera, así como los estudios de funcionamiento y composición del rumen, la metagenómica de glándula mamaria y en cualquier situación

donde la composición y dinámica de las poblaciones microbianas pueden jugar un papel relevante.

PARA RECORDAR

- En la metagenómica el objetivo suele ser comprender la diversidad de organismos presentes, así como sus proporciones en diferentes muestras, a partir de la identificación y cuantificación de ácidos nucleicos, ADN o ARN. En los estudios clásicos de filogenética se emplea el ARNr 16S ya que se trata de una molécula altamente conservada durante la evolución, lo que nos permite apreciar diferencias en grandes tiempos evolutivos.
- Las tecnologías de secuenciación masiva (o *Next Generation Sequencing*, NGS) permitieron extender los estudios metagenómicos a un amplio rango de problemas, incluyendo aquellos de composición y dinámica poblacional microbiana.
- Las tecnologías NGS de segunda generación aumentaron enormemente la capacidad de secuenciado, pero mientras que dos de ellas (**Illumina** y **SOLiD**) eran de fragmentos cortos o muy cortos, la tercera presentaba problemas con los homopolímeros (**454**) y los fragmentos apenas alcanzaban el largo de los obtenidos por el método de Sanger.
- Las tecnologías NGS de tercera generación abarataron aún más el costo de secuenciación, universalizando el acceso a estas tecnologías. Mientras que una de ellas (**IonTorrent**) aún presenta problemas con los homopolímeros, las otras dos (**Oxford Nanopore** y **PacBio**) permiten obtener fragmentos de largos considerables, aún de decenas de miles de pares de bases, con relativamente alta precisión. Esto no solo permite identificar variaciones a nivel de las bases individuales, sino que también permite identificar variabilidad estructural.
- La cantidad de datos generados a bajo costo por las tecnologías de tercera generación y el largo de los fragmentos permiten acceder a otras técnicas, como la genómica funcional o la **metatranscriptómica**, que además de los aspectos composicionales y de dinámica de poblaciones permite identificar vías metabólicas expresadas de forma diferencial, lo que ayuda a comprender los cambios a nivel molecular entre condiciones (por ejemplo, entre ambientes).

7.8 Genes de resistencia

Los antimicrobianos han sido utilizados de forma masiva desde el descubrimiento de la penicilina por Alexander Fleming[17] y han ido incrementando sus aplicaciones, así como el uso de los mismos hasta nuestros días. Hoy en día es frecuente no solo encontrar el uso de antibióticos en seres humanos para curar y prevenir enfermedades, sino que también su uso es frecuente en la producción animal para el tratamiento de patologías, de forma preventiva y rutinaria y hasta como aditivo de crecimiento y producción alterando la microbiota ruminal (por ejemplo, el antibiótico ionóforo **monensina**, derivado de la bacteria *Streptomyces*

[17] Sir Alexander Fleming (6 de agosto de 1881 – 11 marzo de 1955) fue un médico y microbiólogo escocés. Descubridor en 1928 de un compuesto que llamó penicilina y que luego se llamaría bencilpenicilina (o penicilina G), obtenido a partir del hongo *Penicillium rubens*. Fue ganador del Premio Nobel de Medicina y Fisiología en 1945, junto a Howard Florey y Ernst Boris Chain.

cinnamonensis y usado ampliamente en la producción bovina). Pero como suele ocurrir con la vida en nuestro planeta, a medida que aumentamos la presión para eliminar determinados organismos, en las poblaciones que estos conforman comienzan a fijarse variantes que permiten a los individuos sobrevivir. A cada nuevo desarrollo humano en esta lucha por eliminar lo que para nosotros son organismos patógenos (ya sean nocivos para humanos, animales, plantas, etcétera) los organismos se han ido adaptando de diferentes maneras y desarrollando sus propios mecanismos de resistencia a nuestros embates tecnológicos. Más aún, a medida que se aumenta la presión sobre las poblaciones de patógenos, mayor será la necesidad de estos organismos de buscar alternativas para su supervivencia. Por esta razón, más allá del éxito enorme que han tenido los antibióticos para reducir el impacto de enfermedades de origen bacteriano, el mismo comienza a volverse peligrosamente inestable y gradualmente empezamos a perder el control en esta batalla. Nuestras victorias previas son solo el aliciente para una respuesta aún más dura por parte de los patógenos.

Como vimos previamente, los procariotas poseen varias ventajas en lo que hace al desarrollo de resistencia a los antimicrobianos. Por un lado, su dinámica poblacional sumamente flexible, con capacidad de multiplicarse de forma veloz y de alcanzar tamaños poblacionales enormes a partir de uno o pocos individuos. Esta habilidad les confiere la capacidad de soportar tasas de letalidad enormes y aun así poder reconstruir su tamaño poblacional previo apenas disminuye la presión selectiva o se encuentre una forma de evadirla. Por otro lado, existe una gran plasticidad a nivel genómico, así como la capacidad de intercambiar novedades evolutivas con gran facilidad (por ejemplo, a través del intercambio de plásmidos). A la vez, su gran tamaño poblacional y sus breves intervalos de duplicación (en relación con el de eucariotas macroscópicos) los hace candidatos a recibir un gran número de mutaciones por intervalo de tiempo en cada población, por lo que es de esperar la aparición de varios individuos por generación con alguna clase de ventaja selectiva (en un medio en que el cambio de factores de selección ha sido reciente). Además, ciertas bacterias poseen la capacidad de pasar de un modo de vida libre a una forma en que desarrollan un estado de protección mayor (a costo de la actividad), por ejemplo, a través de la formación de esporas o la organización en ***biofilms***, lo que hace que los ataques contra ellas estén destinados al fracaso. Finalmente, el hecho de que un individuo esté constituido por una sola célula, algo que podría verse como algo ya que la muerte de esa célula significa la muerte del organismo en su totalidad, constituye en realidad una gran ventaja en lo que refiere a la transmisión de las nuevas capacidades adquiridas. Por todas estas razones, como era de esperar, las bacterias han ido desarrollando distintos mecanismos para bloquear cada intento que la humanidad ha realizado para eliminarlas.

La resistencia a los antimicrobianos, conocida como **RAM** (**AMR**, por sus siglas en inglés) es la capacidad que desarrollan determinados microbios para resistir la acción de tratamientos que antes eran eficaces para su eliminación. La resistencia a antibióticos es un caso especial de **RAM**, referido a la resistencia adquirida por bacterias.

Existen cinco diferentes mecanismos que han ido evolucionando para este fenómeno de resistencia:

1. Inactivación o modificación de la droga mediante la acción enzimática en la bacteria.
2. Alteración del *target* o del sitio de unión de la droga.
3. Alteración de la vía metabólica en la bacteria, evitando, por ejemplo, el uso de la vía blanco de la droga.
4. Reducción de la concentración de la droga, ya sea por reducción de la permeabilidad a la misma o por aumento de su tasa de eflujo.
5. Desbloqueo y liberación de los ribosomas en bloqueos inducidos por antibióticos.

En el primer caso, un ejemplo conocido es el de la inactivación es la que corresponde a la expresión de β-lactamasas que son capaces de neutralizar los antibióticos β-lactámicos como las penicilinas, las cefalosporinas, monobactámicos y carbapenémicos (carbapenemasas). El elemento común a todas estas moléculas es un anillo químico conocido como anillo β-lactámico, el cual es fundamental para su acción antimicrobiana. Las β-lactamasas son enzimas que tienen la capacidad de hidrolizar dicho anillo. Un ejemplo del segundo mecanismo propuesto son las proteínas de unión a la penicilina, un conjunto de moléculas cuyo rol fundamental es la síntesis de peptidoglicanos para constituir la pared celular de bacterias; ante la presencia de penicilina estas proteínas se unen a la misma por su similitud con los peptidoglicanos, formando un enlace covalente y desactivando la droga. Para el tercer mecanismo, un ejemplo posible sería el de algunas bacterias resistentes a las **sulfonamidas**, que en lugar de usar el ácido para-aminobenzoico en la síntesis de ácido fólico (blanco terapéutico de la droga) y de nucleótido pueden cambiar a usar el ácido fólico disponible. En el caso del cuarto mecanismo, algunas bacterias poseen bombas en la membrana que permiten extraer antibióticos como las **quinolonas** antes de que alcancen una concentración dañina. Por último, a modo de ejemplo para el caso del quinto mecanismo, en la bacteria *Listeria monocytogenes* la secuencia **lmo0762** es un homólogo del gen ***hflX***, que codifica para una HSP (*heat shock protein*) que permite rescatar a los ribosomas bloqueados, separando sus dos subunidades; este mecanismo permite generar resistencia a los antibióticos lincomicina y eritromicina (Duval et al. 2018).

El proceso para la generación de resistencia y su expansión a nivel poblacional es sencillo. Frente a una infección de algún tipo, o simplemente producto de la aplicación rutinaria, los antibióticos generan una presión que tiende a eliminar a todos los organismos susceptibles. En general, las bacterias que poseen alguna clase de resistencia a antibióticos poseen una cierta desventaja selectiva frente al resto de los individuos de su especie ya que los mecanismos de resistencia suelen implicar un costo, la pérdida de una vía o modificaciones no-óptimas, por lo que en ausencia de la presión selectiva en general no predominan. Sin embargo, cuando aparece una fuerte presión selectiva las relaciones de **fitness** cambian de forma sustantiva, y aquellas células que no predominaban pasan ahora a dominar la escena, en particular por la ausencia de competidores, lo que implica que todos los recursos están a disposición de las mismas. El fenómeno puede acontecer aún en ausencia de bacterias de

dicha especie con resistencia ya adquirida en la población, ya sea gracias a su aparición por mutación aleatoria, como por existencia de eventos de intercambio de material genético con bacterias resistentes de otras especies (algunas que pueden ser no-patógenas, por ejemplo). Esto permitirá, en muchos casos, que aparezcan individuos resistentes en la especie y que además rápidamente aumenten su representación en la población. Más aún, frente al uso poco racional de antibióticos de diferentes clases, y a veces luego de una sucesión de fracasos por resistencia, puede generarse resistencia a múltiples drogas en bacterias, lo que se conoce como cepas **MDR** (*mutidrug resistance*).

La pregunta que nos surge es ¿cómo podemos enfrentar el problema de la generación de resistencia a los antibióticos? La mejor forma de enfrentar este problema es estratégica y está al alcance de la mano. Conociendo el organismo que vamos a atacar, así como los mecanismos de resistencia que posee, la mejor forma consiste en elegir un antibiótico de forma efectiva y disminuyendo el riesgo de generar resistencia. Antes se trataba de cultivar el organismo patógeno para luego poder hacer las pruebas bioquímicas que permitiesen identificarlo. Afortunadamente, hoy en día poseemos la tecnología como para resolver este problema en apenas unas horas. Algunos de los secuenciadores de segunda y tercera generación son capaces de generar secuencias en pocas horas y alguno en tiempo real (como el **MinION**) a un costo razonable y con poca necesidad de equipamiento muy costoso, por lo que el problema de diagnóstico puede quedar resuelto y solo sería importante agregar un monitoreo a nivel de toda la población para poder comprender las tendencias a nivel de variantes a dicha escala. Más aún, existen en la actualidad herramientas de proteómica masiva, usando equipos de MALDI-TOF[18], que permiten obtener estos resultados de forma estandarizada y en tiempos mucho menores, lo que refuerza la tendencia en esta área. Por otra parte, en lo que hace a la producción animal, es necesario prohibir el uso sistemático de antibióticos y monitorear también estrechamente la aplicación clínica de los mismos, ya que además de la posibilidad de generar resistencia en patógenos humanos, las excretas suelen estar contaminadas con antibióticos generando resistencia en bacterias ambientales. Sin duda, este es un tema que es motivo de preocupación y responsabilidad de todos, estrechamente vinculado al concepto de **una sola salud** (el cual contempla que todos los organismos estamos conectados al momento de elaborar el concepto de salud).

PARA RECORDAR

- La resistencia a los antimicrobianos (**AMR**) es un fenómeno extendido en el que los microbios desarrollan resistencia frente a tratamientos con drogas que anteriormente resultaban eficaces para su control y eliminación.
- La base genética del desarrollo de esta resistencia está en los elevados tamaños poblacionales de los microbios en general y de los procariotas en particular, así como en la flexibilidad de los genomas procariotas y los distintos mecanismos para compartir novedades evolutivas.
- Existen cinco diferentes mecanismos descritos que han ido evolucionando para este fenómeno de resistencia.
- En algunos casos, debido en parte al mal uso de los antibióticos, por ejemplo, a través de una secuencia de drogas todas con resultados fallidos, se genera en

[18] La tecnología MALDI-TOF (*matrix-assisted laser desorption/ionization time-of-flight*) es una técnica de espectrometría de masas que se utiliza para identificar y analizar moléculas como proteínas, péptidos y otros compuestos químicos. Funciona mediante la ionización de muestras y la medición del tiempo que tardan las partículas cargadas en moverse a través de un campo eléctrico.

las poblaciones bacterianas individuos con resistencia a múltiples drogas de familias diferentes, a los que se les conoce como **MDR** (*MultiDrug Resistant*). Estos organismos se constituyen en una enorme amenaza para los sistemas de salud ya que la posibilidad de diseminación de esta resistencia múltiple aumenta con el mal manejo de las terapias antibióticas.

- Las tecnologías de secuenciación masiva de segunda y especialmente de tercera generación son una excelente herramienta la el diagnóstico rápido de los patógenos causantes de infecciones, así como de las estrategias de resistencia que tienen desarrolladas, por lo que debería ser la alternativa inmediata para un uso racional de los antibióticos. Las tecnologías de proteómica masiva, con equipos MALDI-TOF son otra alternativa barata y ya estandarizada para este manejo racional.

7.9 Introducción a la epidemiología: modelos compartimentales

El estudio de las epidemias es fundamental para conocer la forma de acción y las estrategias de los patógenos, así como para entender la forma de combatirlas. Existen infinidad de modelos matemáticos desarrollados para comprender su evolución, así como métodos para determinar parámetros críticos de dichos modelos. Una clase de modelos frecuentemente usados son los **modelos compartimentales**, que permiten modelar diferentes estrategias y ciclos de una epidemia. En la figura 7.11 vemos un ejemplo de un modelo compartimental sencillo, que se conoce como **SIR con dinámica vital y tamaño poblacional constante**. Este modelo posee compartimentos para las categorías susceptible (S), infectados (I) y recuperados (R), y de ahí proviene su nombre. En el modelo **SIR** más simple, capaz de modelar epidemias como la de la gripe donde la dinámica de infección es tan rápida que no afecta la dinámica de nacimientos y muertes en la población, se suelen omitir estos términos, lo cual simplifica en gran medida las ecuaciones del modelo, pudiéndose describir soluciones explícitas a las ecuaciones diferenciales del mismo. Sin embargo, el modelo representado en la figura 7.11 es algo más complejo, porque además de los 3 compartimentos descritos anteriormente, se permite la muerte de los individuos a una tasa μ independiente del compartimento (es decir, el patógeno no afecta la tasa de mortalidad), por lo cual para mantener el tamaño poblacional constante debemos permitir que los nacimientos equilibren las muertes, o lo que es lo mismo establecer la igualdad $\Lambda = bN = \mu S + \mu I + \mu R = \mu(S + I + R) = \mu N$, donde b es la tasa de nacimientos por individuo en la población.

La dinámica de este sistema es fácil de describir a partir de un conjunto de tres ecuaciones diferenciales ordinarias acopladas unas con otras, cada una describiendo la dinámica de un compartimento. Es decir, la forma de cada una de las ecuaciones diferenciales debe describir la tasa de cambio del compartimento correspondiente, que es la diferencia entre los individuos que entran al compartimento en una unidad infinitesimal de tiempo, menos los individuos que abandonan el compartimento en ese instante. Claramente, si observamos nuevamente la figura 7.11, se trata de las cantidades en las **flechas que entran**, menos las cantidades en las **flechas que salen**. Por ejemplo, para el primer compartimento (S, el de los susceptibles, los individuos que aún no se han contagiado), tenemos que a este compartimento entrarán los nacimientos con número Λ (suponemos que los individuos nacen

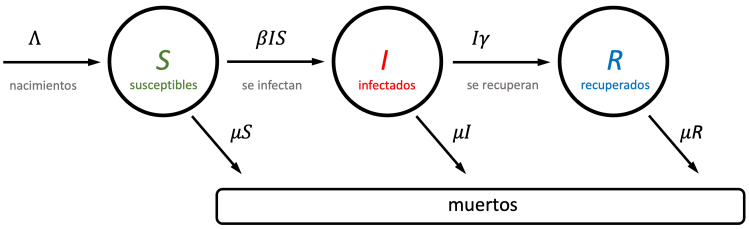


Figura 7.11. Modelo compartimental SIR con dinámica de vida. El modelo tiene 3 compartimentos que se corresponden con los del modelo SIR, más un compartimento que recibe a los muertos con tasa μ en todas las categorías y una tasa de nacimientos Λ.

no-contagiados), al mismo tiempo que abandonan el compartimento de dos formas: por muerte el compartimiento decrece un número μS (la tasa de muerte por la cantidad de individuos en el compartimento) y por infección un número βIS, siendo β la tasa de transmisión (o contacto) entre individuos. Poniendo todo junto, tenemos que la tasa de cambio en este compartimento es igual a

$$\frac{dS}{dt} = \Lambda - \mu S - \frac{\beta IS}{N} \tag{7.77}$$

Antes de seguir con el siguiente compartimento vamos a intentar explicar por qué el número de individuos contagiados es βIS, y por qué a esta forma se le conoce como **ley de acción de masas.** Si suponemos que los contagios se dan de persona contagiada a persona susceptible, y que por lo son proporcionales a las proporciones en ambas categorías, entonces el número de contagiados será proporcional a $N \times (\frac{S}{N}) \times (\frac{I}{N}) = S\frac{I}{N}$. Pero no todos los contactos producen enfermedad y, por lo tanto, precisamos de un parámetro β que nos permita variar la contagiosidad de las distintas enfermedades. Entonces, el número que abandona el compartimento de susceptibles será $\frac{\beta SI}{N}$.

De manera análoga, al compartimento de los infectados (I) entran los $\frac{\beta SI}{N}$ que abandonan el de susceptibles (S) y lo abandonan los que mueren en número μI, más los que pasan al compartimento de recuperados (R). Estos últimos lo hacen en número γI, siendo γ la tasa a la que se recuperan los infectados. Poniendo todo junto, la ecuación que describe este compartimento es

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I - \mu S \tag{7.78}$$

Finalmente el caso del último compartimento, el de los recuperados (R), es fácil de describir. En el mismo entran γI individuos que vienen del compartimento de infectados (I) y lo abandonan μR individuos que mueren, por lo que su correspondiente ecuación diferencial será

$$\frac{dR}{dt} = \gamma I - \mu S \tag{7.79}$$

El tiempo promedio de una infección estará dado por $\frac{1}{(\mu+\gamma)}$, y el número reproductivo básico por

$$R_0 = \frac{\beta}{(\mu+\gamma)} \tag{7.80}$$

El parámetro R_0 es fundamental para determinar la dinámica del sistema, ya que marca el nuevo número de individuos infectados que esperamos por cada infectado en un instante de tiempo.

El modelo que estamos analizando se trata de un sistema de ecuaciones diferenciales ordinarias no-lineal para el cual igualando a cero las ecuaciones anteriores y resolviendo obtenemos dos soluciones. La primera es la correspondiente al equilibrio libre de enfermedad (**DFE**, por sus siglas en inglés), que se alcanza cuando $R_0 \leqslant 1$ y es en este sistema igual a

$$e_1 = (S^*,I^*,R^*) = (N,0,0) \tag{7.81}$$

es decir, sin individuos en los compartimentos de infectados y recuperados. La segunda solución estará dada por $\lim_{t\to\infty}(S(t),I(t),R(t))$ y es igual a

$$e_2 = (S^*,I^*,R^*) = (\frac{N}{R_0},Nc_1(R_0-1),Nc_2(R_0-1)) \tag{7.82}$$

con $c_1 = \mu/\beta$ y $c_2 = \gamma/\beta$. El equilibrio de la ecuación (7.82) se corresponde al equilibrio endémico de la enfermedad, es decir cuando la misma alcanza un estado de equilibrio en el cual no es posible la erradicación, pero tampoco existe avance; este escenario acontece cuando se cumple $R_0 > 1$.

Claramente, el modelo que hemos manejado apenas representa de forma razonable algunos tipos de infección. Por ejemplo, en muchos casos los infectados tienen una tasa de mortalidad muy diferente a la de las otras categorías, por lo que deberíamos agregar un parámetro

al modelo. En otros casos, los recuperados vuelven a ser susceptibles al cabo de un tiempo, es decir, la inmunidad dura determinado tiempo, por lo que debemos introducir una flecha que vaya desde el compartimento R al compartimento S, así como un parámetro para indicar la tasa correspondiente. Además, no hemos modelado en ningún momento la estructura de edades de las poblaciones, y en muchos casos todas las tasas son específicas para cada grupo de edades, lo que claramente agrega muchos parámetros al sistema. Por estas razones al aumentar el realismo del modelo también aumentamos la complejidad del mismos y es necesario recurrir a herramientas computacionales para resolver los modelos. Por otra parte, modelos complejos con malos datos es una combinación catastrófica, ya que la sobreparametrización suele llevarnos a conclusiones completamente inestables e irreproducibles a medida que llegan nuevos datos.

PARA RECORDAR

- Los modelos epidemiológicos compartimentales son una clase de modelos que nos permite entender la dinámica de una epidemia.
- Los modelos **SIR** poseen tres compartimentos: susceptibles (S), infectados (I) y recuperados (R).
- En el caso del modelo **SIR con dinámica de vida y tamaño poblacional constante**, las ecuaciones diferenciales que lo describen son

$$\frac{dS}{dt} = \Lambda - \mu S - \frac{\beta IS}{N} \frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I - \mu S \frac{dR}{dt} = \gamma I - \mu S \tag{7.83}$$

con $\Lambda = bN = \mu(S + I + R) = \mu N$, ya que al ser el tamaño poblacional constante N los nacimientos deben igualar las muertes.

- En el modelo **SIR con dinámica de vida y tamaño poblacional constante** el tiempo promedio de una infección estará dado por $\frac{1}{(\mu+\gamma)}$ y el número reproductivo básico por

$$R_0 = \frac{\beta}{(\mu + \gamma)} \tag{7.84}$$

- Existen dos equilibrios estables para este modelo, uno correspondientes al estado libre de enfermedad (**DFE**) cuya solución es

$$e_1 = (S^*, I^*, R^*) = (N, 0, 0) \tag{7.85}$$

y otro correspondiente al estado endémico, cuya solución es

$$e_2 = (S^*, I^*, R^*) = \left(\frac{N}{R_0}, Nc_1(R_0 - 1), Nc_2(R_0 - 1)\right) \tag{7.86}$$

con $c_1 = \mu/\beta$ y $c_2 = \gamma/\beta$.

7.10 Conclusión

A lo largo de los capítulos anteriores, hemos establecido las bases conceptuales y matemáticas necesarias para abordar el estudio de la genética de poblaciones. Para cerrar esta sección del libro, nos hemos detenido en ciertos aspectos particulares que son de gran relevancia al investigar a los procariotas, los cuales constituyen la mayoría de los seres que habitan nuestro planeta y se agrupan en dos de los tres dominios de la vida (Bacteria y Archaea). Entre estas particularidades podemos resaltar su predominante estado haploide a nivel genómica, su reproducción asexual a través del mecanismo de fisión y sus métodos de intercambio genético, como la transferencia horizontal. Estas particularidades nos obligan a adaptar nuestros enfoques en comparación con los que solemos utilizar en el estudio de los eucariotas, tanto a nivel matemático (con modelos como los que vimos en este capítulo) como conceptual (al pensar en cómo las dinámicas poblacionales pueden generar fenómenos tales como la aparición de resistencia a antimicrobianos). La llegada de tecnologías de secuenciación ha revolucionado notablemente la investigación sobre los organismos procariotas. Dado que la mayoría de los procariotas no pueden ser cultivados en laboratorio, disciplinas como la metagenómica representan un avance significativo en la comprensión de la diversidad microbiana. Mientras que el microscopio nos permitió descubrir la existencia del mundo microbiano, la secuenciación ha permitido, en primer lugar, identificar su gran diversidad, que se manifiesta en dos dominios distintos y una variedad considerable de especies. En la actualidad, la secuenciación nos permite ensamblar miles de genomas y transcriptomas de organismos no cultivables con los cuales interactuamos a diario. En resumen, el estudio de estos organismos es esencial para comprender nuestro entorno cotidiano, y la genética de poblaciones proporciona una lente conceptual potente a través de la cual estudiar la diversidad microbiana y sus complejas dinámicas.

7.11 Actividades

7.11.1 Control de lectura

1. Describe las diferencias entre el modelo exponencial y logístico del crecimiento de poblaciones microbianas. ¿Qué fases existen en el crecimiento del último, y cómo se relacionan con los parámetros empleados en el modelo?
2. ¿Qué es el modelo de Moran? Desarrolla los puntos de similitud y discrepancia con el modelo de Wright-Fisher presentados en el capítulo «Deriva genética».
3. Se denomina a dos o más genes como *homólogos* si existe algún grado de similitud atribuible a la existencia de un ancestro común entre estos. Distintos fenómenos biológicos pueden llevar a la existencia de homología entre genes. Explica las características que definen a un conjunto de genes como ortólogos, parálogos o xenólogos, distinguiendo los escenarios que llevan a estos tipos de homología.
4. Los eucariotas también generan resistencia frente a compuestos químicos que impongan un régimen selectivo (por ejemplo, uso de plaguicidas en cultivos). ¿En términos generales, qué diferencias existen entre el desarrollo de resistencia a nivel poblacional en eucariotas y procariotas? Nombra algunas características biológicas de procariotas relevantes para analizar el fenómeno.
5. ¿Qué es la metagenómica? Nombra al menos un área de estudio de los organismos microbianos en los cuales los avances tecnológicos en esta área jueguen un rol importante.

7.11.2 ¿Verdadero o falso?

1. La replicación de células procariotas acontece por fisión. Un modelo matemático que no tenga en cuenta esto no es realista, en tanto predice un crecimiento exponencial para la población. El modelo logístico tiene en cuenta esto, siendo una aproximación matemática más realista.
2. A diferencia del modelo de Wright-Fisher, en el modelo de Moran la población se reproduce en poblaciones solapantes.
3. La expectativa de tiempos de coalescencia de un conjunto de alelos varía según el modelo matemático para modelar el proceso evolutivo, debido a que los mismos contemplan distintas escalas temporales. No obstante, es posible hacer escalados a nivel temporal para ajustar por este factor.
4. El proceso evolutivo de un conjunto de genes es siempre un fiel reflejo de la evolución de las especies involucradas.
5. Cuando aparece un agente antimicrobiano, la célula bacteriana responde generando una variante que le permite la evasión del mismo por distintos métodos, logrando así su supervivencia. Esto explica fenómenos como la resistencia a antibióticos, donde una población microbiana responde creando variantes que logran evadir la acción y generando así resistencia a la droga.

Solución

1. **Verdadero.** Bajo el modelo de crecimiento logístico la población crecerá de forma sigmoidal, pasando por una fase de crecimiento exponencial, una de crecimiento cuasi-lineal y finalmente disminuyendo su capacidad de crecimiento a medida que se aproxima a la capacidad de carga del sistema (parámetro que conceptualmente puede vincularse a la existencia de recursos limitados para la población). La constante intrínseca de crecimiento es en este contexto un parámetro que permite modelar la dinámica de crecimiento en cada caso específico (es decir tasa de crecimiento en las distintas fases).
2. **Verdadero.** El modelo de Moran se basa en la elección de un organismo de la población por vez, por lo que el recambio de los N individuos de una población no se realiza en un único paso (como sí sucede en la construcción del modelo de Wright-Fisher). En efecto, a *grosso modo* se esperan resultados similares para el modelo de Moran luego de renovar N individuos y las predicciones del modelo de Wright-Fisher. El hecho de que el modelo de Moran no requiera de la asunción de generaciones no-solapantes lo hace un modelo atractivo para el desarrollo de otros modelos teóricos en genética de poblaciones.
3. **Verdadero.** Este factor de escalado temporal de hecho está relacionado al tamaño poblacional efectivo, y varía según los modelos: para poblaciones grandes equivale a N_e en el caso del modelo de Wright-Fisher y el de fisión, mientras que en el caso del modelo de Moran equivale a $N_e^2/2$.
4. **Falso.** Esto acontece solo si el modo de herencia de la información genética es la transferencia vertical, y si no acontecen eventos de pérdida o ganancia de copias génicas. Un caso claro en el que esto no se cumple es cuando acontece la transferencia horizontal de genes, un proceso común en procariotas: los árboles filogenéticos obtenidos del análisis de estos genes reflejarán el proceso evolutivo de los genes, estimándose un mayor grado de similitud entre copias génicas que son producto de un evento de transferencia horizontal y no necesariamente reflejando la cercanía evolutiva entre las especies involucradas.
5. **Falso.** La aparición de resistencia a un agente microbiano en una población bacteriana es un fenómeno poblacional, y no se basa por lo tanto en un sistema explícito de detección de la droga y generación de una respuesta por parte de la célula bacteriana. Esta es una visión común y que parece dar ciertas características humanas al comportamiento microbiano. La preexistencia o la aparición de individuos portadores variantes que son menos susceptibles a la droga por mutación aleatoria es un fenómeno independiente de la presencia del antimicrobiano. Una vez aparece dicho agente, se establece un régimen selectivo donde las variantes menos susceptibles poseen mayor **fitness**, aumentando progresivamente el número de individuos portadores en la población. Variantes más eficientes se pueden fijar en la población, en tiempos que se ven favorecidos por la alta tasa de replicación bacteriana y su plasticidad genómica (existencia de mecanismos de transferencia genética).

7.11.3 Ejercicios

Ejercicio 7.1

En un laboratorio de microbiología se intenta caracterizar la dinámica de crecimiento de una cepa bacteriana. En el medio usado como referencia, la cepa posee una capacidad de carga de $K = 1 \times 10^5$ individuos. Las curvas de caracterización inician siempre con una célula microbiana.

a. Si el tiempo estimado para llegar a la mitad de la capacidad de carga es de 2,5 h, ¿cuál es la constante de crecimiento intrínseca de la cepa en estas condiciones?

b. ¿Cuánto tiempo se estima para alcanzar por lo menos el 99,9 % de K?

Solución

a. Notemos que

$$t_{1/2} = \frac{\ln(\frac{K-N_0}{N_0})}{r_c} \Rightarrow \\ r_c = \frac{\ln(\frac{K-N_0}{N_0})}{t_{1/2}} \tag{7.87}$$

Reemplazando con los valores dados, tenemos

$$r_c = \frac{\ln(\frac{10^5-1}{1})}{50} \\ r_c \approx 0{,}23 \tag{7.88}$$

b. Para calcular la cantidad de horas para alcanzar el valor $0{,}999 \times K$, utilizaremos la ecuación

$$n(t) = \frac{K\ n_0\ e^{r_c\ t}}{K + n_0\ (e^{r_c\ t} - 1)} \tag{7.89}$$

Sustituyendo con los valores dados, tenemos

$$0{,}999 \times K = \frac{K\ n_0\ e^{r_c\ t}}{K + n_0\ (e^{r_c\ t} - 1)} \tag{7.90}$$

$$0{,}999 \times 10^5 = \frac{10^5 \times 1 \times e^{0{,}23 \times t}}{10^5 + 1 \times\ (e^{0{,}23 \times\ t} - 1)} \tag{7.91}$$

$$0{,}999 \times 10^5 = \frac{10^5 \times e^{0{,}23 \times t}}{10^5 + e^{0{,}23 \times\ t} - 1} \tag{7.92}$$

$$(0{,}999 \times 10^5) \times (10^5 + e^{0{,}23 \times\ t} - 1) = 10^5 \times e^{0{,}23 \times t} \tag{7.93}$$

$$0{,}999 \times (10^5 + e^{0{,}23 \times\ t} - 1) = e^{0{,}23 \times\ t} \tag{7.94}$$

$$0{,}999 \times 10^5 + 0{,}999 \times e^{0{,}23 \times\ t} - 0{,}999 = e^{0{,}23 \times\ t} \tag{7.95}$$

$$0{,}999 \times 10^5 - 0{,}999 = e^{0{,}23 \times\ t} - 0{,}999 \times e^{0{,}23 \times\ t} \tag{7.96}$$

$$0{,}999 \times 10^5 - 0{,}999 = 0{,}001 \times e^{0{,}23\times\ t} \tag{7.97}$$

$$\frac{0{,}999 \times 10^5 - 0{,}999}{0{,}001} = e^{0{,}23\times\ t} \tag{7.98}$$

$$ln(\frac{0{,}999 \times 10^5 - 0{,}999}{0{,}001}) = ln(e^{0{,}23\times\ t}) \tag{7.99}$$

$$ln(\frac{0{,}999 \times 10^5 - 0{,}999}{0{,}001}) = 0{,}23 \times\ t \Rightarrow \tag{7.100}$$

$$t = \frac{ln(\frac{0{,}999\times 10^5 - 0{,}999}{0{,}001})}{0{,}23} \tag{7.101}$$

$$t \approx 80{,}1\ h \tag{7.102}$$

El valor obtenido refleja la no-linearidad del proceso de crecimiento: en cerca de cincuenta horas alcanzamos la mitad de la capacidad de carga, pero se requiere bastante menos del doble de tiempo para encontrarse extremadamente cerca de la misma (coherente con los gráficos tipo sigmoide del modelo logístico de crecimiento).

Ejercicio 7.2
(Inspirado en Hartl 2020)

Demuestra que bajo el modelo de Moran la heterocigosidad esperada para un locus bialélico segregando en una población de tamaño N decae el doble de rápido que en una población que lo predicho según el modelo de Wright-Fisher aplicado a una población haploide de igual tamaño.

Solución

Para ver cómo evoluciona la heterocigosidad esperada según el modelo de Moran podemos tomar un enfoque similar al utilizamos cuando detallamos el modelo de Wright-Fisher. Veremos primero cuál es la heterocigosidad esperada luego de un paso en el modelo, y a partir de allí veremos qué predice el modelo para la heterocigosidad luego de un número arbitrario de pasos (en el caso del modelo de Wright-Fisher, esto eran t generaciones arbitrarias). Según el modelo de Moran, si se tienen i copias de un alelo las probabilidades de pasar $i+1$, mantenerse en i copias o pasar a $i+1$ copias son

$$\begin{cases} p_{i,i-1} = \frac{(N-i)\times i}{N^2} \\ p_{i,i} = \frac{i^2+(N-i)^2}{N^2} \\ p_{i,i+1} = \frac{i\times(N-i)}{N^2} \end{cases} \tag{7.103}$$

¿Qué implica esto a nivel de las frecuencias alélicas? Notemos que si el número de copias es i para un alelo, su frecuencia en la población a tiempo t es $p = \frac{i}{N}$, siendo la frecuencia del alelo alternativo $q = \frac{N-i}{N}$ (contamos con N alelos por tratarse de una población haploide).

Notemos que las probabilidades anteriores pasan a expresarse en función de las frecuencias alélicas como

$$\begin{cases} p_{i,i-1} = \frac{(N-i)\times i}{N^2} = (\frac{N-i}{N}) \times (\frac{i}{N}) = (1-p) \times p = qp \\ p_{i,i} = \frac{i^2+(N-i)^2}{N^2} = \frac{i^2}{N^2} + frac(N-i)^2 N^2 = \\ = (\frac{i}{N})^2 + (frac(N-i)N)^2 = p^2 + (1-p)^2 = p^2 + q^2 \\ p_{i,i+1} = \frac{i\times(N-i)}{N^2} = (\frac{i}{N}) \times (\frac{N-i}{N}) = p \times (1-p) = pq \end{cases} \tag{7.104}$$

Luego de una ronda del proceso propuesto por el modelo, la heterocigosidad esperada será

$$H_{t+\delta t} = 2(p + \frac{1}{N})(1 - p - \frac{1}{N})[pq] + 2pq[p^2 + q^2] + 2(p - \frac{1}{N})(1 - p + \frac{1}{N})[qp] \tag{7.105}$$

donde, como en todo cálculo de esperanza, a los valores posibles los multiplicamos por su probabilidad (entre paréntesis rectos en la ecuación anterior). Es importante notar que por cómo se propone el modelo se abarcó en este cálculo a todos los valores de heterocigosidad posibles luego de un paso de replicación y muerte de una célula haploide.

Si desarrollamos el cálculo, llegamos a que

$$H_{t+\delta t} = 2pq(1 - \frac{2}{N^2}) \tag{7.106}$$

Pero en el punto de partida t la heterocigosidad es justamente $H_t = 2pq$, por lo que la ecuación anterior implica

$$H_{t+\delta t} = H_t \times (1 - \frac{2}{N^2}) \tag{7.107}$$

Llegado este punto llegamos a una recursión similar a la que obtuvimos cuando desarrollamos el modelo de Wright-Fisher. Tenemos

$$H_{t+2\delta t} = H_t \times (1 - \frac{2}{N^2})^2 \tag{7.108}$$

Y si repetimos el proceso hasta reemplazar a un número N individuos (la idea detrás del modelo de Wright-Fisher, generar una nueva generación de individuos), obtendremos

$$\begin{aligned} H_{t+N\delta t} = H_t \times (1 - \frac{2}{N^2})^N \approx H_t \times (1 - \frac{2N}{N^2}) \\ H_{t+N\delta t} \approx H_t \times (1 - \frac{2}{N}) \end{aligned} \tag{7.109}$$

El modelo de Wright-Fisher para una población haploide de tamaño N implica un cambio de heterocigosidad de una generación a la siguiente de

$$H_{t+1} \approx H_t \times (1 - \frac{1}{N}) \tag{7.110}$$

donde queda claro que el modelo de Moran predice una pérdida de heterocigosidad el doble de rápido.

Ejercicio 7.3

Un grupo de investigación detectó la existencia de una cepa de *Saccharomyces cerevisiae* que posee un alto potencial industrial para la elaboración de bebidas alcohólicas. Debido a que la cepa no logra aislarse de forma pura al momento, se decide investigar para buscar condiciones que favorezcan su crecimiento en el medio utilizado para la fermentación. Afortunadamente se identificó un gen marcador (al cual llamaremos A) para el cual existen dos alelos que separan claramente a la variante de interés (portador del alelo A_1, de frecuencia p) y otras cepas.

Se realiza un experimento de secuenciación masiva con el objetivo de analizar la proliferación bajo una condición dada, secuenciando a dos tiempos (se estima que a una generación de la población) con tres réplicas cada uno. Se asume que la profundidad de secuenciado (número de bases) en ambos experimentos es la misma.

Variante	$T1_1$	$T1_2$	$T1_3$	$T2_1$	$T2_2$	$T2_3$
A_1	6555	6520	6590	6980	6915	7030
A_2	3460	3510	3435	3135	3180	3055

En las columnas A y B se denotan a las condiciones experimentales estudiadas.

¿En cuál de las condiciones experimentales posee mayor fitness relativo la cepa de interés, y en qué magnitud?

Solución

a. Para determinar en qué condición experimental la cepa de interés tiene un mayor fitness relativo y en qué magnitud, utilizamos la ecuación

$$\Delta_s p = \frac{s_d \times p \times q}{1 + s_d \times p} \tag{7.111}$$

Primero, calcularemos las frecuencias p y q en cada tiempo. Para ello vamos a calcular las frecuencias en cada réplica de cada tiempo, y luego promediar esos valores para obtener una estimación más precisa de las frecuencias de interés.

Para el Tiempo 1 entonces tenemos

$$\begin{cases} p_{\text{T1_1}} = \frac{6555}{6555+3460} \approx 0{,}654 \\ p_{\text{T1_2}} = \frac{6520}{6520+3510} \approx 0{,}650 \\ p_{\text{T1_3}} == \frac{6590}{6590+3435} \approx 0{,}657 \end{cases} \tag{7.112}$$

de lo cual tenemos $\bar{p}_{T1} = \frac{p_{T1_1}+p_{T1_2}+p_{T1_3}}{3} \approx 0{,}6537 \Rightarrow \bar{q}_{T1} \approx 1 - 0{,}6537 = 0{,}3463$.

De forma análoga se llega a $\bar{p}_{T2} \approx 0{,}6903 \Rightarrow \bar{q}_{T2} \approx 1 - 0{,}6903 = 0{,}3097$.

A partir de estos valores podemos calcular $\Delta_s p$, que representa el cambio en la frecuencia de A_1 entre tiempos, el cual atribuimos a la selección. Tenemos

$$\Delta_s p = \bar{p}_{T2} - \bar{p}_{T1} = 0{,}6903 - 0{,}6537 = 0{,}0366 \tag{7.113}$$

Finalmente, despejamos s_d utilizando la ecuación mencionada al comienzo.

$$\Delta_s p = \frac{s_d \times p \times q}{1 + s_d \times p} \tag{7.114}$$

$$\Delta_s p \times (1 + s_d \times p) = s_d \times p \times q \tag{7.115}$$

$$\Delta_s p + \Delta_s p \times p \times s_d = s_d \times p \times q \tag{7.116}$$

$$\Delta_s p = s_d \times p \times q - s_d \times \Delta_s p \times p \tag{7.117}$$

$$\Delta_s p = s_d \times (p \times q - p \times \Delta_s p) \Rightarrow \tag{7.118}$$

$$s_d = \frac{\Delta_s p}{(p \times q - p \times \Delta_s p)} \tag{7.119}$$

$$s_d \approx \frac{0{,}0366}{0{,}6537 \times 0{,}3463 - 0{,}6537 \times 0{,}0366} \tag{7.120}$$

$$s_d \approx 0{,}18 \tag{7.121}$$

En efecto la condición probada parece dar ventaja a la cepa de interés, la cual posee una diferencia en fitness relativo de aproximadamente 0,18.

Ejercicio 7.4

En una población microbiana se encuentra segregando un locus bialélico (al cual nos referiremos como A). La población tiene un tamaño efectivo aproximado de $N_e = 10.000$, siendo las frecuencias alélicas $p = 0{,}54$ y $q = 0{,}46$ para los alelos A_1 y A_2, respectivamente.

a. Compara el valor de heterocigosidad (virtual) esperado según el modelo de Moran respecto al de Wright-Fisher luego de 100, 1000 y 10.000 generaciones.

b. Tomando como punto de partida la situación descrita, ¿cuál es la probabilidad de que dos alelos no coalescan luego de haberse reemplazado 100 individuos de la población (1 % de esta)?

Solución

a. El valor de heterocigosidad *virtual* esperado según el modelo de Moran luego de una generación es de

$$H_{t+1} = H_t \times (1 - \frac{2}{N}) \tag{7.122}$$

por lo que luego de t generaciones esperamos según este modelo un valor de

$$H_t = H_0 \times (1 - \frac{2}{N})^t \tag{7.123}$$

donde H_0 es la heterocigosidad virtual en la primera generación.

Si aplicamos el modelo de Wright-Fisher a una población haploide, tenemos que la caída de heterocigosidad esperada en una generación es

$$H_{t+1} = H_t \times (1 - \frac{1}{N}) \tag{7.124}$$

Por razonamiento análogo, tenemos luego de t generaciones

$$H_t = H_0 \times (1 - \frac{1}{N})^t \tag{7.125}$$

Como ya se vio en otro ejercicio, el modelo de Moran predice una caída de la heterocigosidad dos veces mayor que el modelo de Wright-Fisher cuando se lo aplica a una población haploide. A efectos de realizar las comparaciones pedidas, se verá el cociente entre los valores esperados, el cual corresponde a

$$\begin{aligned} \frac{H_t^{\text{W-F}}}{H_t^{\text{Moran}}} &= \frac{H_0 \times (1 - \frac{1}{N})^t}{H_0 \times (1 - \frac{2}{N})^t} \\ \frac{H_t^{\text{W-F}}}{H_t^{\text{Moran}}} &= \frac{(1 - \frac{1}{N})^t}{(1 - \frac{2}{N})^t} \end{aligned} \tag{7.126}$$

Reemplazando con el valor de tamaño poblacional dado y los tiempos requeridos se obtiene

$$\begin{cases} H_{100}^{\text{W-F}} / H_{100}^{\text{Moran}} \approx 1{,}01 \\ H_{1000}^{\text{W-F}} / H_{1000}^{\text{Moran}} \approx 1{,}11 \\ H_{10000}^{\text{W-F}} / H_{10000}^{\text{Moran}} \approx 2{,}71 \end{cases} \tag{7.127}$$

Vemos por lo tanto que a medida que se empiezan a considerar una mayor cantidad de generaciones la estimación de heterocigosidad virtual comienza a discrepar entre los modelos, llegando a ser la heterocigosidad esperada según el modelo de Wright-Fisher casi tres veces mayor luego de 10.000 generaciones.

b. Recordemos que haciendo un escalado temporal la probabilidad de no coalescencia a tiempo v está dada aproximadamente por la siguiente ecuación

$$P(T_k > v) \approx e^{-\frac{k(k-1)}{2}v} \tag{7.128}$$

El factor de escalado temporal a usar en cada modelo varía y está relacionado con el tamaño poblacional; en el caso del modelo de Moran el escalado factor de escalado es de $N^2/2$. Deshaciendo el escalado temporal, tenemos la probabilidad de no coalescencia según el

modelo de Moran teniendo en cuenta el tamaño de la población, $P(T_k^N > t)$, la cual estará dada por

$$P(T_k^N > t) \approx e^{-\frac{k(k-1)}{2} \times \frac{2}{N^2} \times t}$$
$$P(T_k^N > t) \approx e^{-\frac{k(k-1)}{N^2} \times t} \tag{7.129}$$

donde se deshizo el escalado temporal teniendo en cuenta que $-\frac{k(k-1)}{2} \times t = -x \times \frac{N^2}{2} \times t \Rightarrow x = \frac{k(k-1)}{2} \times \frac{N^2}{2}$.

Recordemos que k equivale al número de alelos; dado que la población consta de 10.000 individuos haploides, el número de alelos será $k = 10.000$. Teniendo esto en cuenta, podremos realizar el cálculo en la escala temporal del modelo de Moran. La probabilidad de no-coalescencia luego de mil reemplazos estará dada por

$$P(T_k^N > 100) \approx e^{-\frac{10.000 \times (10.000-1)}{10.000^2} \times 100}$$
$$P(T_k^N > 100) \approx 3{,}76 \times 10^{-44} \tag{7.130}$$

Ejercicio 7.5

En una población de aproximadamente **54** millones de personas ingresa un patógeno. Luego de un tiempo, un grupo interdisciplinario se dispone a estudiar la dinámica de transmisión en la población, logrando determinar una tasa de transmisión de $\beta = 0{,}12$ contagios per cápita/día, una tasa de recuperación de $\gamma = 0{,}1$ recuperados per cápita/día y una tasa de mortalidad de $\mu = 0{,}001$ muertes per cápita/día.

1. ¿Qué dinámica de transmisión espera en la población con estos datos?
2. Determina la cantidad de individuos sanos, cursando enfermedad y recuperados en el largo plazo, una vez se alcance el equilibrio. Se asume que las condiciones generales de la situación epidemiológica se ven inalteradas.
3. ¿A qué tasa de transmisión máxima se debe llegar si se quiere que la población pueda erradicar la enfermedad? Asume que no es posible aumentar la tasa de recuperados, y que no es deseable aumentar la tasa de mortalidad.

Solución

a. Para determinar la dinámica de transmisión en la población, primero calcularemos el número básico de reproducción (R_0) utilizando la fórmula

$$R_0 = \frac{\beta}{\gamma + \mu} \tag{7.131}$$

Sustituyendo estos valores en la ecuación tenemos

$$R_0 = \frac{0{,}1199}{0{,}1 + 0{,}001} = \frac{0{,}1199}{0{,}101} \approx 1{,}186 \tag{7.132}$$

Valores $R_0 < 1$ indicarían que la enfermedad no puede sostenerse en la población y eventualmente desaparecería. Los valores de $R_0 \geq 1$ sugieren que la enfermedad puede propagarse en la población. En este caso, con un valor cercano a 1 la propagación sería relativamente lenta, pudiendo mantenerse a largo plazo el patógeno circulando en la población.

b. Para determinar la cantidad de individuos sanos, cursando la enfermedad y recuperados a largo plazo (cuando se alcance el equilibrio), utilizaremos la ecuación de equilibrio proporcionada para el caso $R_0 \geq 1$:

$$e_2 = (S^*, I^*, R^*) = (\tfrac{N}{R_0}, N \times c_1 \times (R_0 - 1), N \times c_2 \times (R_0 - 1))$$

donde N es la población total (**54** millones de personas) y las constantes c_1 y c_2 están dadas por $c_1 = \mu \times \beta$ y $c_2 = \gamma \times \beta$.

Sustituyendo los valores, obtenemos

$$\begin{cases} S^* = \frac{N}{R_0} = \frac{54.000.000}{1{,}1861} \approx 45.556.246 \\ I^* = N \times c_1 \times (R_0 - 1) = 54.000.000 \times (0{,}001 \times 0{,}1199) \times (1{,}1861 - 1) \approx 6.276 \\ R^* = N \times c_2 \times (R_0 - 1) = 54.000.000 \times (0{,}1 \times 0{,}1199) \times (1{,}1861 - 1) \approx 67.116 \end{cases} \tag{7.133}$$

siendo estos los valores estimados para S^*, I^* y R^*, la cantidad estimada de personas sanas, cursando la infección y recuperadas en el estado de equilibrio, respectivamente.

c. Para que la población pueda erradicar la enfermedad, es necesario que el número básico de reproducción $R_0 < 1$. Recordemos que $R_0 = \frac{\beta}{\mu+\gamma}$. Por lo tanto, tenemos

$$\begin{aligned} R_0 &< 1 \\ \frac{\beta}{\mu + \gamma} &< 1 \\ \beta < \mu + \gamma\beta &< 0{,}001 + 0{,}1 = 0{,}101 \end{aligned} \tag{7.134}$$

Por lo tanto, si se logra disminuir la tasa de transmisión a menos de 0,101 contagios/día aproximadamente, se espera a largo plazo la erradicación de la enfermedad (si no existe un cambio en las condiciones dadas).

Parte III: Genética cuantitativa

Capítulo 8

El modelo genético básico

De perto ninguém é normal.[1]

Caetano Veloso

En la última parte de nuestro curso nos enfocaremos en la genética cuantitativa, es decir, la genética de características que en principio son cuantificables. En general, a nivel agronómico, la mayor parte de las características que deseamos mejorar son de este tipo y por eso la importancia en nuestro curso. Ejemplos de las mismas que han experimentado un «avance» sostenido durante el siglo XX, avance mayormente determinado por la selección artificial, lo constituyen la producción de leche en bovinos, el tamaño de «pollos parrilleros», la concentración de proteína y grasa en la leche, la producción de la mayor parte de los cultivares y casi todas las características que han mostrado cierta respuesta a la selección.

A diferencia de los modelos que hemos utilizado en la Parte II, en general el modelo de un locus con dos alelos no alcanzará para explicar la variación cuantitativa observada, y deberemos construir modelos más generales, con menos detalle al nivel de locus, pero con mucha mayor capacidad de explicar los aportes de las distintas componentes que afectan la expresión de una característica. En general, vamos a trabajar asumiendo que el fenotipo observado de una característica se encuentra explicado por los aportes de la genética del individuo y su ambiente, así como por la interacción de la genética y el ambiente. Más aún, pasaremos de modelos centrados en el individuo, para los que era necesario conocer la configuración genética y ambiental completa, a pensar en los individuos como integrantes de poblaciones, que tienen una distribución aleatoria particular y, por lo tanto, nos enfocaremos en hacer predicciones sobre los mismos y su descendencia. Como notarás rápidamente, esta última parte del curso recurrirá a un tratamiento estadístico de la mayor parte de los problemas y, por lo tanto, es necesario tener frescos los conceptos de distribuciones, esperanzas, varianzas, momentos, correlación y regresión, entre otros. Si bien existen diferentes materiales que explican mejor o complementan lo escrito por nosotros, gran parte de lo que trabajaremos en esta última parte del curso está basado en el libro de Falconer y Mackay (1996) (o en la vieja versión en castellano, Falconer 1983), que sigue siendo una referencia hoy en día. Un excelente complemento se encuentra en el libro de Lynch y Walsh (1998), aunque el nivel de profundidad es bastante mayor.

Todos somos capaces de observar a diario que más allá de la similitud que presentan determinados organismos cuando los comparamos con otros muy diferentes (por ejemplo, de otra especie), la misma es bastante parcial, y reconocemos fácilmente que ningún ser vivo es

[1] Parte de la letra de la canción *Vaca profana*, que Caetano escribió para Gal Costa y que habla de la diversidad, defendiéndola de los «puretas», destacando que la «normalidad» puede ser apenas el efecto de la distancia del observador. Capaz que viene a cuento en este capítulo que vamos a recurrir a la normalidad de diferentes maneras, en particular a través del teorema del límite central.

exactamente igual a otro ser vivo. Esto nos lleva a intentar entender la base de la similitud y de las diferencias entre organismos. Al mismo tiempo que (en general) somos capaces de distinguir claramente entre hermanos, también reconocemos que cada uno se parece más a un progenitor que a otro; lo que es más, en general se parecen más a alguno de sus progenitores que al resto de las personas del barrio. Además, desde los experimentos de Gregor Mendel, hemos venido confirmando la base genética de muchas características, y aún somos capaces de predecir razonablemente bien cómo será la descendencia de determinada pareja de progenitores. Sin duda esto es posible para características discretas gobernadas por un locus (o a lo sumo muy pocos loci): si se tiene la información referente a los distintos alelos existentes, esta tarea es un ejercicio bastante sencillo. Sin embargo, cuando las características presentan variación continua (por ejemplo, pesos, alturas, producción de leche, volumen, etcétera) el sencillo **modelo mendeliano** básico ya no puede dar cuenta de las diferencias observadas y necesitamos construir un modelo que tenga en cuenta tanto los aspectos genéticos como los efectos ambientales que afectan la expresión de la característica estudiada.

En el presente capítulo vamos a elaborar un modelo que nos permita interpretar la variabilidad fenotípica (es decir, lo observable) a la luz de los componentes genéticos y ambientales que la afectan. Para ello, primero introduciremos el tema de los tipos de variación, continua y discreta, así como el hecho de que la suma de diferentes (muchas) variables aleatorias produce una variación casi-continua (lo que se conoce como el **teorema del límite central**). Esto a su vez nos permitirá vincular el modelo mendeliano de herencia con los enfoques biométricos de la misma, lo que fue un hito en la construcción de la **síntesis darwiniana**, permitiendo resolver varias contradicciones de ambos enfoques previos. Más aún, esto será la base del **modelo infinitesimal**, que supone que las características continuas pueden consistir en la suma infinitesimal de los aportes de un número infinito de loci. Si bien totalmente incorrecto a la luz de nuestro conocimiento (el tamaño del genoma es finito), este enfoque permite construir modelos sencillos pero de gran poder predictivo y para la comprensión de los fenómenos subyacentes.

A partir de nuestro modelo genético básico, que asume que el fenotipo de cualquier individuo es la suma de los efectos genéticos y ambientales en el mismo, más la interacción entre ambos tipos de efectos, iremos disgregando tanto los efectos genéticos como los ambientales en componentes que nos permitan entender sus interrelaciones. Veremos que podemos descomponer los efectos genéticos en *i)* **efectos de carácter aditivo** (es decir que se puede sumar directamente el efecto de cada alelo presente en el individuo), *ii)* los desvíos de la aditividad de la combinación particular de alelos que tiene cada genotipo, que llamaremos **desvío de dominancia**, y finalmente *iii)* los desvíos de interacción de los efectos entre diferentes loci, que llamaremos **efectos epistáticos**. Además, para los efectos aditivos y de dominancia, analizaremos en detalle el modelo de un **gen de efecto mayor**, donde una característica está básicamente gobernada por un solo gen con dos alelos. Finalmente, extenderemos nuestros análisis al nivel de las poblaciones, buscando entender la importancia relativa de las distintas componentes a partir de las varianzas que estas generan.

OBJETIVOS DEL CAPÍTULO

☐ Introducir el marco conceptual de la genética cuantitativa.

☐ Presentar el modelo matemático con el que trabajaremos en capítulos posteriores: el modelo genético básico.

☐ Discutir resultados obtenidos al aplicar el modelo al caso de un locus con dos alelos.

☐ Introducir los conceptos de efecto medio, valor reproductivo y desvío de dominancia.

☐ Tener una primera aproximación a un problema complejo: analizar la interacción entre genotipo y ambiente.

☐ Descomponer la varianza genética para poder analizar el aporte relativo de las fuentes que generan dicha variación.

8.1 Variación continua y discreta

Hasta el momento hemos basado todo el desarrollo del curso casi exclusivamente en razonamientos obtenidos a partir de un locus o de un par de loci con dos alelos. En organismos haploides, suponiendo que el efecto de cada alelo sea constante en un determinado momento en el tiempo (o sea, que no dependa de otros factores), apenas se pueden tener dos genotipos distintos, y poco cambia la situación el que los organismos sean diploides (donde aparecen tres genotipos posibles). Cuando el tipo de característica que analizamos es categórico, por ejemplo, el color verde o amarillo de los guisantes de Mendel (*Pisum sativum*), todo los resultados entran en alguna de las categorías permitidas o definidas, y entonces mientras el número de categorías sea menor o igual al número de genotipos tenemos alguna esperanza de vincular el genotipo con el fenotipo. Es decir, si las categorías son «verde» y «amarillo», es porque asumimos que las pequeñas variaciones de color son irrelevantes para nosotros o sencillamente no existen (cosa que parece totalmente improbable) y solo tenemos que asociar algún (o algunos) genotipos con un color y el resto con el otro. Claramente, en un locus diploide con dos alelos podemos encontrar como máximo tres genotipos, por lo que podría codificarse un nuevo color de guisantes tan extraño como el «tan» (sí, es un color), o un color intermedio entre el verde y el amarillo; las diferencias entre estas alternativas están en la base bioquímica del color, pero no existe en nuestro modelo una imposibilidad matemática para codificar estos genotipos. Sin embargo, cuando la paleta de colores de nuestros guisantes hipotéticos es un poco más amplia (por ejemplo, 14 colores), comienza a existir una imposibilidad para describirla directamente con un **mapeo unívoco** de genotipos a colores. Dicho de otra forma, cada genotipo debería estar asociado a más de un color, lo que implica involucrar otros tipos de factores, ya sean ambientales o genéticos; resulta claro que, en último caso, estos factores genéticos se encontrarían fuera del locus estudiado. Pese

a ello, seguimos frente a un tipo de variación **categórica**, ya que nuestra paleta de colores tiene un número finito de colores, y no tendría sentido en este contexto hablar de matemático la distancia entre esos colores.

Claramente esto podemos «arreglarlo» invocando otros loci, o sencillamente colocando más alelos en nuestro locus original; si suponemos además que cada combinación de dos alelos produce un color diferente, entonces con k alelos podemos formar $n = k(k+1)/2$ genotipos diploides distintos (incluyendo los homocigotos). Por ejemplo, con $k = 5$ alelos podemos formar $n = 5(5 + 1)/2 = 15$ genotipos diferentes, los cuales nos permitirían codificar esos 14 colores, sobrando incluso la capacidad de acomodar uno nuevo. Todo esto es apenas una digresión acerca de las posibilidades matemáticas de nuestros modelos, pero obviamente la base bioquímica y molecular de la herencia del color será la que mande, y nuestro modelo deberá ser capaz de representar esa realidad.

Otro tipo de variables de interés agronómico y biológico suelen producir resultados que son contables pero que, ahora sí, tiene sentido plantear y analizar las distancias entre las observaciones involucradas. Por ejemplo, el número de granos de maíz en una mazorca es en principio cuantificable, y si una mazorca produce 483 granos y otra 497, puede calcularse la diferencia como $497 - 483 = 14$ granos. Claramente, no podía realizar esta operación con los colores, al menos como los definimos previamente en **categorías**.[2] A las características que tienen este tipo de variación se les llama a veces **merísticas**. Poseen una distribución **discreta**, en contraposición a **continua**; son además del tipo de variables cuyo resultado se obtiene de contar el número de elementos que hay en ellas y que los mismos (los elementos) siguen un ordenamiento natural, donde la «distancia» entre elementos tiene sentido. Como vimos antes, en principio alcanzaría con $k = \frac{-1+\sqrt{1+8n}}{2}$ (el resultado de resolver la ecuación $n = k(k + 1)/2$) alelos en un locus para codificar todas las n categorías posibles de una característica discreta como el color. Sin embargo, esto tiene poco sentido en variables discretas donde cada combinación (genotipo) debería acomodarse en el ordenamiento discreto posible. Por ejemplo, para acomodar la posibilidad de tener entre 1 y 528 granos por mazorca alcanzaría en un locus diploide con 32 alelos, pero sus combinaciones deberían de alguna manera tener un ordenamiento lógico (a nivel bioquímico) que se traduzca en conteos del 1 al 528.

Una alternativa más razonable parece ser invocar más loci y asumir, en principio, que los efectos de los alelos son aditivos (es decir, se pueden sumar), tanto dentro de cada locus como entre loci. En la figura 8.1 podemos ver el resultado de una simulación particular de este modelo, para un solo locus (izquierda arriba), así como para 5, 10 y 50 loci en un gen de organismos diploide, con dos alelos en cada loci y frecuencias de $p = q = \frac{1}{2}$ en cada uno. Para cada situación se simularon los valores de 100.000 individuos, suponiendo además que cada copia de uno de los dos alelos suma una unidad (1) a la característica, mientras que el otro alelo no suma nada (0).

Claramente, con 1 locus solo hay 3 genotipos posibles, es decir 0, 1 o 2 copias del alelo que suma una unidad y, por lo tanto, los valores posibles son 0, 1 y 2. Con 5 loci tenemos mucha más variación posible. Primero, los límites se darán cuando se consideren 0 copias

[2] Los colores pueden llevarse a vectores numéricos que permiten una aproximación más cuantitativa a los mismos, permitiendo distinguir entre los infinitos matices que existen entre los mismos (¡los *inuit* cuentan con cincuenta palabras para los tonos de blanco!). Ejemplos de esta aproximación a los colores son formas de cuantificación como los modelos de color aditivo RGB y CYMK. De todas formas, resulta claro que entrar en este grado de detalle y análisis cuantitativo no tendría sentido para analizar la herencia de colores planteada en el esquema mendeliano clásico.

del alelo que suma (es decir, valor total 0) para el valor mínimo, y como valor máximo para el caso donde en los 5 loci se encuentren en homocigosis las copias que suma (es decir, valor total de $5 \times 2 = 10$ unidades). Pero por otro lado, comenzamos a apreciar un fenómeno interesante: la distribución de los valores no es uniforme, y los valores centrales tienen mayor cantidad de datos (individuos que tienen ese valor genotípico), al tiempo que a medida que se consideran los extremos del rango de valores posibles las barras disminuyen de tamaño (es decir, hay menos individuos que presentan estos valores).

La situación se ve aún más clara cuando tenemos 10 loci. Con la misma cuenta que antes para 5, tenemos que los límites de valores serán 0 cuando el individuo no tenga ningún alelo que aporte valor, a un máximo de 20 (2 copias en los 10 loci). Además, la tendencia sobre la forma de la distribución instalada en la gráfica de 5 loci se ve plenamente confirmada en la gráfica de 10 loci, con la masa concentrada en los valores centrales y disminuyendo hacia los extremos. En la última gráfica, la correspondiente a 50 loci se aprecia otro fenómeno interesante: si bien los límites son ahora de 0 y de 100, ya no apreciamos que ninguna barra se encuentra cerca de esos valores, al menos con 100.000 individuos simulados. De hecho, es fácil calcular la probabilidad de cualquiera de los dos eventos extremos en las colas. Por ejemplo, para que el valor de un individuo sea 100 es necesario que los 50 loci sean homocigotos para el alelo que aporta una unidad (y, por lo tanto, aportan 2 unidades por locus). Asumiendo que los loci se encuentran en equilibrio de Hardy-Weinberg con una probabilidad $p = q = \frac{1}{2}$, como hemos usado en la figura 8.1, la probabilidad en un locus de que sea homocigoto para el alelo que aporta una unidad es $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$. Asumiendo que todos los loci son independientes, la probabilidad de que en todos se dé esta situación es el producto de cada una de las probabilidades, o lo que es lo mismo $\left(\frac{1}{4}\right)^{50} \approx 7{,}89 \times 10^{-31} \approx 0$. En términos prácticos, ya que apenas hemos muestreado solo 100 mil individuos, resulta claro que es muy remota la posibilidad de observar un individuo con estas características; siendo más exactos, esperamos ver $\approx 7{,}89 \times 10^{-31} \times 100.000 \approx 7{,}89 \times 10^{-26}$ individuos. La otra cosa interesante acerca de esta distribución es que a medida que aumentamos el número de loci en consideración, también aumenta el número de categorías de valores (barras en el gráfico) y rápidamente el gráfico empieza a insinuar que una línea continua podría ser trazada sin gran pérdida de precisión.

Más aún, si recuerdas tus cursos de estadística notarás que la forma en la gráfica de 50 loci se asemeja bastante al de una distribución normal. Esto no nos debería sorprender para nada si recordamos el enunciado del **teorema del límite central**: la suma de variables aleatorias independientes e idénticamente distribuidas tiende a la distribución normal a medida que el número de variables sumadas se incrementa. Es importante tomarse el tiempo necesario para refrescar este concepto, ya que resultará central para este y el resto de los capítulos que conforman el libro.

El fenómeno que describimos anteriormente es mucho más relevante de lo que imaginamos. Durante el comienzo del siglo XX se desarrolló un intenso debate entre la escuela mendeliana (que enfocaba el problema de la herencia en los términos en los que Mendel había definido el problema de los guisantes) y la escuela biométrica (mucho más enfocada en lo continuo de la variabilidad), ambos manteniendo posiciones irreconciliables sobre los mecanismos de la herencia, que en ese momento eran desconocidos. Sin embargo, la posibilidad de aproximarse a distribuciones continuas simplemente a partir de la suma de variables discretas genera un puente conceptual entre las dos visiones. Más aún, si a esta variabilidad casi-continua le solapamos el efecto de la variabilidad ambiental, dada la «discretitud» (permítasenos acuñar este término) de las escalas «reglas» con las que usualmente medimos,

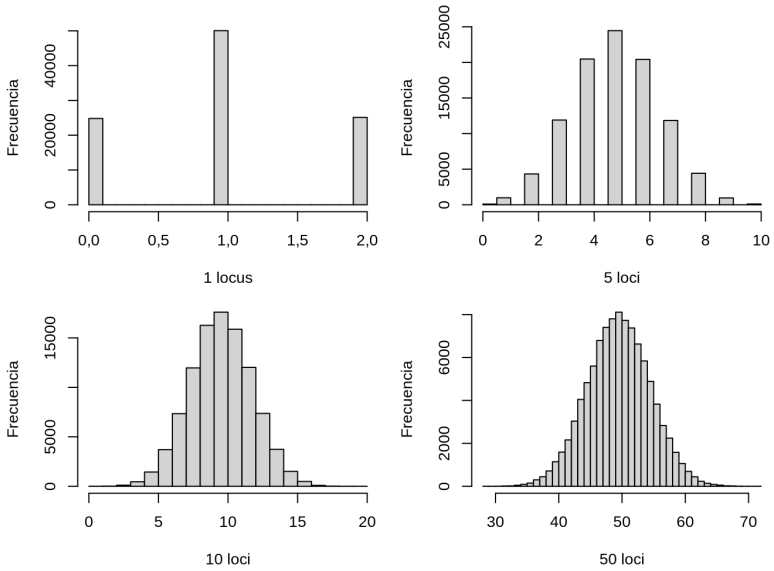


Figura 8.1. Distribución de valores al sumar los efectos de diferente número de loci, De arriba izquierda a abajo derecha, 1 locus, 5 loci, 10 loci y 50 loci. En todos los casos se muestrearon 100.000 individuos en los que cada uno de los loci fueron muestreados independientemente, bajo equilibrio de Hardy-Weinberg y frecuencia de $p = q = \frac{1}{2}$ y con valores 0, 1, 2 para los tres genotipos.

entonces estamos frente a una situación donde resulta mucho más difícil distinguir si se trata de una variable discreta o continua. Aportes sustantivos en este sentido fueron los realizados por George Udny Yule[3] (1902) y Wilhelm Johannsen (1903), que fueron fundamentales para los posteriores trabajos que llevaron a la ley de Hardy-Weinberg (Castle 1903; Hardy 1908; Weinberg 1908) y que contribuyeron a reconciliar ambas posiciones en lo que luego terminaría siendo la síntesis neodarwinista.

Muchas de las características en genética cuantitativa pueden describirse por alguno de los dos tipos de variables que vimos antes, **discretas** y **continuas**. Algunas otras características, sin embargo, tienen una distribución donde es posible tomar uno solo de unos pocos valores posibles y que los mismos tienen un ordenamiento natural, o que se trata de características del tipo presencia-ausencia. Por ejemplo, en el caso de enfermedades o patologías con cierta base genética, ciertas veces la causa es multifactorial y su base genética es poligénica, es decir, influenciada por muchos genes. En esos casos, también es posible pensar que existe una característica subyacente no observable, que es continua y que, al sobrepasar un determinado valor de umbral, entonces la enfermedad se manifiesta. Esta forma de pensar las características y modelarlas, los modelos **umbral** (Wright 1934a), puede tener una base molecular y funcional real o puede ser sencillamente un truco adecuado para modelarla, pero el hecho es que suele ser una buena alternativa en mejoramiento genético para varias características (Gianola 1982).

[3]George Udny Yule (18 de febrero de 1871 – 26 junio de 1951) fue un estadístico escocés, alumno de Karl Pearson y conocido por la distribución que lleva su nombre, además de por sus trabajos fundacionales en bioestadística.

PARA RECORDAR

- El **teorema central del límite** o **teorema del límite central** indica que, en condiciones muy generales, en la suma de variables aleatorias independientes, con media y varianza finitas, la función de distribución «se aproxima bien» a una distribución normal.
- El **modelo infinitesimal** supone que las características continuas pueden consistir en la suma infinitesimal de los aportes de un número infinito de loci.
- Cuando nos enfrentamos a una variación de tipo **categórica** (es decir con un número finito y contable de las mismas), donde a su vez asumimos que cada combinación de dos alelos producirá una categoría distinta, con k alelos podemos formar $n = k(k+1)/2$ genotipos diploides diferentes (incluyendo los homocigotos).
- Los **modelos umbral** se construyen a partir de la suposición de que existe una característica subyacente no observable, que es continua y que, al sobrepasar un determinado valor, cierta característica se manifiesta.

8.2 El modelo genético básico

En general, para la mayoría de las características que poseen una componente genética en su expresión estamos limitados en un principio a observar su manifestación **fenotípica**, sin poder acceder a identificar directamente el genotipo de cada individuo. Esto es así, por ejemplo, en las características productivas de interés económico en ganado como el peso en distintos momentos de su vida, la producción de leche, grasa y proteína, pero también en humanos (la altura, el peso, la presión sanguínea) o en plantas. Claramente, lo que observamos puede tener un componente genético, pero también existe en esas observaciones un «ruido» superpuesto. En general, este ruido es imposible de eliminar en este nivel de análisis, y «contamina» nuestras estimaciones sobre el efecto de genes y alelos. Es decir, aún si una característica cuantitativa se encontrara 100 % determinada por los efectos genéticos, las mediciones que realizamos de la misma tienen, de forma ineluctable, un error de medición (como toda medida). Esto nos lleva a considerar que nuestros modelos de lo que observamos deberían incluir, al menos, estos dos factores: el genotipo y los errores de medición. Más en general, el ambiente tiene una influencia importante o determinante en la mayor parte de las características cuantitativas de interés. Si bien todos reconocemos que la altura de los padres es de alguna manera heredable, también todos sabemos que la altura de las poblaciones humanas ha ido aumentando a medida que el ambiente (alimentación, deportes, hábitos, salud) han ido cambiando, en particular durante el siglo XX. Más directamente vinculado a la producción, todos sabemos que la producción de leche por vaca ha aumentado dramáticamente durante la segunda mitad del siglo XX, en gran parte por el mejoramiento genético, pero sin embargo, aun la misma vaca incorrectamente alimentada apenas producirá una fracción de todo su potencial genético. De hecho, como nosotros en este curso estamos mayormente interesados en la genética, podemos considerar que los errores de medición son a nuestros efectos parte del ambiente. Eso nos lleva a una primera aproximación, algo rústica, para el modelado de los fenotipos. Consideraremos el **fenotipo de cada individuo**, al que identificamos con la letra P (por su nombre en inglés, *phenotype* y para diferenciarlo claramente de la F que hemos usado para el índice de fijación).

El fenotipo es en nuestro modelo la **suma** de los **efectos genéticos** (que notamos con G) y los **efectos ambientales** (que notamos con E, del inglés *environment*). Tenemos entonces,

$$P = G + E \tag{8.1}$$

A decir verdad, para que quedase más claro el significado de esta ecuación deberíamos escribirla como $P_i = G_i + E_i$, a efectos de explicitar que se trata de una ecuación correspondiente al individuo i. Sin embargo, como usaremos subíndices para indicar otras cosas, omitiremos el subíndice referente al individuo. Pese a eso, debe quedar claro que las ecuaciones corresponden a un individuo.

El genotipo de un individuo es el arreglo particular de alelos en los distintos loci, ya sea en todos ellos o, al menos, en los relevantes para el estudio realizado. Dependiendo del nivel de sofisticación de nuestro tratamiento necesitaremos conocer también las fases de los mismos, es decir qué alelos de cada loci se encuentran localizados con qué alelos de los otros loci en el mismo cromosoma (heredado por vía materna o paterna). Esto es relevante para poder identificar los **haplotipos** presentes, aunque generalmente nosotros no vamos a considerar este nivel de análisis. En la figura 8.2 podemos ver una representación esquemática de una característica que está gobernada (al menos) por dos loci, con al menos dos alelos en cada uno de ellos.

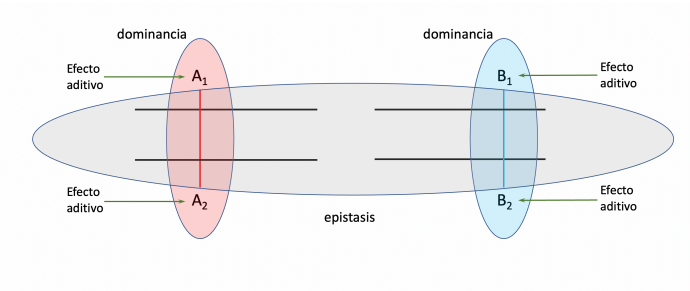


Figura 8.2. Representación esquemática de los efectos genéticos en el modelo genético básico. En dos o más loci con dos o más alelos en cada uno, tenemos los efectos aditivos de cada alelo por separado, que simplemente suman el efecto de cada uno presente (sin importar su combinación), los efectos de dominancia que son los desvíos **intralocus** de la aditividad (marcados acá con los óvalos de color rosado y celeste) y finalmente los efectos de epistasis, que son las desviaciones de interacción correspondientes a las diferentes combinaciones **interloci** (el óvalo gris).

Por un lado tenemos el efecto que cada alelo confiere a la característica, independiente de la combinación en que se encuentre, tanto dentro de su locus como considerando lo que hay en otros loci. Llamamos a esto **efectos aditivos**, ya que como no dependen de combinaciones simplemente podemos sumar el aporte de cada uno de ellos para tener la contribución de todos al valor genotípico y fenotípico. Por ejemplo, para conocer el valor aditivo A del individuo que aparece representado en la figura alcanza con sumar los efectos de los 2 alelos presentes en cada loci, es decir $A = A_1 + A_2 + B_1 + B_2$. Notemos que los subíndices se refieren a los alelos presentes, y pueden ir en cada locus desde el 1 al número de alelos presentes en el mismo. Siempre sumaremos los efectos de dos alelos por locus, porque siempre hay dos alelos en un locus diploide (pueden ser el mismo, y en ese caso lo

sumamos dos veces) y, por lo tanto, si definimos $i = 1 : 2$ el recorrido por los alelos de cada locus y n el número de loci, entonces tenemos

$$A = \sum_{i=1,j=1}^{i=2,j=n} A_{i,j} \tag{8.2}$$

Pero, como vimos previamente en el capítulo «Selección natural», el **fitness** de un individuo heterocigoto no necesariamente tiene que ser el punto medio entre los genotipos homocigotas, por lo que no alcanza con conocer estos dos valores, en general, para poder representar los 3 genotipos. Una forma matemática de acomodar esto es expresando los desvíos de cada genotipo respecto a su **valor aditivo** (es decir, a la suma de los dos alelos que lo componen) y le llamamos a esto **efecto de dominancia**, o **desvío de dominancia**. En el caso de que los valores de todos los genotipos (3 en el caso de dos alelos, $k(k+1)/2$ en general) sean exactamente iguales a la suma de los alelos que lo conforman, entonces el efecto de dominancia será cero y decimos que el locus tiene un **comportamiento aditivo**. Muchas son las posibles causas bioquímicas/moleculares que pueden subyacer al efecto de dominancia. Un ejemplo ilustrativo y fácil de comprender es la no-linealidad de la actividad enzimática respecto al número de copias del alelo; dicho de otra forma, dos copias del mismo alelo «ventajoso» no generan el doble de actividad enzimática (por ejemplo, porque la misma muestra saturación), por lo que no existe una relación lineal entre número de copias de ese alelo y el valor genotípico/fenotípico. Puesto en términos matemáticos, como la dominancia es un efecto **intralocus** que depende de la combinación específica de alelos en cada uno de los loci, si llamamos l a la combinación específica que tiene nuestro individuo en cada locus, podemos escribir el efecto total de dominancia como

$$D = \sum_{j=1}^{n} D_{l,j} \tag{8.3}$$

Por otro lado, como se ve en la figura 8.2, también pueden existir desvíos de los valores de determinadas combinaciones respecto a los **efectos aditivos** y también a los efectos de los **desvíos de dominancia**, ya que ninguno de ellos ha considerado las combinaciones de alelos **entre** loci. A estos desvíos le llamamos **epistáticos**, y se corresponden con interacciones alélicas intergénicas. Las causas más usuales a nivel molecular para estos efectos son la **pleiotropía** y el **ligamiento**. Estas interacciones pueden ser de muchos tipos, por ejemplo, entre dos loci tenemos interacciones de efectos aditivos en un locus con los efectos aditivos del otro, I_{AA}, de los aditivos con los de dominancia, I_{AD} y de los de dominancia con los de dominancia, I_{DD}. A medida que aumentamos el número de loci crecen las combinaciones y a fines prácticos agrupamos todos en el término I, es decir

$$I = I_{AA} + I_{AD} + I_{DD} + I_{AAA} + I_{AAD} + ... \tag{8.4}$$

Por lo general solo se tienen en cuenta las interacciones entre dos componentes ya que las superiores suelen tener poco efecto en el valor global y en términos prácticos implica estimar una gran cantidad de componentes, lo que los datos no suelen garantizar.

Veamos ahora un ejemplo de dos loci que se comportan en forma aditiva. Sin pérdida de generalidad, supongamos que ambos loci se encuentran en equilibrio Hardy-Weinberg, con frecuencias iguales de los dos alelos en cada locus, lo que nos lleva a que las frecuencias sean $f(A_1A_1) = 1/4$, $f(A_1A_2) = 1/2$ y $f(A_2A_2) = 1/4$ (lo mismo para el locus B). En el locus A tenemos dos alelos, con valores $A_1 = 2$, $A_2 = -2$, mientras que, para el segundo locus, sus alelos aportan $B_1 = 1$ y $B_2 = -1$. En el cuadro siguiente tenemos todas las 9 combinaciones posibles de los genotipos en los dos loci y los valores correspondientes, que surgen de sumar los efectos de los alelos presentes en cada uno:

		A_1A_1	A_1A_2	A_2A_2	**Promedio**
	Frecuencia	$1/4$	$1/2$	$1/4$	
B_1B_1	$1/4$	6	2	-2	2
B_1B_2	$1/2$	4	0	-4	0
B_2B_2	$1/4$	2	-2	-6	-2
Promedio		4	0	-4	

Como el efecto de ambos genes es aditivo, tanto a nivel de filas como de columnas, el pasaje de una celda a la consiguiente es constante (diferentes constantes en filas que en columnas) y el valor del cambio es igual a la diferencia de valor del alelo que cambia. Por ejemplo, si miramos la primera fila, el pasaje de A_1A_1 a A_1A_2 significa perder una copia de A_1 con valor 2 y ganar una de A_2 con valor -2, es decir, el cambio es de $-2-2 = -4$, que es lo que observamos, ya que pasamos de un valor de 6 a uno de 2 y, por lo tanto, $\Delta = 2-6 = -4$. El mismo valor vamos a obtener en cualquier movida de una celda a su contigua a la derecha, sin importar la fila, al tiempo que si movemos de derecha a izquierda su valor será el opuesto (4) en este caso. Para las columnas ocurre lo mismo, pero ahora el valor de $\Delta = -2$ si nos desplazamos a la contigua inferior, mientras que $\Delta = 2$ si nos desplazamos a la contigua superior.

Veamos ahora qué ocurre cuando aparece el efecto de dominancia en uno de los locus. En el cuadro siguiente los efectos aditivos son idénticos que en el cuadro anterior, pero ahora aparece un efecto de dominancia en el locus A. Mientras que, si nos movemos verticalmente, es decir cambiamos en el número de copias de los alelos B_1 y B_2, las distancias entre celdas contiguas se mantienen, es decir hay aditividad, si nos movemos en sentido horizontal la situación cambia. No solo cambia respecto al cuadro anterior, sino que también cambia dependiendo de en qué columna me pare. Por ejemplo, la distancia entre cualquier celda de la primera columna (que corresponde al genotipo A_1A_1) a la contigua de la segunda columna (que corresponde al genotipo A_1A_2) es de $\Delta = -3$ unidades, mientras que de la segunda a la tercera columna (que corresponde al genotipo A_2A_2) el cambio es de $\Delta = -5$ unidades. Claramente el promedio no coincide en ningún caso con el valor del genotipo de B correspondiente a esa fila, como sí ocurría en el cuadro anterior (aditividad en los dos loci, con las frecuencias particulares elegidas).

		A_1A_1	A_1A_2	A_2A_2	**Promedio**
	Frecuencia	1/4	1/2	1/4	
B_1B_1	1/4	6	3	−2	2,50
B_1B_2	1/2	4	1	−4	0,50
B_2B_2	1/4	2	1	−6	−1,50
Promedio		4	0	−4	

Finalmente, en el cuadro siguiente le hemos agregado al cuadro de efectos aditivos un efecto de interacción **aditivo × aditivo**, sencillamente como el producto de los valores del efecto aditivo de los dos loci para cada par de genotipos (las funciones de interacción pueden ser infinitas, esta es una sola, elegida por sencilla para representar el fenómeno). El primer valor de cada celda corresponde al efecto aditivo y el segundo al **efecto de epistasis** del tipo I_{AA}. Además de haberse alterado los valores de casi todas las celdas, el cambio entre celdas contiguas experimenta ahora algunos valores leves y otros abruptos, en los dos sentidos. Por ejemplo, pasar de A_1A_1/B_1B_1 a A_1A2/B_1B_1 significa $\Delta = 14 - 2 = 12$ unidades, mientras que de A_1A_1/B_2B_2 a A_1A_2/B_2B_2 el cambio correspondiente es de solo $\Delta = -6 - (-2) = -4$ unidades. Notemos que aquellos genotipos cuyos efectos aditivos son del mismo signo tienen un incremento debido a la interacción, lo que lleva, por ejemplo, a que el A_2A_2/B_2B_2 deje de ser la peor combinación. Claramente, los efectos epistáticos pueden cambiar totalmente la configuración de los genotipos respecto a la aditividad o a la dominancia.

		A_1A_1	A_1A_2	A_2A_2	**Promedio**
	Frecuencia	1/4	1/2	1/4	
B_1B_1	1/4	$6+8=14$	$2+0=2$	$-2-8=-10$	2
B_1B_2	1/2	$4-0=4$	$0+0$	$-4-0=-4$	0
B_2B_2	1/4	$2-8=-6$	$-2+0=-2$	$-6+8=2$	−2
Promedio		4	0	−4	

Teniendo en cuenta todo lo anterior, podemos ahora escribir nuestro modelo genético de una forma algo más complicada, pero menos rústica, describiendo mejor los componentes del genotipo y sus diversas interacciones. Si notamos con A los **efectos aditivos**, con D los **desvíos de dominancia** y con I los **efectos epistáticos**, entonces podemos describir el fenotipo de cada individuo como

$$P = G + E = A + D + I + E \tag{8.5}$$

Claramente, esto asume que la relación entre lo genético y lo ambiental es aditiva, lo que no necesariamente es así y, por lo tanto, debemos entrar en el campo de las interacciones entre ambos componentes, como veremos en más detalle en la sección «Interacción genotipo-ambiente». Dicho de otra forma, en términos matemáticos, no alcanza con sumar los efectos de genotipo y ambiente para describir de la mejor forma posible el comportamiento diferencial de los genotipos en los ambientes y es necesario incorporar un parámetro adicional que describa esa diferencia de comportamientos en los distintos ambientes. Se trata de una interacción, y vamos a usar también la letra I como en el caso de epistasis, pero ahora con

un subíndice que refiera específicamente a la clase de interacción genotipo-ambiente, por lo que el modelo de la ecuación (8.5) quedará ahora

$$P = G + E + I_{G\times E} = A + D + I + E + I_{G\times E} \tag{8.6}$$

que será una de las maneras más explícitas de expresar el fenotipo de un individuo que usaremos en el curso.

PARA RECORDAR

- Definiremos al **fenotipo** (P) de un individuo como la suma de los efectos genéticos (G), los ambientales (E) y la interacción entre los mismos ($I_{G\times E}$).
- El **genotipo** del individuo, definido como el arreglo particular de alelos en los distintos loci, consistirá de la suma de los efectos aditivos (A), de dominancia (D) y epistáticos (I).
- Los **efectos aditivos** corresponden al efecto que cada alelo confiere a la característica, independiente de la combinación en que se encuentre, tanto dentro de su locus como considerando lo que hay en otros loci. Tomando como referencia un locus diploide, entonces

$$A = \sum_{i=1,j=1}^{i=2,j=n} A_{i,j} \tag{8.7}$$

- El **efecto o desvío de dominancia** expresa el desvío de cada fenotipo respecto a su valor aditivo. La dominancia es un efecto intralocus que depende de la combinación específica de alelos en cada uno de los loci. Si llamamos l a la combinación específica que tiene nuestro individuo en cada locus, podemos escribir el efecto total de dominancia como

$$D = \sum_{j=1}^{n} D_{l,j} \tag{8.8}$$

- Los **efectos o desvíos epistáticos** se corresponden con interacciones alélicas intergénicas, es decir, entre distintos loci. Las causas más usuales a nivel molecular para estos efectos son la pleiotropía y el ligamiento. Este efecto corresponde a la suma de todas las interacciones entre los loci y podemos resumirlo mediante el término I

$$I = I_{AA} + I_{AD} + I_{DD} + I_{AAA} + I_{AAD} + \ldots \tag{8.9}$$

8.3 Modelo genético básico: un locus con dos alelos

Como hemos mencionado antes, en general, la mayor parte de las características de interés para la genética cuantitativa son de herencia poligénica, es decir, controlada por el efecto de alelos en varios (muchos) genes. Esto, sumado a la imposibilidad de conocer para cada uno de esos genes sus estados alélicos, además de los efectos de cada alelo, ha llevado a la genética cuantitativa del siglo XX a desarrollarse basada en una aproximación estadística simple y potente, llamada **el modelo infinitesimal**. La misma traslada el problema de conocer el efecto de cada gen y de cada alelo a un problema de distribuciones estadísticas, que asumiendo un número de loci razonablemente grande y una pequeña contribución de cada

loci, nos permite invocar el teorema del límite central, como ya vimos antes también. Una porción importante de esta **Parte III** del libro estará basada en dicho modelo, por lo que será de gran relevancia para nosotros.

Por el contrario, los genes de **efecto mayor**, aquellos genes en que algunas o todas las variantes (alelos) marcan una diferencia importante en el fenotipo, son relativamente escasos en la genética cuantitativa, aunque a veces determinantes en diferencias importante en producción, o entre razas de la misma especie. En una revisión reciente, D. Wright (2015) identifica varios de ellos en especies animales domésticas. Por ejemplo, en el perro el largo de patas estaría gobernado en gran medida por el gen $FGF4$, mientras que el tamaño corporal estaría determinado por el gen $IGF1$. En ganado vacuno el crecimiento muscular estaría determinado en gran medida por el gen $MSTN$, el (mal) olor en la leche por el gen $FMO3$, mientras que la producción de leche estaría influenciada por el gen $DGAT1$ y la composición influenciada por las variantes en el gen $ABCG1$. En ovinos, la tasa de ovulación incrementada y la tasa de melliceros estaría influenciada por el gen $BMP15$, mientras que el conocido gen del efecto *Booroola* estaría determinado por las variantes en el gen $BMPR1B$. En caballo, mientras tanto, el tipo de paso estaría determinado por el gen $DMRT3$. Finalmente, en cerdos, la hipertermia maligna viene dada por variantes en el gen $RYR1$, el contenido de glicógeno en músculo por el gen $PRKAG3$, el tamaño de las orejas por el gen $PPARD$, la muscularidad y la grasa dorsal influenciados por el gen $IGF2$, mientras que la terneza de la carne estaría gobernada por el gen $CAST$.

Veamos, a partir de un ejemplo ficticio, como actúa un gen de efecto mayor y hasta dónde podemos llegar con el análisis del mismo. La figura 8.3 muestra el comportamiento **fenotípico** de un gen en un organismo diploide, con dos alelos, A_1 y A_2, con frecuencias $p = 0{,}6$ y $q = 0{,}4$, respectivamente, en una característica como peso al año en una raza vacuna. Supongamos que se trata de una característica en que conocemos la base molecular de la misma y que tenemos un **test genético** para determinar a qué genotipo corresponde cada animal (por ejemplo, a partir de un **chip de genotipado** o a partir de un **ensayo de PCR**). El genotipo A_1A_1 produce animales con un peso de 180 kg, el genotipo A_2A_2 produce animales de 120 kg de peso, mientras que el heterocigoto produce animales de 160 kg de peso, cada uno de ellos marcado en la figura 8.3 como una línea rayada vertical del color correspondiente al genotipo. Sin embargo, como existe **variabilidad ambiental**, definida como **variabilidad no-genética**, los animales de cada genotipo se distribuirán alrededor del valor correspondiente en forma aleatoria. En este caso asumimos que la variabilidad ambiental no depende del genotipo, es decir que es igual para todos los animales, algo que discutiremos más adelante y, por lo tanto, como de acuerdo con la ecuación (8.1) el fenotipo es igual a $G + E$, si G es un valor fijo para cada genotipo (G_{11}, G_{12} y G_{22}), entonces los fenotipos correspondientes a cada genotipo seguirán una distribución aleatoria con media en su correspondiente valor genotípico y varianza igual a la varianza ambiental (que llamaremos σ^2_E). Es decir,

$$P_{11} = G_{11} + N(0,\sigma^2_E) = N(G_{11},\sigma^2_E) \tag{8.10}$$

$$P_{12} = G_{12} + N(0,\sigma^2_E) = N(G_{12},\sigma^2_E) \tag{8.11}$$

$$P_{22} = G_{22} + N(0,\sigma^2_E) = N(G_{22},\sigma^2_E) \tag{8.12}$$

Esto se aprecia claramente en la figura como las curvas **gaussianas** de diferentes colores, centradas en los correspondientes valores genotípicos.

En nuestro caso particular, hemos asumido además que la población se encuentra en equilibrio de Hardy-Weinberg, por lo que las frecuencias de cada genotipo serán

$$\text{fr}(A_1A_1) = p^2 = 0{,}6^2 = 0{,}36 \tag{8.13}$$

$$\text{fr}(A_1A_2) = 2pq = 2 \times 0{,}6 \times 0{,}4 = 0{,}48 \tag{8.14}$$

$$\text{fr}(A_2A_2) = q^2 = 0{,}4^2 = 0{,}16 \tag{8.15}$$

lo que se aprecia claramente por el área de cada curva en la figura. Si juntamos las tres curvas, es decir unimos las subpoblaciones con distinto genotipo, entonces tenemos toda la población, representada por la curva negra. La media de la población general se encuentra representada por una línea continua de color negro, a la derecha de la línea rayada violeta que corresponde al genotipo heterocigoto. Finalmente, esta figura también muestra un aspecto interesante, que suele confundir a muchos cuando trabajamos con modelos lineales, por ejemplo. Pese a que las distribuciones de valores de los individuos pertenecientes a cada genotipo es normal y con la misma varianza, la distribución resultante de combinar los individuos de todos los genotipos **no es normal**. Esto no suele representar ningún problema cuando el modelo es infinitesimal porque no existe un número pequeño de clases, como en nuestro caso de la figura (8.1) y como cada clase está formada por diferentes combinaciones (suma) de contribuciones infinitesimales, entonces la distribución es aproximadamente normal en esos casos. En estadística, el supuesto de normalidad en modelos lineales (en ANOVA, por ejemplo) se refiere a la distribución de los residuos (es decir, luego de descontados los efectos del modelo), no de los datos originales.

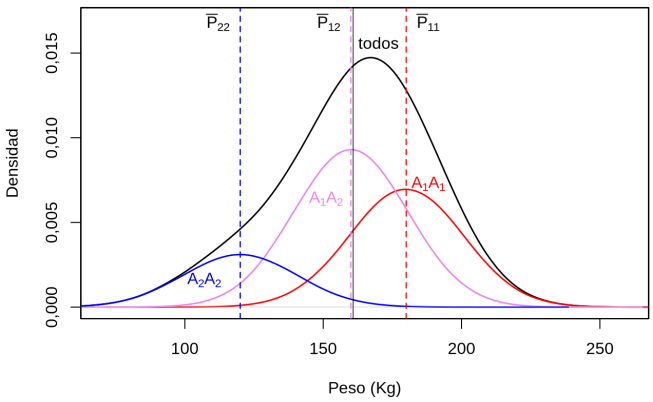


Figura 8.3. Distribución de pesos en una población compuesta de tres genotipos para un gen de **efecto mayor**, con dos alelos. La frecuencia del alelo A_1 es $p = 0{,}6$ y la población se encuentra en equilibrio Hardy-Weinberg. Los pesos promedio de los tres genotipos son $\bar{P}_{22} = 120$ kg, $\bar{P}_{12} = 160$ kg y $\bar{P}_{11} = 180$ kg, mientras que el efecto aleatorio del ambiente se muestrea de una distribución $N(\mu = 0{,}\sigma^2 = 400\ \text{kg}^2)$. Casi pegado a la línea rayada violeta, puede verse en una línea vertical continua de color negro la media de la población general, es decir la media **fenotípica** de la población.

En la figura 8.3 hemos colocado una línea vertical en el valor correspondiente a cada uno de los genotipos, pero hasta ahora no hemos mencionado de dónde salieron esos valores.

Es decir, el **test genético** lo único que es capaz de decirme son los genotipos de cada animal al que le realizamos el ensayo, pero no puede decirme cuánto vale ese genotipo. ¿De dónde sale entonces esa información? Si consideramos que los valores de los genotipos son independientes de otro tipo de información genómica y que la distribución de valores **fenotípicos** alrededor de cada valor **genotípico** es la misma, una distribución normal producto de la **variabilidad ambiental**, entonces las **medias de los fenotipos de cada clase** se corresponderá con el consiguiente **valor genotípico**. Es decir, si se cumplen nuestros supuestos, entonces, usando los resultados de la ecuación (8.12), tenemos que

$$\bar{P}_{11} = \mathbb{E}(P_{11}) = \mathbb{E}[G_{11} + N(0,\sigma^2_E)] = \mathbb{E}[N(G_{11},\sigma^2_E)] = G_{11} \tag{8.16}$$

$$\bar{P}_{12} = \mathbb{E}(P_{12}) = \mathbb{E}[G_{12} + N(0,\sigma^2_E)] = \mathbb{E}[N(G_{12},\sigma^2_E)] = G_{12} \tag{8.17}$$

$$\bar{P}_{22} = \mathbb{E}(P_{22}) = \mathbb{E}[G_{22} + N(0,\sigma^2_E)] = \mathbb{E}[N(G_{22},\sigma^2_E)] = G_{22} \tag{8.18}$$

Dicho de otra forma, si se conocen los pesos de cada animal y su genotipo, es posible separar a los animales por genotipo y calcular el promedio depara cada grupo, obteniendo entonces así el valor esperado para ese genotipo, que se asumimos se corresponde sin error su «verdadero» valor.[4] Pasamos entonces, ahora, de manejarnos de una distribución de **valores fenotípicos** a un conjunto de 3 valores genotípicos. Antes de comenzar a trabajar con estos valores genotípicos, veamos si podemos reducir el número de parámetros de 3 a 2, de forma tal de simplificar nuestras ecuaciones futuras. De todas las transformaciones imaginables, posiblemente la más utilizada en los textos de genética cuantitativa es la que aparece representada en la figura 8.4.

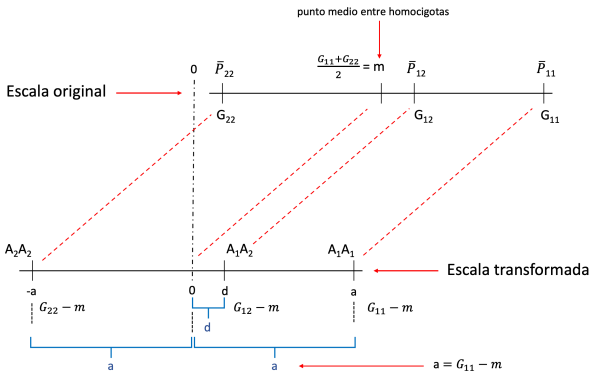


Figura 8.4. Transformación de escalas entre las mediciones originales, en la **escala original** (arriba) de la característica y la correspondiente transformación lineal para ubicar los genotipos extremos simétricamente al cero, en la **escala transformada** (abajo). El punto medio entre los homocigotos en la escala original es $m = \frac{G_{11}+G_{22}}{2}$, que si se lo restamos a cada uno de los genotipos originales nos da la correspondiente ubicación en la escala transformada. El cambio de notación de los genotipos en la escala superior a la inferior, con los alelos específicos es para resaltar el pasaje al modelo conceptual.

En la parte superior de la figura vemos representados los valores de los tres genotipos (G_{11}, G_{12}, G_{22}), o lo que es lo mismo las medias de las clases fenotípicas ($\bar{P}_{11}$, $\bar{P}_{12}$, $\bar{P}_{22}$) en la

[4] Obviamente que toda estimación tiene error, pero el concepto de «verdadero» es parte de la base filosófica del problema, ya que para la escuela estadística bayesiana lo único real y «verdadero» son los datos, no el valor de un parámetro **ficticio**, que depende del modelo que estemos usando.

escala original, es decir, la escala en la que se tomó la medida. En nuestro caso, con los valores de la figura 8.3, $G_{11} = 180$ kg, $G_{12} = 160$ kg y $G_{22} = 120$ kg. También sobre la misma escala, podemos ahora identificar el **punto medio entre los homocigotas**, que llamaremos m y cuyo valor está dado por

$$m = \frac{G_{11} + G_{22}}{2} \tag{8.19}$$

Es importante notar que m, el **punto medio entre los homocigotas, no es** en general igual a la media de la población, que es un concepto totalmente diferente. Mientras que m solo depende de los valores de G_{11} y G_{22} (en forma aritmética es el promedio de estos dos valores, mientras que geométricamente es el punto medio del segmento que los une), la media de la población $\mathbb{E}(P) = \bar{P}$ depende de los 3 valores genotípicos y de sus frecuencias. Si ahora, a cada uno de los valores genotípicos le restamos el valor del punto medio, tenemos ahora que en la escala transformada que aparece en la parte inferior de la figura, los valores correspondientes a cada uno de los genotipos homociogotas estarán dados por

$$A_1A_1 = G_{11} - m = G_{11} - \frac{G_{11} + G_{22}}{2} = \frac{2G_{11} - G_{11} - G_{22}}{2} = \frac{G_{11} - G_{22}}{2} \tag{8.20}$$

$$A_2A_2 = G_{22} - m = G_{22} - \frac{G_{11} + G_{22}}{2} = \frac{2G_{22} - G_{11} - G_{22}}{2} = \frac{G_{22} - G_{11}}{2} \tag{8.21}$$

Si ahora definimos $a = \frac{G_{11} - G_{22}}{2}$, usando los resultados de la ecuación anterior, el valor genotípico en la nueva escala para los homocigotas será

$$\begin{aligned} A_1A_1 &= \frac{G_{11} - G_{22}}{2} = a \\ A_2A_2 &= \frac{G_{22} - G_{11}}{2} = -a \end{aligned} \tag{8.22}$$

El último resultado sale de multiplicar a por -1, es decir $-1 \times a = -a = -1 \times \frac{G_{11} - G_{22}}{2} = \frac{G_{22} - G_{11}}{2}$. Como para definir los valores de dos genotipos (G_{11} y G_{22}) usamos un solo parámetro (a, ya que en el caso de G_{22} es solamente su opuesto), entonces hemos conseguido la reducción del número de parámetros que deseábamos. Si ahora, además, le llamamos d a la diferencia $G_{12} - m$, tenemos entonces finalmente que los valores de los tres genotipos estarán dados por

$$A_1A_1 = \frac{G_{11} - G_{22}}{2} = a \tag{8.23}$$

$$A_1A_2 = \frac{2G_{12} - G_{11} - G_{22}}{2} = d \tag{8.24}$$

$$A_2A_2 = \frac{G_{22} - G_{11}}{2} = -a \tag{8.25}$$

como aparece representado en la escala transformada (inferior) de la figura 8.4. De más está decir que se trata de una transformación lineal que conserva las distancias originales entre puntos y que la posibilidad de reducir un parámetro es a costa de expresar los valores como desvío de un punto medio (m), con ningún otro significado importante que permitirnos una manipulación algebraica más sencilla.

Ejemplo 8.1

En un estudio sobre el efecto de los genes de la miostatina y la μ-calpaína sobre características de la carcasa, Bennett y colaboradores (Bennett et al. 2019) eligen el marcador (SNP) **rs110065568** para el gen de la **miostatina** (en el cromosoma **BTA2**), que se trata de una sustitución de Fenilalanina (**F**) por Leucina (**L**) en la posición del aminoácido **94** (**F94L**). El gen de la miostatina es el responsable principal del fenotipo **doble musculado** común a varias razas de ganado, como el toro de la raza **belgian Blue** que aparece en la figura 8.5 y aún a diferentes especies de interés doméstico.

Figura 8.5. Toro de la raza **belgian blue** mostrando su impresionante musculatura (de Wikipedia, autor Mastiff, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, vía Wikimedia Commons).

En el cuadro siguiente se aprecian los promedios de mediciones para dos características que obtuvieron los autores del trabajo, de acuerdo con el genotipo de los animales (G_{FF} homocigotos para Fenilalanina, G_{LL} para Leucina y G_{FL} heterocigotos).

Característica	G_{FF}	G_{FL}	G_{LL}
Peso al nacimiento (kg)	40,7	41,4	43,3
Área del ojo del bife (cm^2)	89	93	103

Asumiendo que estos valores son los promedios de un gran número de animales y que por lo tanto el promedio de los efectos ambientales es cero, transformar los valores genotípicos de las características a la escala en la que el cero es el punto medio entre los homocigotos.

Para la primera característica, peso al nacimiento, el punto medio entre los homocigotas los calculamos como

$$m = \frac{G_{LL} + G_{FF}}{2} = \frac{43{,}3 + 40{,}7}{2} = 42 \text{ kg} \tag{8.26}$$

Por lo tanto, los valores de los genotipos expresados en esta escala transformada serían

$$A_{LL} = G_{LL} - m = (43{,}3 - 42) \text{ kg} = 1{,}3 \text{ kg} = a \tag{8.27}$$

$$A_{FL} = G_{FL} - m = (41{,}4 - 42) \text{ kg} = -0{,}6 \text{ kg} = d \tag{8.28}$$

$$A_{FF} = G_{FF} - m = (40{,}7 - 42) \text{ kg} = -1{,}3 \text{ kg} = -a \tag{8.29}$$

Con la misma idea, para la segunda característica tenemos

$$m = \frac{G_{LL} + G_{FF}}{2} = \frac{103 + 89}{2} = 96 \text{ cm}^2 \tag{8.30}$$

y los genotipos transformados

$$A_{LL} = G_{LL} - m = (103 - 96) \text{ cm}^2 = 7 \text{ cm}^2 = a \tag{8.31}$$

$$A_{FL} = G_{FL} - m = (93 - 96) \text{ cm}^2 = -3 \text{ cm}^2 = d \tag{8.32}$$

$$A_{FF} = G_{FF} - m = (89 - 96) \text{ cm}^2 = -7 \text{ cm}^2 = -a \tag{8.33}$$

Media de la población

A partir del cambio de escala que realizamos previamente, los valores genotípicos quedarán expresados como desvíos del punto medio (m) entre los homocigotas. En general, dicho punto no coincide con la media de la población ya que el mismo se desplazará de acuerdo con la importancia de d en relación con a y de las frecuencias de los alelos.[5] Para calcular la media procedemos como de costumbre, multiplicando el valor de cada categoría por su frecuencia relativa. Asumiendo que los genotipos están en equilibrio de Hardy-Weinberg, las frecuencias de los 3 genotipos (A_1A_1, A_1A_2 y A_2A_2) serán p^2, $2pq$ y q^2, mientras que los valores genotípicos correspondientes (en la escala transformada) serán a, d y $-a$. Poniendo todo junto en un cuadro, lo anterior queda resumido de la siguiente manera:

Genotipo	Frecuencia	Valor genotípico	Producto
A_1A_1	p^2	a	p^2a
A_1A_2	$2pq$	d	$2pqd$
A_2A_2	q^2	$-a$	$-q^2a$
Suma			$M = a(p - q) + 2pqd$

[5] Recordemos que la media no equivale al promedio de valores presentes en la población, ya que para obtener el valor medio debemos ponderar por la frecuencia de cada uno de los valores que observamos. La persona interesada en ahondar en este tema puede referirse al apéndice matemático de este libro.

El resultado final, $M = a(p-q) + 2pqd$, deriva de sumar todos los términos de la última columna y simplificar, es decir

$$M = p^2a + 2pqd - q^2a = a(p^2 - q^2) + 2pqd = a(p+q)(p-q) + 2pqd \ \therefore \qquad (8.34)$$

$$M = a(p-q) + 2pqd \qquad (8.35)$$

En el caso particular de un locus con dos alelos, la restricción de que $p+q = 1 \Leftrightarrow q = 1-p$ nos permite ir un poco más lejos y expresar la media solo en función de a,d y p, de la siguiente manera

$$M = a(p-q) + 2pqd = a[p-(1-p)] + 2dp(1-p) = a(2p-1) + 2d(p-p^2) \quad (8.36)$$

En la figura 8.6 podemos apreciar cómo varía la media genotípica de la población en función de la frecuencia de alelo A_1 (p) y del desvío del punto medio del heterocigota (d).

Para hacer el gráfico independiente del valor de a (porque en general no podemos interpretar gráficos en 4 dimensiones) fijamos este en 1 $a = 1$, y por lo tanto $-1 \leqslant d \leqslant 1$, asumiendo que no hay subdominancia ni sobredominancia. Claramente, a medida que aumenta el valor de p siempre aumenta la media de la población, lo que debería resultar obvio ya que la frecuencia del homocigota con valor positivo (a) es p^2, es decirlo cual implica que crece con p (y, por lo tanto, la frecuencia del otro homocigota —que tiene valor $-a$-, decrece con p). Por otro lado, también es posible observar que la media de la población decrece a medida que d va de $+a$ (1) a $-a$ (-1). Sin embargo, la interacción entre p y d no es lineal, ya que en la ecuación (8.36) hay un término que involucra su producto. De hecho, podemos entender la ecuación (8.36) como el aporte de dos componentes a la media, el correspondiente al aporte de a, que es ($a(p-q)$), y el aporte correspondiente al desvío del heterocigota (por dominancia), que es de $2d(p-p^2) = 2pqd$, que no es más que el desvío d multiplicado por la frecuencia de heterocigotas ($2d(p-p^2) = 2pqd$). Pero como hemos visto repetidas veces, $2pq$ es máximo cuando $p = q = \frac{1}{2}$, en cuyo caso $2pq = 2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$, y en ese caso el aporte será $2pqd = \frac{d}{2}$. A medida que las frecuencias de p o de q se acerquen a cero, entonces habrá cada vez menos heterocigotas, y por lo tanto $2pqd \to 0$. En los casos extremos, cuando $p = 1$, entonces $M = a(p-q)+2pqd = a(1-0)+2\times1\times0\times d = a$ y en el caso opuesto, cuando $q = 1$, entonces tenemos $M = a(p-q)+2pqd = a\times(0-1)+2\times0\times1\times d = -a$, lo que es obvio ya que en cada uno de estos casos solo hay un genotipo.

En general, si asumimos que no hay subdominancia ni sobredominancia (es decir, el valor genotípico del heterocigota se encuentra entre el de los dos homocigotas), la amplitud total de la variación genotípica es $a-(-a) = a+a = 2a$

En esta sección nos hemos manejado con un locus de un organismo diploide que cuenta solo con dos alelos. Sin embargo, si asumimos que no existe interacción entre loci, es decir, los efectos epistáticos no existen o son despreciables a los fines de nuestro interés (una simplificación), entonces es posible generalizar el resultado de la ecuación (8.36) a todos los loci involucrados en la característica de interés. El supuesto anterior implica que el efecto de todos los loci en conjunto es aditivo. Dicho de otra forma, asumimos que los efectos de cada locus se suman para obtener el efecto total. Si ahora consideramos $a = \sum_i a_i$, es decir, la

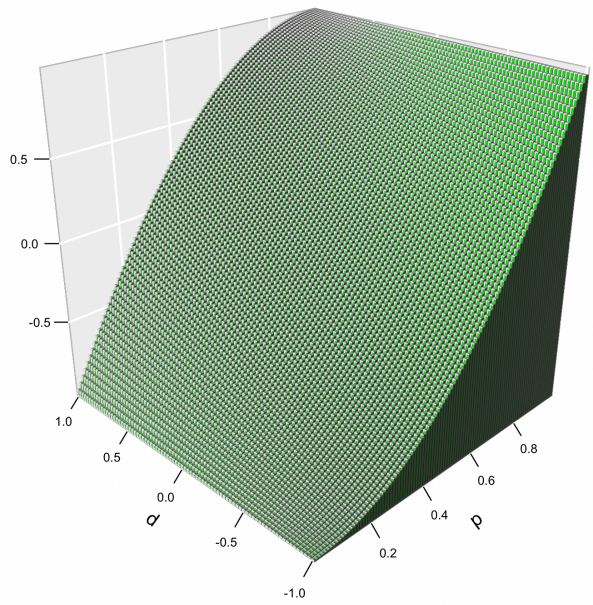


Figura 8.6. Media genotípica de la población como función de la frecuencia del alelo A_1 (p) y del valor de dominancia d. El valor de a fue fijado en 1 y, por lo tanto, $-1 \leqslant d \leqslant 1$, asumiendo que no hay subdominancia ni sobredominancia. Se observa claramente que mientras la media crece siempre con el valor de p (ya que p^2 es la frecuencia del homocigota con valor positivo, a), como d puede variar en un sentido u otro alrededor del cero, su valor aumentará o decrecerá la media respecto a la situación de ausencia de dominancia, cambiando también la forma (concavidad) de la curva.

diferencia entre los homocigotos fijados para todos los alelos positivos y el cero (en la escala transformada, con el cero el punto medio de cada uno de ellos), entonces la ecuación (8.36) se transforma en

$$M = \sum_i a_i(p_i - q_i) + 2\sum_i p_i q_i d_i \tag{8.37}$$

que es la generalización de la ecuación (8.36) asumiendo aditividad entre loci.

Ejemplo 8.2

En el mismo estudio de Bennett et al. (2019), los autores encuentran que en una de las razas evaluadas la frecuencia del alelo L (**F94LaL**) es de 20,7 %. Calcular las medias de la población para las dos características analizadas en el ejemplo 8.1.

Si $p = 0{,}207 \Rightarrow q = 1 - p = 1 - 0{,}207 = 0{,}793$, con los valores de la primera característica (peso al nacimiento) $a = 1{,}3$ y $d = -0{,}6$ y usando la ecuación (8.35), tenemos

$$M = a(p - q) + 2pqd = 1{,}3(0{,}207 - 0{,}793) + 2 \times 0{,}207 \times 0{,}793 \times (-0{,}6) = -0{,}9588 \tag{8.38}$$

respecto al punto medio entre los homocigotos, que era $m = 42$ kg, por lo que sumando ambos tenemos la media expresada en la escala original, es decir

$$42 + (-0{,}9588) = 42 - 0{,}9588 = 41{,}0412 \text{ kg} \tag{8.39}$$

Aplicando idéntico razonamiento para la segunda característica, pero ahora con los valores de $a = 7$ y $d = -3$, tenemos

$$M = a(p - q) + 2pqd = 7(0{,}207 - 0{,}793) + 2 \times 0{,}207 \times 0{,}793 \times (-3) = -5{,}0869 \text{ cm}^2 \tag{8.40}$$

respecto al punto medio entre los homocigotos ($m = 96$ cm^2), por lo que sumando ambos nos queda

$$96 + (-5{,}0869) = 96 - 5{,}0869 = 90{,}9131 \text{ cm}^2 \tag{8.41}$$

como el valor de la media para la segunda característica, expresada en la escala original.

PARA RECORDAR

- Los **genes de efecto mayor** son aquellos en que algunas o todas las variantes (alelos) marcan una diferencia importante en el fenotipo, son relativamente escasos en la genética cuantitativa.
- Si consideramos que los valores de los genotipos son independientes de otro tipo de información genómica y que la distribución de valores fenotípicos alrededor de cada valor genotípico es la misma (una distribución normal producto de la variabilidad ambiental), entonces las medias de los fenotipos de cada clase se corresponderá con el consiguiente valor genotípico.

- Podemos transformar de escala los valores genotípicos, ubicándolos en torno a cero (la media de los homocigotas en la escala original), para trabajar con un parámetro menos en nuestros cálculos.
- Los valores de los tres genotipos transformados serán

$$A_1A_1 = \frac{G_{11} - G_{22}}{2} = a \tag{8.42}$$

$$A_1A_2 = \frac{2G_{12} - G_{11} - G_{22}}{2} = d \tag{8.43}$$

$$A_2A_2 = \frac{G_{22} - G_{11}}{2} = -a \tag{8.44}$$

- Usando los valores transformados, podremos entonces calcular la media de la población como

$$M = a(p - q) + 2pqd \tag{8.45}$$

8.4 Efecto medio

Hasta ahora hemos discutido la separación de lo genético y lo ambiental, asumiendo que el genotipo es lo relevante desde el punto de vista genético, ya que lo que identifica a los individuos desde ese punto de vista es no solo la suma de los valores de los alelos presentes sino también el aporte de las combinaciones específicas, tanto **intralocus** como **interloci**. Sin embargo, cuando nos interesa entender la capacidad de transmitir un determinado valor genético de un individuo a sus descendientes, entonces debemos enfocarnos en los mecanismos de herencia y en la forma en que los individuos transmiten información a sus descendientes. En los organismos diploides con dos sexos, cada individuo transmite a su descendencia solamente la mitad de la información que este tendrá para cada posición del **genoma autosomal** (esto no es cierto para los cromosomas sexuales, por ejemplo, ni para los genomas de organelos como la mitocondria), viniendo la otra mitad del otro padre (o madre). Dicho de otra forma, los padres/madres transmiten a sus hijos uno de los dos alelos que cada uno de ellos posee en cada loci del genoma. Como nunca transmiten el par de alelos que cada padre/madre posee, entonces nunca transmiten genotipos y, por lo tanto, mal podrían transmitir su correspondiente valor.

En este sentido, es necesario entender cuál es el impacto en el cambio de valor medio de una población que proporciona un alelo cuando es introducido en la misma. Supongamos que tenemos una población que tiene dos alelos, A_1 y A_2, con frecuencias p y q, tanto en machos como en hembras. Supongamos ahora que elegimos como reproductor macho uno o más animales que son homocigotos para el alelo A_1, es decir es A_1A_1. Eso nos asegura que los gametos producidos por los machos serán A_1 (el mismo razonamiento vale si intercambiamos el rol de los sexos). Esos gametos se unirán a los gametos producidos por las hembras, que como son sacados al azar de la población estarán en frecuencias p el alelo A_1 y $q = 1 - p$ el alelo A_2.

Si ponemos esta información en un cuadro, donde la frecuencia del macho aparece en la primera fila y las frecuencias de los gametos en las hembras en las columnas, tenemos el siguiente resumen de la situación:

	$\mathbf{A_1}$	$\text{fr}(A_1) = p$	$\mathbf{A_2}$	$\text{fr}(A_2) = q$
$\mathbf{A_1}$, $\text{fr}(A_1) = 1$	A_1A_1	p	A_1A_2	q

Es decir, como el macho produce gametos A_1 con frecuencia igual a uno, los únicos genotipos posibles serán A_1A_1 y A_1A_2, con frecuencias respectivas $1 \times p = p$ y $1 \times q = q$. Como los valores de los genotipos A_1A_1 y A_1A_2 son, respectivamente, a y d, entonces multiplicando cada valor por su frecuencia obtenemos la media de valores genotípicos de la población luego de haber asegurado que uno de los alelos de los individuos es A_1 y el otro está tomado al azar, es decir

$$M_1 = a \times p + d \times q = pa + dq \tag{8.46}$$

Como vimos antes, el valor genotípico de la población antes del apareamiento con el animal A_1A_1 era

$$M_0 = a \times p^2 + d \times 2pq + (-a) \times q^2 = a[p^2 - q^2] + 2pqd \ \therefore \tag{8.47}$$

$$M_0 = a[(p+q)(p-q)] + 2pqd = a(p-q) + 2pqd \tag{8.48}$$

La diferencia, por lo tanto, atribuible a la introducción de alelo A_1 como uno presente en todos los genotipos es

$$M_1 - M_0 = pa + dq - [a(p-q) + 2pqd] = pa + qd - pa + qa - 2pqd \Leftrightarrow \tag{8.49}$$

$$q[d + a - 2pd] = q[a + d(1-2p)] = q[a + d(p+q-2p)] = q[a + d(q-p)] \tag{8.50}$$

Si definimos $\alpha = a + d(q-p)$, tenemos entonces que el **efecto medio del alelo** A_1, que llamaremos α_1, es

$$\alpha_1 = q\alpha \tag{8.51}$$

De la misma manera, para estudiar el efecto de asegurar la presencia del otro alelo en todos los genotipos de los descendientes (de la población original), podemos hacer el mismo cuadro que antes, pero ahora el o los machos serán homocigotos para el alelo A_2, es decir serán A_2A_2, mientras que los gametos de las hembras serán nuevamente extraídos al azar de la población (con frecuencias p y q). Poniendo todo junto, tenemos el siguiente cuadro:

	A_1	$\text{fr}(A_1) = p$	A_2	$\text{fr}(A_2) = q$
A_2, $\text{fr}(A_2) = 1$	A_1A_2	p	A_1A_2	q

En este caso, la media de la nueva población estará dada por

$$M_1 = d \times p - a \times q = dp - aq \tag{8.52}$$

La diferencia, por lo tanto, atribuible a la introducción de alelo A_2 como uno presente en todos los genotipos es

$$M_1 - M_0 = dp - qa - [a(p-q) + 2pqd] = dp - qa - pa + qa - 2pqd \Leftrightarrow \quad (8.53)$$

$$\Leftrightarrow -p[a - d + 2qd] = -p[a + d(2q-1)] = -p[a + d(2q - p - q)] = -p[a + d(q-p)] \quad (8.54)$$

y usando la definición previa de $\alpha = a + d(q-p)$, entonces el cambio de valor genotípico en la población producto de asegurar la presencia del alelo A_2 en los genotipos será

$$\alpha_2 = -p[a + d(q-p)] = -p\alpha \quad (8.55)$$

que es el **efecto medio del alelo** A_2.

Resumiendo, podemos juntar ambos cuadros y tenemos el cambio en la media genotípica de la población correspondiente a cada alelo:

Gameto	A_1A_1	A_1A_2	A_2A_2	Valor medio	Media poblacional	Efecto medio del gen
$\mathbf{A_1}$	p	q		$pa + qd$	$a(p-q) + 2pqd$	$q[a + d(q-p)]$
$\mathbf{A_2}$		p	q	$pd - qa$	$a(p-q) + 2pqd$	$-p[a + d(q-p)]$

Otra forma de entender lo que implican estos cambios es a través del concepto de **efecto medio de sustitución de alelos**, que se puede aplicar cuando tenemos dos alelos en un gen. Supongamos que podemos reemplazar al azar alelos A_2 de la población de individuos, **sustituyéndolos** con alelos A_1. Los genotipos afectados serán los A_1A_2, cuya mitad son alelos A_2 y que tienen frecuencia $2pq$, porque su frecuencia de cambios será pq y los A_2A_2, con frecuencia q^2. Asumiendo que cambiamos un solo alelo por genotipo, los cambios serán del genotipo A_1A_2 al A_2A_2 y del A_2A_2 al A_1A_2. Poniendo todo junto tenemos

$$\frac{pq(a-d) + q^2[d - (-a)]}{q^2 + pq} = \frac{pq(a-d) + q^2(d+a)}{q(p+q)} = \frac{pq(a-d) + q^2(d+a)}{q} = \quad (8.56)$$

$$\begin{aligned} \frac{q[p(a-d) + q(d+a)]}{q} &= p(a-d) + q(d+a) \\ &= pa - pd + qd + qa = a(p+q) + d(q-p) \\ &= a + d(q-p) = \alpha \end{aligned} \quad (8.57)$$

como lo definimos previamente.

De hecho, usando la notación A_1A_1, A_1A_2 y A_2A_2 **para los valores** de los correspondientes genotipos, la primera fracción en (8.4) se puede escribir como

$$\alpha = \frac{pq(A_1A_1 - A_1A_2) + q^2(A_1A_2 - A_2A_2)}{q^2 + pq} = \frac{q[p(A_1A_1 - A_1A_2) + q(A_1A_2 - A_2A_2)]}{q} = \quad (8.58)$$

$$p(A_1A_1 - A_1A_2) + q(A_1A_2 - A_2A_2) \quad (8.59)$$

que es una forma más conceptual de ver el significado del **efecto de sustitución alélica.** Otra forma de apreciar esto mismo es aún más conceptual e implica ver el efecto de α al sustituir un alelo por el otro como la diferencia entre los **efectos medios** de los alelos que se intercambian (es decir, α_1y α_2). Utilizando las ecuaciones (8.51) y (8.55), entonces α sería

$$\alpha_1 - \alpha_2 = q\alpha - [-p\alpha] = q\alpha + p\alpha = \alpha(p+q) = \alpha \tag{8.60}$$

que es lo que queríamos mostrar.

Ejemplo 8.3

Usando los datos previos del estudio de Bennett et al. (2019), calcular el **efecto medio** de los alelos L y F para la característica peso al nacimiento, así como el **efecto de sustitución** correspondiente.

Tenemos varias formas de llegar a lo mismo, pero una de ellas es primero estimar el **efecto de sustitución** (α) y luego usarlo para calcular α_1 y α_2.De acuerdo con la definición $\alpha = a + d(q-p)$. Como ya tenemos todos los valores se trata solo de sustituirlos

$$\alpha = a + d(q-p) = 1{,}3 + (-0{,}6)(0{,}793 - 0{,}207) = 1{,}3 - 0{,}6(0{,}586) = 0{,}9484 \text{ kg} \tag{8.61}$$

Luego, usando los resultados de las ecuaciones (8.51) y (8.55), tenemos

$$\alpha_1 = q\alpha = 0{,}793 \times 0{,}9484 \text{ kg} = 0{,}7521 \text{ kg} \alpha_2 = -p\alpha = -0{,}207 \times 0{,}9484 \text{ kg} = -0{,}1963 \text{ kg} \tag{8.62}$$

Obviamente, como verificación trivial, $\alpha = \alpha_1 - \alpha_2 = [0{,}7521 - (-0{,}1963)]$ kg $= 0{,}9484$ kg $= \alpha$.

PARA RECORDAR

- En una población que posee dos alelos A_1 y A_2, el **efecto medio** del alelo A_1, que llamaremos α_1, es $\alpha_1 = q\alpha$. Como contraparte, el **efecto medio** del alelo A_2, que llamaremos α_2, es $\alpha_2 = -p[a + d(q-p)] = -p\alpha$.
- El efecto medio de la sustitución de un alelo por otro (cuando tenemos dos alelos en un gen) será $\alpha_1 - \alpha_2 = q\alpha - [-p\alpha] = q\alpha + p\alpha = \alpha(p+q)$, es decir α.

8.5 Valor reproductivo (o valor de cría)

Como vimos antes, en organismos diploides con el modo de reproducción sexual, estos transmiten a su descendencia la mitad de sus alelos. Este proceso es aleatorio en el sentido de que la combinación particular de alelos que va en cada gameto está determinada por el azar, ya que además de los procesos de recombinación a nivel de cada cromosoma, la combinación de cromosomas que va a cada gameto es también al azar (en el sentido de que puede ir la copia «paterna» o «materna» que tiene el organismo que se reproduce, no en qué número de cromosomas van). Más aún, como vimos antes, los organismos diploides de reproducción sexual pasan a los descendientes solamente alelos, no genotipos, por lo que fue necesario definir el **efecto medio de un alelo** precisamente como el efecto teórico de introducir dicho alelo de forma obligatoria en todos los genotipos de la nueva generación.

En genética animal, por ejemplo, una posibilidad de conocer el **valor reproductivo** (también llamado valor de cría o EBV (*estimated breeding value*, según sus siglas en inglés) de un individuo determinado sería realizar en la práctica el experimento conceptual que nos permitió definir el concepto de **efecto medio de un alelo**, es decir, aparear a nuestro candidato con una muestra lo más amplia y representativa posible de la población. Los hijos de nuestro candidato se distribuirán en forma aleatoria en torno a un valor medio particular que en general no coincidirá con la media de la población. El desvío de la media de los hijos de nuestro candidato respecto a la media de la población será un estimador de **la mitad del valor cría** (o del valor reproductivo) del candidato. Será la mitad porque el candidato les aporta a sus hijos solo la mitad de sus alelos, la otra mitad viniendo al azar del otro progenitor. Es de destacar que cuanto más descendientes se evalúen más cerca estaremos del **verdadero valor de cría** del candidato (que se alcanzaría perfectamente si el mismo se aparease con un número infinito de individuos de la población) y por esto el término **estimador** (que como todo estimador posee una incertidumbre asociada, algo que veremos más adelante).

Como el valor reproductivo se trata del efecto que cada alelo aporta al individuo, en nuestro modelo se trata de **efectos aditivos** y debemos sumarlos. En el caso de un locus con dos alelos las posibilidades se reducen al cuadro siguiente:

Genotipo	Valor alelo 1	Valor alelo 2	Valor reproductivo
$\mathbf{A_1A_1}$	$\alpha_1 = q\alpha$	$\alpha_1 = q\alpha$	$2\alpha_1 = 2q\alpha$
$\mathbf{A_1A_2}$	$\alpha_1 = q\alpha$	$\alpha_2 = -p\alpha$	$\alpha_1 + \alpha_2 = (q-p)\alpha$
$\mathbf{A_2A_2}$	$\alpha_2 = -p\alpha$	$\alpha_2 = -p\alpha$	$2\alpha_2 = -2p\alpha$

Por ejemplo, el genotipo A_1A_1 está formado por dos alelos A_1, que cada uno aporta un valor de $\alpha_1 = q\alpha$, por lo que el **valor de cría** de los individuos con este genotipo es de $2\alpha_1 = 2q\alpha$. En el caso de un individuo con genotipo A_1A_2 el mismo tendrá un alelo de cada tipo, por lo que su valor de cría será $\alpha_1 + \alpha_2 = q\alpha + (-p\alpha) = (q-p)\alpha$. Finalmente, en el caso de un individuo con genotipo A_2A_2, el mismo tendrá dos alelos A_2 y, por lo tanto, su valor de cría será $2\alpha_2 = -2p\alpha$.

En general, el valor de cría se expresa como desvío de la media de la población, y como esta depende de las frecuencias de los alelos, entonces los **valores de cría** (igual que el efecto medio de cada alelo) **son específicos a una población de referencia**. El valor reproductivo

medio de la población viene dado por la multiplicación de los valores reproductivos de cada genotipo por su frecuencia correspondiente, es decir

$$\bar{A} = p^2(2q\alpha) + 2pq[(q-p)\alpha] + q^2(-2p\alpha) = 2\alpha(p^2q + pq^2 - p^2q - pq^2) = 0 \quad (8.63)$$

Es decir, el valor de cría promedio de una población, en un modelo de un locus con dos alelos es cero, lo que no sorprende ya que realizamos todas las cuentas a partir de desvíos de la media de la población.

En el caso del valor de cría, al tratarse de la suma de los efectos aditivos de los alelos, el concepto es inmediatamente generalizable, tanto a un número indeterminado de loci como a un número de alelos igual o mayor a 2 en cada uno de esos loci. En cada locus un individuo puede tener solo dos alelos, por más que existan mayor cantidad de alelos posibles para ese locus, por lo que si utilizamos los números 1 y 2 para referirnos a los dos lugares en cada locus que pueden ser ocupados por alelos y utilizamos la letra k para los n loci que influencian la característica de interés, entonces tenemos

$$A = \sum_{k=1}^{n} \alpha_{k,1} + \sum_{k=1}^{n} \alpha_{k,2} = \sum_{j=1,k=1}^{j=2,k=n} \alpha_{k,j} \quad (8.64)$$

Es decir, en la ecuación (8.64) estamos sumando los aportes de cada locus, que son a su vez de dos alelos, que pueden ser iguales o distintos entre ellos, entre un número variable de alelos en cada locus. La figura 8.7 muestra una representación gráfica del significado de la ecuación (8.64). En la misma se observan 3 cromosomas (representados por diferentes colores) que albergan un total de 6 loci (identificados por letras de la **A** a la **E**) que contribuyen a una característica cuantitativa. El valor de cría será la suma de los efectos medios de los alelos presentes en cada uno de los loci, es decir

$$A = \sum_{j=1,k=1}^{j=2,k=6} \alpha_{k,j} = (\alpha_{A1} + \alpha_{A3}) + (\alpha_{B5} + \alpha_{B2}) + (\alpha_{C8} + \alpha_{C4}) + (\alpha_{D7} + \alpha_{D2}) + \quad (8.65)$$

$$(\alpha_{E1} + \alpha_{E1}) + (\alpha_{F9} + \alpha_{F2}) \quad (8.66)$$

donde los distintos α son los **efectos medios** de los alelos correspondientes.

Si bien hemos dado dos definiciones diferentes del valor de cría, una de base teórica basada en el **efecto medio de los alelos** y otra más práctica, basada en la descendencia del candidato, ambas son estrictamente equivalentes cuando no existen efectos epistáticos, ya que si existe desequilibrio de ligamiento los alelos de diferentes loci van a ir juntos en relaciones distintas al producto de sus frecuencias y, por lo tanto, determinadas combinaciones (efectos epistáticos) serán más o menos frecuentes.

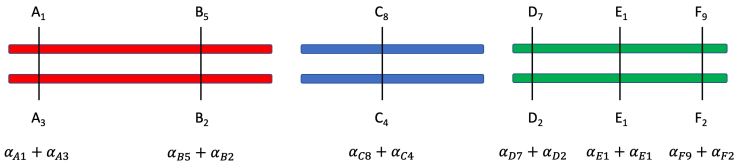


Figura 8.7. Valor de cría (o reproductivo) de un individuo diploide. Tres cromosomas, marcados por colores diferentes, presentan un total de 6 loci (nombrados A-E) que contribuyen al valor genotípico y fenotípico de una característica. El **valor de cría** está completamente determinado por la suma de los aportes individuales de todos los alelos presentes en los 6 loci. Cada loci tiene un número distinto de alelos, pero para cada uno de ellos siempre habrá dos alelos en particular, identificados por los números en el subíndice.

Ejemplo 8.4

En el trabajo de Bennett et al. (2019), los autores describen, además del gen de la miostatina, el aporte de los diplotipos (un par determinado de haplotipos) **CC-CC**, **CC-NN** y **NN-NN** en el gen de la μ-calpaína. Para este gen se utilizaron dos marcadores (SNP), el CAPN1_316 (BTA29; rs17872000) y el CAPN1_4751 (BTA29; rs17872050). El marcador CAPN1_316 segrega los alelos C y G, mientras que CAPN1_4751 segrega los alelos C y T. Un haplotipo común es el **CC** (o sea, una C en el CAPN1_316 junto con una C en el CAPN1_4751), por lo que el resto se consideraron en conjunto como **NN** (es decir, alelos no-C en los dos marcadores).

En un estudio de simulación (hipotético) sobre datos del estudio de Bennett et al. (2019), se llega a la conclusión de que con frecuencias del alelo L de $p_{MSTN} = 0{,}20$, el efecto de sustitución en dicha población es de $\alpha_{MSTN} = 6{,}6$ cm^2 para área del ojo del bife, mientras que para el diplotipo CC, con frecuencia de $p_{CAPN1} = 0{,}45$, el efecto de sustitución en dicha población es de $\alpha_{CAPN1} = 0{,}3$ cm^2. Calcular los valores de cría de los 9 genotipos posibles (3 genotipos de **MSTN** x 3 diplotipos de **CAPN1**).

Primero, usando la información de los efectos de sustitución y de las frecuencias para cada gen, calculamos los efectos medios de los alelos (los haplotipos pueden considerarse alelos). Es decir, para la miostatina (MSTN), tenemos

$$\alpha_{MSTNaL} = q(\alpha_{MSTN}) = (1 - 0{,}20)6{,}6 \text{ cm}^2 = 0{,}8 \times 6{,}6 \text{ cm}^2 = 5{,}28 \text{ cm}^2 \quad (8.67)$$

$$\alpha_{MSTNaF} = -p(\alpha_{MSTN}) = -0{,}20 \times 6{,}6 \text{ cm}^2 = -1{,}32 \text{ cm}^2 \quad (8.68)$$

Para el gen de la μ-caspaína (CAPN1), tenemos, de la misma forma

$$\alpha_{CAPN1aCC} = q\alpha_{CAPN1} = (1 - 0{,}45)0{,}3 \text{ cm}^2 = 0{,}55 \times 0{,}3 \text{ cm}^2 = 0{,}165 \text{ cm}^2 \quad (8.69)$$

$$\alpha_{CAPN1aNN} = -p\alpha_{CAPN1} = -0{,}45 \times 0{,}3 \text{ cm}^2 = -0{,}135 \text{ cm}^2 \quad (8.70)$$

Poniendo todo junto en el cuadro, recordando que los efectos aditivos de los alelos presentes se suman, tenemos (expresados en cm^2)

	$CC-CC$	$CC-NN$	$NN-NN$
LL	$2\times 5{,}28+2\times 0{,}165=10{,}89$	$2\times 5{,}28+0{,}165-0{,}135=10{,}59$	$2\times 5{,}28-2\times 0{,}135=10{,}29$
LF	$5{,}28-1{,}32+2\times 0{,}165=4{,}29$	$5{,}28-1{,}32+0{,}165-0{,}135=3{,}99$	$5{,}28-1{,}32-2\times 0{,}135=3{,}69$
FF	$-2\times 1{,}32+2\times 0{,}165=-2{,}31$	$-2\times 1{,}32+0{,}165-0{,}135=-2{,}61$	$-2\times 1{,}32-2\times 0{,}135=-2{,}91$

Derivación de los efectos medios por mínimos cuadrados

Una forma alternativa de derivar los efectos de sustitución es a partir del método de mínimos cuadrados, buscando los valores de los parámetros $\alpha 1$ y $\alpha 2$ de tal manera que minimicen las distancias entre los genotipos y la recta que pasa por los valores reproductivos. Esto podemos hacerlo tanto en la **escala transformada** (con el cero igual al punto medio entre los homocigotas) o como desvíos de la media M. De hecho, es posible la transformación entre ambas escalas de forma muy sencilla, como veremos en la siguiente sección. Por lo tanto, haremos la derivación en la escala transformada y al final, cambiaremos de escala para verificar que es el mismo resultado obtenido previamente como desvíos de la media poblacional.

Si, siguiendo la notación de John H. Gillespie (2004), llamamos $Q(\alpha 1,\alpha 2)$ a la función que establece la distancia de los genotipos a la recta de valores reproductivos, entonces

$$Q(\alpha 1,\alpha 2)=p^2(a-2\alpha_1)^2+2pq(d-\alpha_1-\alpha_2)^2+q^2(-a-2\alpha_2)^2 \tag{8.71}$$

Para encontrar los valores de los parámetros α_1 y α_2 que minimicen esta función debemos derivar e igualar a cero. Como la derivada de una suma es la suma de las derivadas, usando la regla de la cadena y teniendo en cuenta que cuando derivamos respecto a α_1, α_2 pasa a ser tratada como una constante (y viceversa), entonces

$$\frac{\partial Q}{\partial \alpha_1}=-4[p^2(a-2\alpha_1)+pq(d-\alpha_1-\alpha_2)]=0 \tag{8.72}$$

$$\frac{\partial Q}{\partial \alpha_2}=-4[pq(d-\alpha_1-\alpha_2)]+q^2(-a-2\alpha_2)]=0 \tag{8.73}$$

Operando con la primera ecuación de (8.73), tenemos

$$-4[p^2(a-2\alpha_1)+pq(d-\alpha_1-\alpha_2)]=-4p[p(a-2\alpha_1)+q(d-\alpha_1-\alpha_2)] \tag{8.74}$$

y dividiendo entre $-4p$, usando además el hecho de que $p\alpha_1+q\alpha_2=0$ ya que es el efecto aditivo medio, tenemos

$$p(a-2\alpha_1)+q(d-\alpha_1-\alpha_2)=pa-2p\alpha_1+qd-q\alpha_1-q\alpha_2= \tag{8.75}$$

$$pa-p\alpha_1-p\alpha_1+qd-q\alpha_1-q\alpha_2=pa-(p\alpha_1+q\alpha_2)-\alpha_1(p+q)+qd= \tag{8.76}$$

$$pa+qd-\alpha_1 \tag{8.77}$$

e igualando a cero el resultado de la ecuación (8.77) tenemos que

$$pa + qd - \alpha_1 = 0 \Leftrightarrow \alpha_1 = pa + qd \tag{8.78}$$

De manera análoga, operando con la segunda ecuación de (8.73), dividiendo entre $-4q$, tenemos

$$p(d - \alpha_1 - \alpha_2)] + q(-a - 2\alpha_2)] = pd - p\alpha_1 - p\alpha_2 - qa - 2q\alpha_2 = \tag{8.79}$$

$$pd - (p\alpha_1 + q\alpha_2) - \alpha_2(p + q) - qa = pd - qa - \alpha_2 \tag{8.80}$$

e igualando a cero el resultado de la ecuación (8.80) tenemos que

$$pd - qa - \alpha_2 = 0 \Leftrightarrow \alpha_2 = pd - qa \tag{8.81}$$

Para verificar que el resultado obtenido en la **escala transformada** es idéntico al obtenido previamente como desvío de la media poblacional, teniendo en cuenta que la media de la población es $M = a(p - q) + 2pqd$, tenemos entonces

$$\alpha_1 = p(a - M) + q(d - M) = pa + qd - pM - qM = \tag{8.82}$$

$$= pa + qd - M(p + q) = pa + qd - M = pa + qd - [a(p - q) + 2pqd] \ \therefore \tag{8.83}$$

$$\alpha_1 = pa + qd - pa + qa - 2pqd = q[a + d - 2pd] = q[a + d(1 - 2p)] \ \therefore \tag{8.84}$$

$$\alpha_1 = q[a + d(1 - p - p)] = q[a + d((1 - p) - p)] = q[a + d(q - p)] \tag{8.85}$$

que es el mismo resultado que habíamos obtenido previamente, o usando la definición original de $\alpha = a + d(q - p)$, $\alpha_1 = q\alpha$.

De la misma manera, para α_2 en la escala como desvío de la media poblacional, tenemos

$$\alpha_2 = p(d - M) - q(a - M) = pd - qa - M = pd - qa - [a(p - q) + 2pqd] = \tag{8.86}$$

$$= pd - qa - pa + qa - 2pqd = -p[a + d(2q - 1)] = -p[a + d(q + q - 1)] \ \therefore \tag{8.87}$$

$$\alpha_2 = -p[a + d(q - (1 - q))] = -p[a + d(q - p)] \tag{8.88}$$

que es el mismo resultado obtenido previamente, o $\alpha_2 = -p\alpha$, mostrando la consistencia de las dos aproximaciones.

PARA RECORDAR

- El **valor de cría** es un estimador del mérito genético de un animal. El mismo se expresa como desvío de la media de la población, y como esta depende de las frecuencias de los alelos, entonces los mismos son específicos a una población de referencia.

- Dado que en cada locus un individuo puede únicamente tener dos alelos (que denominaremos 1 y 2 haciendo referencia al lugar del locus que ocupan), y usando la letra k para los n loci que influencian la característica de interés

$$A = \sum_{k=1}^{n} \alpha_{k,1} + \sum_{k=1}^{n} \alpha_{k,2} = \sum_{j=1,k=1}^{j=2,k=n} \alpha_{k,j} \tag{8.89}$$

El **valor de cría** será la suma de los efectos medios de los alelos presentes en cada uno de los loci.

8.6 Desvío de dominancia

En la sección anterior avanzamos en la comprensión de uno de los componentes del genotipo, el que tiene que ver con los valores aditivos de los alelos que lo constituyen. Pero como hemos discutido previamente, en general la expresión de los alelos presentes en un locus trasciende el valor de cada uno en particular. La razón de esto es fácil de imaginar en la biología subyacente al problema. Por un lado, no necesariamente las dos copias presentes en un locus determinado tienen por qué transcribirse/traducirse (por ejemplo, en determinadas regiones de cromosomas como el X). Por el otro, la cinética de las reacciones, en el caso de que se trate de un locus que produce un sustrato enzimático, no suele ser simplemente aditiva ya que el comportamiento es varias veces sigmoidal. Todas estas razones nos llevan a pensar que, en general, debemos considerar en forma matemática el desvío correspondiente a la interacción entre alelos dentro del locus como una posibilidad bastante razonable.

De acuerdo con nuestro modelo genético para el caso de un solo locus, como no consideramos los efectos epistáticos ya que no hay otro locus con el cual interactuar, entonces $G = A + D * I = A + D$ y, por lo tanto,

$$G = A + D \Leftrightarrow D = G - A \tag{8.90}$$

Es decir, de acuerdo con nuestro modelo, la interacción entre alelos del mismo locus lo modelamos como un desvío de los valores genotípicos respecto a los valores de cría (o valores aditivos). En la sección precedente expresamos los valores reproductivos como desvíos de la media de la población, por lo que si queremos mantener la consistencia de nuestras derivaciones del desvío de dominancia debemos entonces expresarlo de la misma forma. Para eso vamos a expresar primero los valores genotípicos como desvíos de la media de la población, lo que resulta inmediato al restar $M = a(p-q) + 2pqd$ a cada uno de los tres valores genotípicos en la escala transformada (a, d y $-a$). Operando, si denotamos con γ a los desvíos genotípicos respecto a la media poblacional, es decir, $\gamma_{ij} = G_{ij} - M = G_{ij} - [a(p-q) + 2pqd]$, tenemos entonces

$$\gamma_{11} = a - [a(p-q) + 2pqd] = a(1-p+q) - 2pqd = a(q+q) - 2pqd = 2qa - 2pqd \;\therefore \tag{8.91}$$

$$\gamma_{11} = 2q(a - pd) \tag{8.92}$$

$$\gamma_{12} = d - [a(p-q) + 2pqd] = a(q-p) + d(1-2pq) \tag{8.93}$$

$$\gamma_{22} = -a - [a(p-q) + 2pqd] = -a(1-q+p) - 2pqd = -a(p+p) - 2pqd \ \therefore \quad (8.94)$$

$$\gamma_{22} = -2p(a+qd) \quad (8.95)$$

Pero de acuerdo con nuestra definición $\alpha = a + d(q-p) \Leftrightarrow a = \alpha - d(q-p)$, por lo que sustituyendo en la ecuación (8.95) tenemos para los homocigotas las siguientes expresiones

$$\gamma_{11} = 2q(a-pd) = 2q[\alpha - d(q-p) - pd] = 2q(\alpha - qd + pd - pd) \ \therefore \quad (8.96)$$

$$\gamma_{11} = 2q(\alpha - qd) \quad (8.97)$$

$$\gamma_{22} = -2p(a+qd) = -2p[\alpha - d(q-p) + qd] = -2p(\alpha - qd + pd + qd) \ \therefore \quad (8.98)$$

$$\gamma_{22} = -2p(\alpha + pd) \quad (8.99)$$

Para los heterocigotas el principio es el mismo, aunque la derivación nos lleva un poquito más de trabajo para finalmente arribar a una expresión sencilla

$$\gamma_{12} = a(q-p) + d(1-2pq) = [\alpha - d(q-p)](q-p) + d(1-2pq) = \quad (8.100)$$

$$= \alpha(q-p) - d(q-p)^2 + d - 2pqd = \alpha(q-p) - d(q^2 - 2pq + p^2) + d - 2pqd = \quad (8.101)$$

$$\alpha(q-p) - d(q^2+p^2) + 2pqd + d - 2pqd = \alpha(q-p) + d[1-(p^2+q^2)] = \quad (8.102)$$

$$= \alpha(q-p) + d(2pq) \ \therefore \quad (8.103)$$

$$\gamma_{12} = \alpha(q-p) + 2pqd \quad (8.104)$$

Una forma sencilla de entender el procedimiento que hemos utilizado es organizar toda la información en un cuadro como el siguiente:

	$\mathbf{A_1A_1}$	$\mathbf{A_1A_2}$	$\mathbf{A_2A_2}$
Frecuencia H-W	p^2	$2pq$	q^2
Valor genotípico	a	d	$-a$
Genotípico como desvío de M	$2q(\alpha - qd)$	$(q-p)\alpha + 2pqd$	$-2p(\alpha + pd)$
Valor Reproductivo (de cría)	$2q\alpha$	$(q-p)\alpha$	$-2p\alpha$
Desvío de dominancia	$-2q^2d$	$2pqd$	$-2p^2d$

En un locus con dos alelos tenemos 3 genotipos posibles, que si además le incorporamos la restricción de que las frecuencias se encuentren en equilibrio de Hardy-Weinberg, estarán en las clásicas proporciones p^2, $2pq$ y q^2 (para una frecuencia p del alelo A_1). Los valores, como ya hemos visto, en la escala transformada son a, d y $-a$. A partir de las ecuaciones (8.95) a (8.104), los valores genotípicos como desvíos de la media M son los que aparecen en la tercera línea del cuadro. Si ahora le restamos a esos valores genotípicos los correspondientes valores reproductivos que aparecen en la cuarta línea (y que dedujimos en la sección precedente), tenemos ahora los **desvíos de dominancia** que aparecen en la última línea. Por ejemplo, para el genotipo A_1A_1, su valor como desvío de la media de la población (como vimos en la ecuación (8.99)) es $\gamma_{11} = 2q(\alpha - qd) = 2q\alpha - 2q^2d$, por lo que si ahora le restamos su valor reproductivo $A_{11} = 2q\alpha$, tenemos el valor $-2q^2$.

Lo mismo se aplica para los otros genotipos, por lo que resumiendo tenemos

$$\delta_{11} = \gamma_{11} - A_{11} = 2q\alpha - 2q^2 d - 2q\alpha = -2q^2 d \tag{8.105}$$

$$\delta_{12} = \gamma_{12} - A_{12} = (q-p)\alpha + 2pqd - (q-p)\alpha = 2pqd \tag{8.106}$$

$$\delta_{22} = \gamma_{22} - A_{22} = -2p\alpha - 2p^2 d - (-2p\alpha) = -2p^2 d \tag{8.107}$$

Es importante notar que al igual que los valores reproductivos, los desvíos de dominancia también dependen de las frecuencias de los alelos, así como de los valores de a y d, por lo que son específicos de cada población. Una forma gráfica de entender la relación entre los valores genotípicos, los valores reproductivos y los desvíos de dominancia se aprecia en la figura 8.8. En la misma, podemos ver los tres genotipos representados por los círculos negros y sus correspondientes valores en la escala transformada (a, d y $-a$) en el eje de las ordenadas a la izquierda. Si hacemos la regresión de los valores genotípicos en el número de copias del alelo A_1, por ejemplo, usando la técnica de mínimos cuadrados, entonces obtenemos la línea roja a trazos que vemos en la figura. El cambio en las ordenadas de esta línea con un cambio de una unidad en el número de copias de A_1 es la pendiente de dicha recta y se corresponde con nuestra definición del **efecto de sustitución** y es, por lo tanto, de α. La recta no queda equidistante de los puntos negros porque cada punto tiene una «inercia» diferente, dada por la frecuencia del correspondiente genotipo: en el caso de la figura 8.8, $q^2 = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$ pesa mucho menos que $p^2 = \left(\frac{3}{4}\right)^2 = \frac{9}{16}$ y, por lo tanto, la diferencia de distancias de los puntos negros a la recta. Ahora, si observamos el valor de los tres puntos rojos que están sobre la recta en la escala de la derecha, que es la escala como desvíos de la media poblacional $M = a(p-q) + 2pqd$, entonces tenemos los valores reproductivos que se corresponden con cada genotipo. Por lo tanto, la distancia entre los puntos rojos y los puntos negros, es decir la ordenada de los puntos negros (genotipos) menos la de los correspondientes puntos rojos (valores reproductivos) son los desvíos de dominancia. Notemos que en la ecuación (8.107) todos los desvíos de dominancia depende exclusivamente de la frecuencia alélicas (p o $q = 1 - p$) y del valor del desvío del genotipo heterocigota respecto del punto medio entre los homocigotas (o lo que es lo mismo, del apartamiento de este genotipo del cero en la escala transformada), por lo que cuando el mismo sea cero ($d = 0$) los valores genotípicos serán idénticos a los valores reproductivos.

Para ilustrar el efecto del cambio en el desvío de dominancia con la frecuencia del alelo A_1 y del desvío del heterocigoto respecto al punto medio de los homocigotos en la figura 8.9 ilustramos diferentes combinaciones de estos parámetros. En la misma figura, los círculos azules representan los valores genotípicos, mientras que las líneas rojas a trazos representan las regresiones de los mismos en el número de copias del alelo A_1 (lo que nos da los valores de cría, o aditivos, de cada genotipo). En la línea superior de la figura se ven los gráficos cuando el desvío del heterocigoto respecto del punto medio es cero ($d = 0$), es decir que el comportamiento de los genotipos es totalmente aditivo (alcanza con conocer el número de copias de A_1 en cada uno de ellos y la pendiente α para predecir con exactitud su valor genotípico), por eso los puntos están sobre la recta siempre. En la fila del medio apreciamos la situación con dominancia parcial ($d = 0{,}75a$), donde el genotipo A_1A_2 está mucho más cerca en valor del A_1A_1 que del A_2A_2 y por eso está por encima de la recta de regresión. A su vez, al movernos de izquierda a derecha aumenta la frecuencia del alelo A_1 (de $p = 0{,}50$, $p = 0{,}75$ a $p = 0{,}90$), por lo que la importancia del genotipo A_1A_1 (frecuencia p^2) en la determinación de la pendiente crece, en particular respecto al A_2A_2.

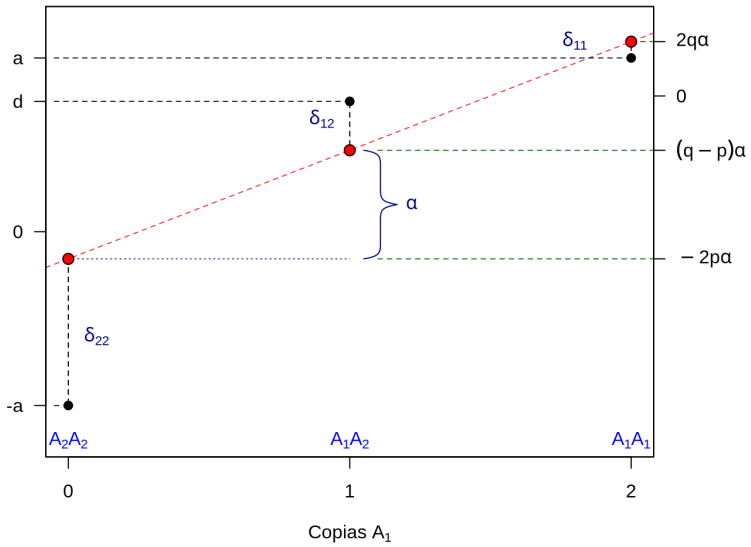


Figura 8.8. Representación gráfica de los valores genotípicos, círculos negros, valores de cría (o reproductivos), círculos rojos y la desviación entre ambos valores correspondiente al desvío de dominancia. La línea roja a trazos representa la recta de regresión de los valores genotípicos en el número de copias del alelo A_1. La pendiente de dicha recta es α (el efecto de sustitución) y los valores de las ordenadas de dicha recta para los puntos en las abscisas 0, 1 y 2 se corresponden con los valores de cría de los genotipos con dichos número de copias del alelo A_1. Mientras que el eje de la izquierda refiere a los valores de los genotipos en la escala transformada, los valores del eje a la derecha se corresponden con desvíos de la media $M = a(p-q) + 2pqd$. Los valores usados para la representación fueron $d = \frac{3}{4}a$ y $p = \frac{3}{4}$. Figura de elaboración propia sobre idea en Douglas S. Falconer (1983).

Por esta razón, a medida que nos vamos hacia la derecha en los gráficos la pendiente parece mucho más determinada por los valores de A_1A_1 y A_1A_2 que por los de A_2A_2 y mientras que la distancia de A_1A_1 a la recta decrece, la de A_2A_2 aumenta. Este efecto se ve aún más magnificado cuando tenemos efecto de sobredominancia, como en la última fila de la figura ($d = 1{,}75a$), donde a altas frecuencia del alelo A_1 llega a invertirse el signo de la pendiente, casi ignorando el aporte del genotipo A_2A_2 en la regresión ya que su proporción es ínfima ($q^2 = (1-p)^2 = (1-0{,}90)^2 = 0{,}10^2 = 0{,}01 = 1\,\%$). Claramente, cuanto más cerca se encuentre la recta de regresión de los puntos, mejor explican los efectos aditivos los valores genotípicos.

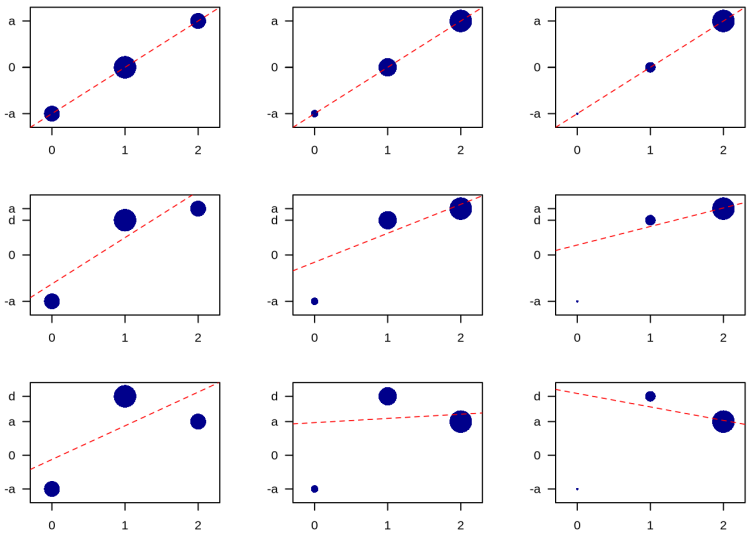


Figura 8.9. Representación gráfica de los valores genotípicos (círculos azules) y la correspondiente recta de regresión de los valores genotípicos en el número de copias del alelo A_1 (cuya pendiente es α) para diferentes combinaciones de frecuencias del alelo A_1 (de izquierda a derecha, $p = 0{,}50$, $p = 0{,}75$ y $p = 0{,}90$) y del desvío del heterocigoto respecto al punto medio de los homocigotos (de arriba a abajo, $d = 0$, $d = 0{,}75a$ y $d = 1{,}75a$). Las áreas de cada círculo son proporcionales a la frecuencia de cada genotipo (dentro de cada gráfico). Figura de elaboración propia sobre idea en Michael Lynch y Walsh (1998).

En el caso del valor de cría, vimos previamente en la ecuación (8.63) que el valor medio de la población era igual a cero. Veamos ahora qué ocurre con el desvío de dominancia promedio. Nuevamente, multiplicando los valores de desvío de cada genotipo por las frecuencias correspondientes y sumando tenemos

$$\bar{\delta} = p^2(-2q^2d) + 2pq(2pqd) + q^2(-2p^2)d = 2p^2q^2d(-1+2-1) = 2p^2q^2d \times 0 = 0 \tag{8.108}$$

por lo que también el valor promedio del desvío de dominancia en nuestra población será cero. Esto no nos debería sorprender ya que, por construcción, es decir, por propiedades del ajuste por mínimos cuadrados el desvío promedio de la recta es cero. Finalmente, un punto importante a entender es que también por construcción, bajo apareamientos al azar (es decir, equilibrio de Hardy-Weinberg) los valores reproductivos son ortogonales a los

valores de desvío dominante. Dicho de otra forma, no existe correlación entre unos valores y los otros. Para demostrar hacemos la suma de los productos cruzados, multiplicado cada uno por su frecuencia correspondiente, es decir

$$p^2[(2q\alpha)(-2q^2d)] + 2pq[(q-p)\alpha(2pqd)] + q^2[(-2p\alpha)(-2p^2d)] = \tag{8.109}$$

$$= 4p^2q^2d\alpha[-q + (q-p) + p] = 4p^2q^2d\alpha[0] = 0 \tag{8.110}$$

o, dicho de otra forma, no existe correlación entre los valores de cría de los genotipos y los desvíos de dominancia, que es lo que queríamos demostrar.

Ejemplo 8.5

Asumiendo que los genotipos se encuentran en equilibrio de Hardy-Weinberg, calcular el desvío de dominancia para los tres genotipos del gen de la miostatina, para la característica área del ojo del bife, con los datos del estudio de Bennett et al. (2019) que aparecen en el ejemplo 8.1 y recordando que la frecuencia del alelo L era $p = 0{,}207 \Rightarrow q = 1 - p = 0{,}793$.

Usando los resultados del ejemplo 8.1, tenemos que los genotipos en la escala transformada eran

$$A_{LL} = G_{LL} - m = (103 - 96)\ \text{cm}^2 = 7\ \text{cm}^2 = a \tag{8.111}$$

$$A_{FL} = G_{FL} - m = (93 - 96)\ \text{cm}^2 = -3\ \text{cm}^2 = d \tag{8.112}$$

$$A_{FF} = G_{FF} - m = (89 - 96)\ \text{cm}^2 = -7\ \text{cm}^2 = -a \tag{8.113}$$

Ahora, a partir de la ecuación (8.107), tenemos

$$\delta_{LL} = -2q^2d = -2 \times 0{,}793^2(-3\ \text{cm}^2) = 3{,}773\ \text{cm}^2 \tag{8.114}$$

$$\delta_{FL} = 2pqd = 2 \times 0{,}207 \times 0{,}793(-3\ \text{cm}^2) = -0{,}985\ \text{cm}^2 \tag{8.115}$$

$$\delta_{FF} = -2p^2d = -2 \times 0{,}207^2(-3\ \text{cm}^2) = 0{,}257\ \text{cm}^2 \tag{8.116}$$

Claramente, como la frecuencia del genotipo LL es menor que la del FF, entonces la incidencia en la regresión (valores de cría) es menor para LL y, por lo tanto, su desvío de dominancia debe ser mucho mayor.

PARA RECORDAR

- Podemos modelar la interacción entre alelos del mismo locus como un desvío de los valores genotípicos respecto a los valores de cría (o valores aditivos). Por lo tanto si $d = 0$, los valores genotípicos serán iguales a los valores reproductivos.
- Los **desvíos de dominancia** correspondientes a los genotipos A_1A_1, A_1A_2 y A_2A_2 serán $-2q^2d$, $2pqd$ y $-2p^2d$ respectivamente.
- Al igual que los valores reproductivos, los desvíos de dominancia también dependen de las frecuencias de los alelos, así como de los valores de a y d, por lo que son específicos de cada población.

- Por construcción y bajo apareamiento al azar, no existe correlación entre los valores de cría de los genotipos y los desvíos de dominancia.

8.7 Interacción genotipo-ambiente

Como discutimos en la propuesta original del modelo genético básico, consideramos el fenotipo de un individuo como la resultante de los aportes de su genética y del ambiente en que expresa la característica. Sin embargo, vimos que considerar exclusivamente estas dos fuentes de variabilidad a partir de la suma de los aportes de cada una podía ser extremadamente limitado para describir las observaciones fenotípicas. De hecho, el mismo genotipo suele tener un comportamiento fuertemente dependiente del ambiente en que se expresa y los distintos genotipos son diferentes también en cómo responden a estos cambios de ambiente.

La figura 8.10 muestra un ejemplo de falta de interacción y de una importante interacción **G×E** en ganado lechero.

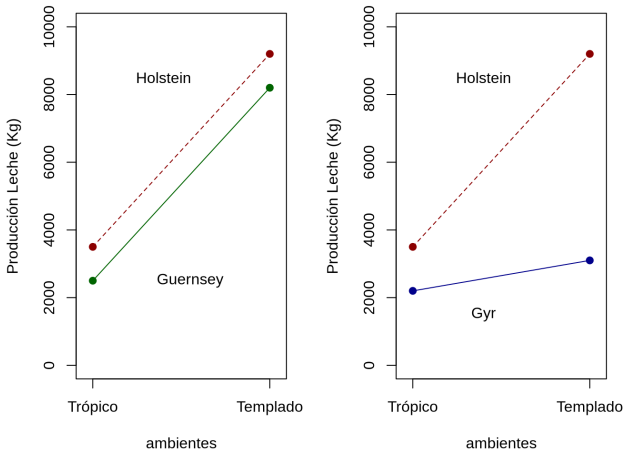


Figura 8.10. Interacción G×E en ganado lechero. En la gráfica de la izquierda, dos razas típicamente lecheras (guernsey y holstein) expresan su producción en el trópico y en un clima templado y si bien en ambas la producción cambia enormemente con el ambiente, las distancias entre razas se mantienen, es decir, las líneas son paralelas y **no existe** interacción **G×E**. En la gráfica de la derecha, ahora comparando holstein contra una raza cebuina (gyr), si bien ambas razas experimentan un decrecimiento en la producción en el trópico respecto a condiciones ambientales más moderadas, mientras que en holstein el decrecimiento es enorme, en gyr el cambio es sumamente menor.

Las razas lecheras tradicionales, como holstein, jersey, guernsey, etcétera, producirán leche en grandes cantidades si cuentan con comida *ad libitum* en climas templados y mejor aún en climas fríos. En la gráfica de la izquierda vemos la comparación de dos razas lecheras en dos ambientes contrastantes, el trópico y un clima templado. Claramente, ambas razas sufren del estrés térmico asociado a la producción en el trópico y su producción es sustancialmente menor que en clima templado, pero las diferencias entre razas en los dos ambientes se

mantienen y, por lo tanto, las líneas son paralelas, por lo que no existe interacción genotipo-ambiente. Por otro lado, la raza cebuina gyr, aún con comida *ad libitum* producirá bastante menos leche aún en óptimas condiciones climáticas ya que posee menor capacidad genética para producirla. En la gráfica de la derecha vemos la comparación de gyr y holstein en los mismos dos ambientes contrastantes. Cuando comparamos la producción de ambas razas en el trópico, donde la temperatura puede ser agobiante, la enorme diferencia que podemos observar en climas templados se ve completamente reducida en el trópico. Claramente, ambas razas producen menos en el trópico que en climas templados, pero mientras que en el trópico la producción de la vaca holstein caerá enormemente, el animal gyr continuará produciendo leche en forma razonable en proporción a su máxima capacidad genética.

En general, los animales (al menos los de interés productivo) tienen cierta capacidad de modular estos efectos «eligiendo» el ambiente, en la medida de sus posibilidades. Ninguna vaca holando se va a quedar a pleno rayo de sol, en un día típico de verano, al mediodía, en Salto, pudiendo resguardarse en el montecito que tiene el potrero. Las plantas, por el contrario, tienen limitada capacidad de desplazamiento (por no decir incapacidad), por lo que en general sufrirán todas las consecuencias del ambiente y, por lo tanto, el componente de interacción genotipo-ambiente se verá en toda su expresión. Dada la complejidad del tratamiento de este tema, le dedicaremos todo el capítulo final («Normas de reacción e interacción genotipo-ambiente»).

8.8 La varianza en el modelo genético básico

Hasta el momento hemos trabajado en este capítulo considerando lo que le ocurre a cada individuo en particular y hemos visto cómo podemos modelar su fenotipo, es decir, la observación o medición directa de una característica en el individuo. Sin embargo, como también hemos discutido en forma abundante, los individuos pertenecen a poblaciones y los distintos fenotipos de sus integrantes están sujetos a distintos factores estocásticos, tanto en lo que hace a la genética de los mismos, como a los ambientes. Con esto en mente, resulta muy claro que no alcanza con las medias de cada uno de los componentes para describir la importancia relativa de las distintas fuentes de variabilidad y debemos recurrir a otros descriptores de estas.

En general, con distribuciones que **se comportan en forma razonable**, la varianza (el segundo momento centrado de una distribución) es un buen indicador de la dispersión y variabilidad de las observaciones. Más aún, para conocer la importancia relativa de las distintas fuentes de variabilidad y su importancia estadística, por ejemplo, la genética en relación con la ambiental, la forma usual de hacerlo es comparando las varianzas, una aproximación que introdujo Ronald Fisher con su **análisis de varianza**. Si recordamos nuestro modelo genético básico, en una de sus formas más simplificadas establecía que el fenotipo de un individuo era la suma de los efectos del genotipo y del ambiente en el mismo, más un componente referido a la interacción genotipo ambiente, es decir $P = G + E + I_{G\times E}$. Como cada individuo de nuestra población podemos considerarlo como una extracción al azar de los factores genéticos y ambientales, entonces los fenotipos serán un también una extracción al azar de una distribución aleatoria.

Por lo tanto, aplicando la regla para el cálculo de la varianza de una suma de variables aleatorias, tenemos que

$$V_P = \mathbb{V}(G + E + I_{G\times E}) = V_G + V_E + V_{G\times E} + 2Cov_{GE} \qquad (8.117)$$

Es decir, la varianza fenotípica (la varianza de lo que observamos) es la suma de la varianza genética, la varianza ambiental, la varianza de la interacción genotipo-ambiente ($V_{G\times E}$), más dos veces la covarianza **entre** genotipo y ambiente. La varianza de interacción es la correspondiente a ese término del modelo genético básico, como lo discutimos previamente y parece ser bastante ubicua en la biología de animales y plantas al menos. Ya vimos previamente que cuando el cambio de ambientes implicaba diferentes pendientes en dos genotipos era necesario agregar este término de interacción, que se trata de sumar valores a combinaciones específicas de genotipos y ambientes, por lo que este «número» tendrá una varianza. En el caso de la covarianza entre genotipo y ambiente la misma se refiere al hecho de que muchas veces los «ambientes» no se distribuyen al azar respecto a los genotipos. Un ejemplo usual de esto es cuando a determinados animales de «elite» se le asignan mejores recursos (alimentación, alojamiento, tratamiento, etcétera) que al resto de los animales. En general no es posible desentrañar esta covarianza por lo que de existir puede afectar las estimaciones de ambos componentes, genéticos y ambientales.

Como vimos en la introducción del modelo genético básico, la componente genética del mismo (G) puede descomponerse en los aportes de los efectos aditivos, los efectos de dominancia y los efectos epistáticos, es decir $G = A + D + I$, por lo que sustituyendo en la ecuación (8.118), tenemos que la varianza fenotípica se puede descomponer de la siguiente manera

$$V_P = \mathbb{V}(G + E + I_{G\times E}) = (V_A + V_D + V_I) + V_E + V_{G\times E} + 2Cov_{GE} \qquad (8.118)$$

que es una de las maneras de descomponerla que más usaremos en el curso (tanto la varianza de efectos epistáticos, como la varianza ambiental se pueden descomponer aún más, pero el uso de estas divisiones será muy pautado).

Componentes genéticos de la varianza: un locus con dos alelos

En el caso de la varianza genética, cuando reducimos nuestro modelo al caso de un locus con dos alelos, podemos continuar el análisis que realizamos previamente para deducir el valor reproductivo y los desvíos de dominancia. Comencemos entonces por la varianza aditiva. Como los valores reproductivos ya están medidos como desvíos de la media de la población, entonces la varianza es solamente la suma de los productos de las frecuencias por el cuadrado del valor reproductivo de cada genotipo.

Si organizamos la información de los valores reproductivos y las frecuencias de cada genotipo en un cuadro, así como los cuadrados de los valores reproductivos y el producto de las frecuencias por los cuadrados, tenemos la siguiente información:

Genotipo	Frecuencia	Valor reproductivo	Cuadrado	Producto
$\mathbf{A_1A_1}$	p^2	$2q\alpha$	$4q^2\alpha^2$	$4p^2q^2\alpha^2$
$\mathbf{A_1A_2}$	$2pq$	$(q-p)\alpha$	$(1-4pq)\alpha^2$	$(2pq-8p^2q^2)\alpha^2$
$\mathbf{A_2A_2}$	q^2	$-2p\alpha$	$4p^2\alpha^2$	$4p^2q^2\alpha^2$
Suma				$2pq\alpha^2$

Los cuadrados de los valores reproductivos de los genotipos A_1A_1 y A_2A_2 son triviales de obtener, mientras que el del valor reproductivo del heterocigota es bastante sencillo si tenemos en cuenta que $p^2+2pq+q^2=1 \Leftrightarrow p^2+q^2=1-2pq$, por lo que

$$[(q-p)\alpha]^2=(q-p)^2\alpha^2=(p^2+q^2-2pq)\alpha^2=(1-2pq-2pq)\alpha^2=(1-4pq)\alpha^2 \tag{8.119}$$

La suma de los valores de la última columna es la varianza aditiva de la característica explicada por nuestro locus. Para obtenerla hacemos

$$V_A=4p^2q^2\alpha^2+(2pq-8p^2q^2)\alpha^2+4p^2q^2\alpha^2=2pq\alpha^2[2pq+1-4pq+2pq]= \\ =2pq\alpha^2(1)\therefore V_A=2pq\alpha^2 \tag{8.120}$$

Para determinar la varianza de dominancia podemos aplicar el mismo procedimiento, ya que también se trata de desvíos de la media de la población. Organizando la información en un cuadro tenemos

Genotipo	Frecuencia	Desvío dominancia	Cuadrado	Producto
A_1A_1	p^2	$-2q^2d$	$4q^4d^2$	$4p^2q^4d^2$
A_1A_2	$2pq$	$2pqd$	$4p^2q^2d^2$	$8p^3q^3d^2$
A_2A_2	q^2	$-2p^2d$	$4p^4d^2$	$4p^4q^2d^2$
Suma				$(2pqd)^2$

Esta vez los cuadrados de los valores de desvíos de dominancia son todos triviales, por lo que, si pasamos a sumar los productos por las correspondientes frecuencias, en la última columna, tenemos

$$V_D=4p^2q^4d^2+8p^3q^3d^2+4p^4q^2d^2=4p^2q^2d^2[q^2+2pq+p^2]=4p^2q^2d^2(1)=4p^2q^2d^2\ \therefore \tag{8.121}$$

$$V_D=(2pqd)^2 \tag{8.122}$$

Si ignoramos la varianza de los efectos epistáticos, cosa que podemos hacer en este caso ya que, al tratarse de una característica explicada por un solo locus, entonces

$$V_G=V_A+V_D \tag{8.123}$$

y, por lo tanto, haciendo uso de las ecuaciones (8.120) y (8.122), poniendo todo junto, tenemos una expresión para la varianza genética en el caso de características gobernadas por un gen de efecto mayor y la misma es

$$V_G = 2pq\alpha^2 + (2pqd)^2 = 2pq[a + d(q - p)]^2 + (2pqd)^2 \qquad (8.124)$$

En la figura 8.11 se ilustra la partición de la varianza genética en sus componentes aditivo y de dominancia, de acuerdo con la frecuencia del alelo A_1 (p) y del desvío del heterocigota respecto del punto medio entre los homocigotas (d). A la izquierda, $d = 0{,}50a$, es decir el heterocigota se encuentra a medio camino entre lo que le correspondería si el locus fuese aditivo y el valor del homocigota A_1A_1. Se aprecia claramente que con ese apartamiento de la aditividad la varianza de dominancia es relativamente pequeña respecto a la aditiva y, por lo tanto, la proporción de la varianza de dominancia en el total de la varianza genética es pequeña. Por otro lado, con $d = 0{,}95a$, es decir que la dominancia de A_1 es casi total, la varianza de dominancia ya no es pequeña y puede llegar a ser una proporción importante del total de varianza genética cuando la frecuencia del alelo A_1 es elevada (notemos que para $p > 0{,}75 \Rightarrow V_D > V_A$).

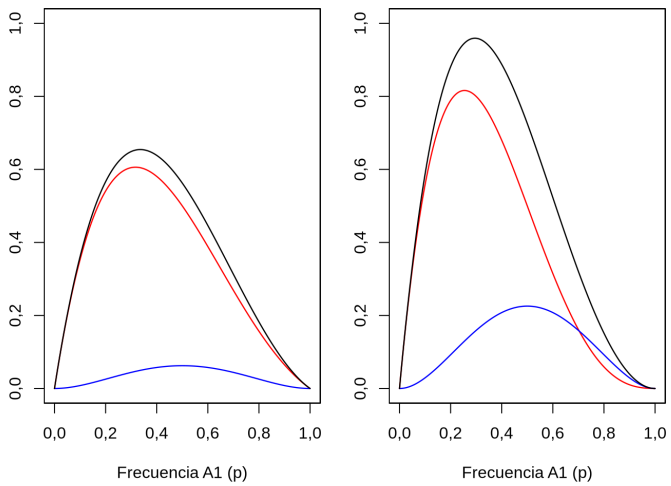


Figura 8.11. Partición de la varianza genética de acuerdo con la frecuencia del alelo A_1 (p) en dos situaciones del desvío de dominancia (desvío del heterocigota, $d = 0{,}50a$ a la izquierda, $d = 0{,}95a$ a la derecha). En rojo se representa la varianza de efectos aditivos, en azul la varianza del desvío de dominancia y en negro la varianza genética total.

Ejemplo 8.6

Calcula la varianza genético-aditiva, de dominancia y varianza genética total si el fenotipo de la característica área del ojo del bife en la raza limousin (figura 8.12) estuviese genéticamente determinada exclusivamente por los alelos L y F del marcador $F94L$ del gen de la miostatina, con frecuencias y valores genotípicos como en el Ejercicio 8.1. Discute la importancia relativa de la varianza aditiva y de la de dominancia.

Figura 8.12. Toro de la raza limousin, originaria de la región con dicho nombre en el centro de Francia, con vacas de la raza pastando detrás (de Wikipedia, autor: Velela, figura de dominio público).

De acuerdo con los datos del Ejercicio 8.1, $a = 7$ cm^2 y $d = -3$ cm^2, mientras que $p = 0{,}207$. Utilizando las ecuaciones (8.120), (8.122) y (8.123), tenemos que

$$V_A = 2pq[a + d(q - p)]^2 = 2 \times 0{,}207 \times 0{,}793 \times [7 - 3(0{,}586)]^2 = 9{,}0213 \text{ cm}^4 \quad (8.125)$$

$$V_D = (2pqd)^2 = [2 \times 0{,}207 \times 0{,}793 \times -3]^2 = 0{,}9700 \text{ cm}^4 \quad (8.126)$$

$$V_G = V_A + V_D = (9{,}0213 + 0{,}9700) \text{ cm}^2 = 9{,}9913 \text{ cm}^4 \quad (8.127)$$

La proporción de la varianza genética total debida la varianza aditiva es $V_{A\,\%} = \frac{9{,}0213}{9{,}9913} = 0{,}9029 = 90{,}29\,\%$ y por lo $V_{D\,\%} = 100\,\% - 90{,}29\,\% = 9{,}71\,\%$.

PARA RECORDAR

- Podemos descomponer el modelo genético básico en términos de varianza de la siguiente manera: $V_P = V(G + E + I_{G\times E}) = (V_A + V_D + V_I) + V_E + V_{G\times E} + 2Cov_{GE}$
- En el caso de características gobernadas por un gen de efecto mayor podemos expresar la varianza genética como $V_G = 2pq\alpha^2 + (2pqd)^2 = 2pq[a + d(q - p)]^2 + (2pqd)^2$

La varianza ambiental

La varianza ambiental es, por definición, la varianza de todos los efectos que no sean genéticos. De otra forma, en el «ambiente» englobamos todo aquello que no podemos modelar como genético. Sin duda se trata de una definición de ambiente apenas operativa, es decir, por conveniencia y que nada tiene que ver con las definiciones de ambiente más aceptadas en otras disciplinas científicas. Por ejemplo, como usualmente no podemos incorporar explícitamente los errores de medición (ya que no los conocemos, porque suelen ser aleatorios), esto pasa a formar parte de la variación no-genética, y nuestra única posibilidad de acomodarla en el modelo genético básico es que forme parte de los «efectos del ambiente».

Esta forma de considerar la partición de efectos y varianzas, si entendemos el significado e interpretación de la misma, no tiene más consecuencias que el sobreuso de un término para referirnos de manera genérica a todo lo que no es genético, que es el foco de nuestro curso. De hecho, nada nos impide escribir todos los términos de efectos y varianzas ambientales que se nos ocurra o nos guste en nuestro modelo sin que ninguna de las conclusiones se vea alterada, en la medida en que los mismos no afecten de manera novedosa lo genético. Por ejemplo, si una característica se ve afectada por la temperatura y yo la registro de forma consistente, nada me impide incluirla de forma explícita en mis modelos.

En otras partes del curso procederemos a descomponer la varianza ambiental de diferentes formas, de acuerdo con las necesidades o conveniencias de cada tema o enfoque. Por ejemplo, en medidas que se repiten en la vida de un organismo es posible descomponer la varianza ambiental en temporal y permanente, la primera referida a los cambios que atañen a una sola observación, la segunda a aquellos que afectan todas las mediciones del organismo. Es decir,

$$V_E = V_{Ep} + V_{Et} \tag{8.128}$$

Por ejemplo, un árbol frutal puede ver afectado su crecimiento en edad temprana y esto afectará todos sus futuros registros productivos, por lo que, con nuestro criterio operativo, se tratará de un **efecto ambiental permanente**. Además, cada año el árbol experimentará variaciones ambientales que cambiarán su producción de año en año y, por lo tanto, se trata de **efectos ambientales temporales**. Lo mismo puede ocurrir con el crecimiento de un cordero que es gestado como mellizo y que es criado por su madre también como mellizo, lo que afectará definitivamente su crecimiento; por lo tanto, este sería un **efecto ambiental permanente**. Ambiental porque no es genético, aunque la preponderancia de animales melliceros tenga una base genética, la misma no hace al crecimiento del cordero. Luego, este efecto «ambiental» afectará los registros productivos de lana todos los años, más allá de que otros efectos ambientales temporales también los afecten (y por eso los registros de cada año son diferentes).

En otros casos consideraremos la partición de la varianza ambiental en componentes que nos ayuden con el tratamiento estadístico de los datos a los fines específicos. Por ejemplo, en diseños experimentales con grupos de parientes con una estructura determinada, un tema que veremos en el capítulo «Parentesco y semejanza entre parientes», aparecen ciertos componentes de covariación que pueden atribuirse a efectos ambientales compartidos. Por ejemplo, en un diseño experimental de hermanos enteros en especies multíparas (cerdos, perros, etcétera), los animales de la misma camada comparten un ambiente que será mucho más «parecido» que el ambiente compartido entre animales de distintas camadas. Por eso,

a este efecto ambiental se le llama del **ambiente común** (E_c) y la varianza la representamos como V_{Ec}.

En general, no necesitamos asumir una distribución particular para la varianza ambiental y la misma se debería ajustar a la distribución de los residuos luego de ajustados los efectos principales del modelo, ya que hemos asumido que «ambiental» implica todo lo no-genético. A pesar de esto, para características continuas o casi-continuas, con distribución simétrica y densidad (probabilidad) importante en el centro de la distribución se suele asumir una distribución normal con media cero.

La varianza de la interacción

Como vimos antes, los efectos epistáticos son aquellos productos de la interacción entre dos o más loci. Como hasta ahora hemos ido construyendo el modelo genético capa por capa, es decir, primero los efectos aditivos explicando el máximo posible de la varianza (con la restricción de aditividad), sobre estos estimamos los efectos de dominancia como desvíos del comportamiento aditivo dentro de cada locus y luego recién estamos por estimar los efectos epistáticos, los mismos serán desvíos, tanto del comportamiento aditivo dentro de cada locus como de la suma de los mismos y de los efectos de dominancia correspondientes a los genotipos en los diferentes loci. Claramente, podemos a su vez ir construyendo la estimación de estos efectos usando un diferente número de loci, empezando por las interacciones entre todos los pares de loci, pasando luego a interacciones entre tríos de loci y así sucesivamente. A cada capa que vamos agregando quedará menos varianza por explicar, por lo que no suelen considerarse las interacciones de más de a dos loci por vez.

Por otro lado, dentro las posibles interacciones entre dos loci podemos distinguir tres tipos: 1) interacciones entre los efectos aditivos en cada loci, 2) las interacciones entre efectos aditivos en un locus y los efectos de dominancia en el otro locus y 3) las interacciones entre efectos de dominancia en ambos loci. Poniendo todo esto junto y asumiendo que despreciamos los efectos de orden 3 y mayores, tenemos que la varianza de los efectos epistáticos puede escribirse como

$$V_I = V_{AA} + V_{AD} + V_{DD} + ... \approx V_{AA} + V_{AD} + V_{DD} \tag{8.129}$$

En general, resulta bastante complejo de estimar por separado los distintos componentes de la varianza epistática, que ya de por sí es complicada de estimar en la mayor parte de los diseños experimentales tradicionales, al menos en el caso del modelo infinitesimal. Cuando en el próximo capítulo desarrollemos la covarianza entre parientes, veremos que algunas relaciones de parentesco incluyen estos efectos y que en teoría podemos llegar a estimar sus efectos con determinadas combinaciones de estructuras de parentesco.

PARA RECORDAR

En general, utilizamos el término **varianza ambiental** para describir de manera genérica todo aquello que no esté englobado en lo genético.

- Podemos descomponer la varianza ambiental en permanente (V_{Ep}) y temporal (V_{Et}), dependiendo si los fenómenos afectan al organismo en un registro o si lo afectarán durante toda su vida.

Respecto a los **efectos epistáticos**, podemos distinguir tres tipos de variaciones entre dos loci: interacciones entre los efectos aditivos en cada loci, interacciones entre efectos aditivos en un locus y los efectos de dominancia en el otro locus, e interacciones entre efectos de dominancia en ambos loci.

8.9 Conclusión

La relación entre el genotipo de un individuo y su fenotipo no es lineal. Podemos pensar en el caso de organismos de reproducción clonal para pensar esto: aun cuando todos posean el mismo genotipo (si despreciamos los efectos mutacionales), resulta claro que existirán variaciones fenotípicas entre los mismos. El estudio de la variación fenotípica y la capacidad de transmisión de caracteres de interés de una generación a otra se vuelve entonces complejo, ya que muchos factores tendrían que ser tomados en cuenta; aspirar a adoptar una aproximación de modelado explícita (como hicimos hasta ahora en este libro, teniendo en cuenta las frecuencias alélicas en la población, desequilibrio de ligamiento, tasa mutacional, etcétera) se vuelve inviable. ¿Cómo podemos aproximarnos a esta problemática? En este capítulo introducimos una forma de plantear este problema que utilizaremos de ahora en más: realizaremos una aproximación estadística, observando las desviaciones respecto a la media de generación a generación, así como la cantidad de variación existente. Para ello introdujimos un modelo (modelo genético básico) que tiene en cuenta las principales clases de variación que deben ser contempladas cuando se plantea el estudio de la herencia de caracteres fenotípicos complejos. En los siguientes capítulos de esta parte del libro nos volcaremos a emplear este modelo para explorar qué tanto se heredan estos caracteres de generación a generación, y en qué medida podemos hacer un estudio sistemático respecto a la influencia que podemos tener sobre los mismos (es decir, ejercer un proceso de *selección artificial* de caracteres de interés, objetivo primario del mejoramiento genético).

8.10 Actividades

8.10.1 Control de lectura

1. Una de las formas más explícitas de expresar el fenotipo de un individuo nombradas en este capítulo es la ecuación $P = G + E + I_{G\times E} = A + D + I + E + I_{G\times E}$. ¿A qué se refiere cada uno de sus términos?
2. ¿De qué manera se modifica el modelo descrito en el punto anterior si queremos evaluar sus componentes a nivel poblacional?
3. ¿Qué es el efecto medio de un alelo? ¿Qué relación tiene con el efecto de sustitución?
4. ¿Qué es el valor de cría de un individuo? ¿Cómo se vincula este concepto con el de efecto medio de un alelo?
5. ¿De qué manera afecta al fenotipo el ambiente?

8.10.2 ¿Verdadero o falso?

1. En un organismo diploide, se dan interacciones entre los alelos de un locus. Esto genera en el fenotipo una desviación que es conocida como efecto de epistasis (término acuñado de la genética clásica).
2. El cambio en la media genotípica de una población donde las frecuencias alélicas varían se verá afectada por la existencia de dominancia.
3. Una población de ranas se encuentra compuesta por individuos RR, Rr y rr. Cada uno de estos genotipos corresponden a un locus que codifica el largo de la cola y los valores de largo asociados son $RR = 0{,}2$ mm, $Rr = 0{,}4$ mm y $rr = 0{,}1$ mm por lo que la característica presenta sobredominancia.
4. Cada gen hará un aporte particular al fenotipo, dependiendo de lo que codifique (por ejemplo, alguna proteína o ARN con una función particular), lo cual se ve reflejado en lo que se denomina efecto medio. Esta medida es, por lo tanto, característica y constante para cada gen en un organismo dado.
5. El valor de cría de un individuo es dependiente de la población en la cual se encuentre.

Soluciones

1. **Falso.** El apartamiento en el fenotipo de un organismo heterocigota para un locus dado se denomina desvío de dominancia (concepto similar al parámetro h visto en capítulos de la segunda sección de este libro). El efecto de epistasis refiere al apartamiento del fenotipo respecto a lo esperado para el conjunto de genes, y está dada por la interacción que pueden tener estos productos génicos (entre sí o por su co-ocurrencia, además de las consecuencias de dichas interacciones).
2. **Verdadero.** Como se muestra en la figura 8.6, el cambio en la media genotípica de una población depende de las frecuencias alélicas, pero la forma en la cual se da el cambio depende a su vez del desvío de dominancia (d).
3. **Verdadero.** El valor intermedio entre los genotipos homocigotas es de $\frac{0{,}2+0{,}1}{0{,}2} = \frac{0{,}3}{0{,}2} = 1{,}5$ mm. Notemos que los individuos heterocigotas poseen un largo de cola

mayor (0,25 mm). En efecto, poseen largos de cola mayor a cualquiera de los heterocigotas; esto indica un modo de acción génica de sobredominancia (como vimos en el capítulo «Selección natural»).

4. **Falso.** El efecto medio de un alelo es dependiente de la población en la que se encuentra, ya que depende de su frecuencia en la población (ver subsección «Efecto medio»). Por lo tanto, este valor no puede ser característico de un alelo (si lo fuera, no variaría).

5. **Verdadero.** El valor de cría (aditivo) es la suma de los efectos medios de los alelos presentes en cada uno de los loci del individuo que aportan a una característica dada. Por lo tanto, al ser el efecto medio dependiente de las frecuencias alélicas en la población, el valor de cría también será dependiente de estas.

8.10.3 Ejercicios

Ejercicio 8.1

Un locus bialélico que afecta a un carácter poligénico de interés se encuentra segregando en una población (alelos A_1 y A_2). Mediante herramientas de secuenciación se logra estimar sus frecuencias alélicas.

a. La media poblacional (en escala transformada) para la medida de interés en una población con frecuencias $p = 0{,}56$ y $q = 0{,}44$ es de 0,59. ¿Cuál es el modo de acción génica para el locus considerado? Ten en cuenta que los genotipos homocigotas tienen un valor genotípico estimado de magnitud $|0{,}42|$

b. ¿Cómo cambiaría el valor de cría para los genotipos en esa población si la frecuencia del alelo A_1 aumentara un 20 %? Se asume que el valor de d no varía.

Solución

a. La media poblacional en escala transformada corresponde a $M = a(p - q) + 2pqd$. El valor genotípico de los homocigotas es a su vez equivalente en magnitud a a, por lo que podemos obtener d operando algebraicamente

$$M = a(p - q) + 2pqd \Rightarrow \tag{8.130}$$

$$2pqd = M - a(p - q) \tag{8.131}$$

$$d = \frac{M - a(p - q)}{2pq} \tag{8.132}$$

$$d = \frac{0{,}59 - 0{,}42 * (0{,}56 - 0{,}44)}{2 \times 0{,}56 \times 0{,}44} \tag{8.133}$$

$$d \approx 1{,}095 \tag{8.134}$$

El valor obtenido implica que el modo de acción génica es la sobredominancia $-a < a < d$

b. El valor de cría de los posibles genotipos se calcula según

$$\begin{cases} \text{EBV}(A_1A_1) = 2q\alpha \\ \text{EBV}(A_1A_2) = (q-p)\alpha \\ \text{EBV}(A_2A_2) = -2p\alpha \end{cases} \tag{8.135}$$

donde $\alpha = a + d(q-p)$

En el primer momento (antes del cambio de frecuencias alélicas) tenemos

$$\begin{cases} \text{EBV}(A_1A_1) = 2q\alpha = 2 \times 0{,}44 \times [0{,}42 + 1{,}095 \times (0{,}44 - 0{,}56)] \approx 0{,}254 \\ \text{EBV}(A_1A_2) = (q-p)\alpha = (0{,}44 - 0{,}56) \times [0{,}42 + 1{,}095 \times (0{,}44 - 0{,}56)] \approx -0{,}0346 \\ \text{EBV}(A_2A_2) = -2p\alpha = -2 \times 0{,}56 \times [0{,}42 + 1{,}095 \times (0{,}44 - 0{,}56)] \approx -0{,}323 \end{cases} \tag{8.136}$$

Notemos que p aumenta a $p = 0{,}56 + 0{,}20 \times 0{,}56 = 0{,}672$, por lo que $q = 1 - p \Rightarrow q = 1 - 0{,}672 = 0{,}328$. Con estos datos (y asumiendo que d no varía) podemos calcular los valores estimados de valor de cría para los genotipos nuevamente

$$\begin{cases} \text{EBV}(A_1A_1) = 2q\alpha = 2 \times 0{,}328 \times [0{,}42 + 1{,}095 \times (0{,}328 - 0{,}672)] \approx 0{,}0284 \\ \text{EBV}(A_1A_2) = (q-p)\alpha = (0{,}328 - 0{,}672) \times [0{,}42 + 1{,}095 \times (0{,}328 - 0{,}672)] = -0{,}0149 \\ \text{EBV}(A_2A_2) = -2p\alpha = -2 \times 0{,}672 \times [0{,}42 + 1{,}095 \times (0{,}328 - 0{,}672)] \approx -0{,}0582 \end{cases} \tag{8.137}$$

Ejercicio 8.2

El dingo (*Canis dingo*) es un canino salvaje australiano que, pese a sus características predatorias contra animales domésticos, ha sido clasificado por la UICN (Unión Internacional para la Conservación de la Naturaleza) como vulnerable (es decir, presenta una alta probabilidad de pasar a ser una especie en peligro de extinción). Los dingos, al igual que los lobos pero a diferencia de los perros, se reproducen una única vez por año, obteniendo camadas de 5 cachorros en promedio según reporta la literatura. Una estrategia de conservación para aumentar el número de animales de esta especie consiste en aumentar el número de cachorros por camada. Con este fin, una investigación ha descubierto que existe un solo locus con dos alelos que controla esta característica (D y d). A su vez, se ha determinado que una población de referencia en el sur de Australia (que se encuentra en equilibrio de Hardy-Weinberg) tiene los siguientes números de dingos hembra por genotipo: $DD = 18$, $Dd = 24$ y $dd = 8$. Luego de un monitoreo extensivo, detectaron que el genotipo DD tenía camadas de 8 cachorros, el genotipo Dd de 6 y dd de apenas 2 cachorros por camada.

a. Bajo este esquema de apareamiento, ¿cuántos cachorros nacen en promedio por año?
b. El equipo de conservación de la especie propone introducir solo machos DD a la población y evitar que otros machos se apareen con las hembras. Si fueran exitosos, ¿cuántos cachorros nacerían?, ¿cuál es el efecto del alelo D?
c. Si por el contrario ahora las hembras se aparearan únicamente con machos dd, ¿cuántos cachorros nacerían?, ¿cuál es el efecto del alelo d?

Solución

a. Para saber en promedio cuantos cachorros nacen en esta población basta con hacer la media ponderada de cuantos cachorros nacen por madre

$$\frac{18 \times 8 + 24 \times 6 + 8 \times 2}{18 + 24 + 8} = 6{,}08 \tag{8.138}$$

Lo mismo obtendremos si utilizamos los valores de los genotipos en la escala transformada y luego les sumamos m tal que

$$m = \frac{8 + 2}{2} = 5 \tag{8.139}$$

$$a = 8 - 5 = 3 \tag{8.140}$$

$$-a = 2 - 5 = -3 \tag{8.141}$$

$$d = 6 - 5 = 1 \tag{8.142}$$

Y sabiendo que $p = \frac{18+12}{50} = 0{,}6$ por lo tanto $q = 1 - p = 0{,}4$, entonces $M = a(p - q) + 2pqd = 3(0{,}6 - 0{,}4) + 2 \times 0{,}6 \times 0{,}4 \times 1 = 1{,}08$, por lo que la media en la escala original sería $1{,}08 + 5 = 6{,}08$.

b. En el paso anterior determinamos que $p = 0{,}6$ y, por lo tanto, $q = 0{,}4$ (es decir, la frecuencia de $D = 0{,}6$ y la de $d = 0{,}4$).

Podemos entonces generar un cuadro que nos ayude a visualizar el apareamiento de esta población con los machos DD.

		Frec. nuevos machos
		$D = 1$
Frec. población	$D = 0{,}6$	$DD = 0{,}6$
	$d = 0{,}4$	$Dd = 0{,}4$

Como se puede observar, al introducir machos homocigotos DD, la nueva generación no contará con individuos dd. A su vez, la nueva media poblacional será de $0{,}6 \times 8 + 0{,}4 \times 6 = 7{,}2$ cachorros por camada. Por lo tanto, el efecto medio del alelo D será de $7{,}2 - 6{,}08 = 1{,}12$, ya que corresponde al efecto producido por la introducción del alelo a la población.

c. Al igual que en el caso anterior podemos determinar las nuevas frecuencias a partir de una tabla:

		Frec. nuevos machos
		$d = 1$
Frec. población	$D = 0{,}6$	$Dd = 0{,}6$
	$d = 0{,}4$	$dd = 0{,}4$

La nueva media de esta población sería entonces $0{,}6 \times 6 + 0{,}4 \times 2 = 4{,}4$ y el efecto medio del alelo d sería $4{,}4 - 6{,}08 = -1{,}68$.

Ejercicio 8.3

a. Comprueba los resultados de efecto medio de los alelos del ejercicio anterior usando el concepto de efecto de sustitución.

b. Calcula el valor de cría de un macho con genotipo Dd.

Solución

a. En principio, podemos calcular el efecto de sustitución (α) según la ecuación (8.4)
$\alpha = a + d(q - p) = 3 + 1(0{,}4 - 0{,}6) = 2{,}8$

A partir del mismo, podemos entonces calcular el efecto medio de cada gen tal que

el efecto medio del alelo D que llamaremos α_1 será $\alpha_1 = q\alpha = 0{,}4 \times 2{,}8 = 1{,}12$, mientras que el efecto medio de d que llamaremos α_2 será $\alpha_2 = -p\alpha = -0{,}6 \times 2{,}8 = -1{,}68$ comprobando entonces los resultados obtenidos previamente.

b. El valor de cría será igual a la suma de los efectos medios de los alelos que componen el genotipo, tal que

$$VC_{Dd} = \alpha_1 + \alpha_2 = 1{,}12 - 1{,}68 = -0{,}56 \tag{8.143}$$

Ejercicio 8.4

a. En una población salvaje de perros salchicha, hace 20 años se determinó que existe un locus con dos alelos (con valores $F = 10$ cm y $f = 4$ cm), que determina el largo de la cola de estos animales. Se ha determinado a su vez que el valor del fenotipo Ff en escala transformada es 0, y al genotipar esta población salvaje se determina que 17 perros poseen genotipo FF, 102 Ff y 153 ff. Se asume que esta característica no se ve afectada por el ambiente.

Responde:

1. ¿Cuánto mide la cola de los perros de cada genotipo presente en la población?
2. ¿Cuál es el promedio de largo de cola para los perros de esta población?
3. Basados en esta información, ¿qué efectos genéticos afectan esta característica?

b. Este año, un nuevo estudio descubrió que existe otro locus con efecto en la característica. Este locus también posee dos alelos, sus valores son $M = 2$ cm y $m = -3$ cm, y se ha determinado que el valor del fenotipo Mm en escala transformada es de 3 cm. No existe interacción entre los alelos de los loci que influyen en el largo de la cola. Debido a este nuevo descubrimiento, se realizó un nuevo genotipado de los individuos de la población (descubriendo que se había visto muy reducida debido al aumento de las aves rapaces) y se encontró la siguiente distribución de individuos por genotipo:

	FF	Ff	ff
MM	2	6	6
Mm	8	9	8
mm	5	5	1

1. ¿Cuánto mide la cola de los perros de cada genotipo presente en la población sabiendo que la característica está controlada por los dos loci?
2. ¿Cuál es el promedio de largo de cola para estos perros sabiendo que la característica está controlada por los dos loci?
3. Basados en esta nueva información, ¿se modificaron los efectos genéticos que afectan esta característica?

Solución

a.

1. En primer lugar se asume que la característica no está afectada por el ambiente, lo que determinará que solo los componentes genéticos afectan la misma. Dado que la característica se encuentra determinada por los alelos de un solo locus, también podemos descartar la influencia de la epistasis en su expresión. A su vez, al transformar los valores se determina que el genotipo Ff posee valor 0 cm; esto implica que el valor del mismo coincide con el valor promedio de los genotipos homocigotas (es decir, no existe dominancia de un alelo sobre el otro). Esto nos determina que el largo de la cola es una característica de carácter totalmente aditivo, por lo que basta sumar el valor o «aporte» de cada uno de los alelos para determinar el valor del fenotipo.

Por lo tanto, tenemos que

$$FF = 10 + 10 = 20 \text{ cm} \tag{8.144}$$

$$Ff = 10 + 4 = 14 \text{ cm} \tag{8.145}$$

$$ff = 4 + 4 = 8 \text{ cm} \tag{8.146}$$

2. En este caso podemos desarrollar dos alternativas: realizar el cálculo de la media ponderada de forma explícita, o recurrir a la ecuación

$$M = a(p - q) + 2pqd \tag{8.147}$$

que se encuentra expresada en términos de la escala transformada por lo que se le debería sumar el valor de m para obtener la media en la escala original. Comenzaremos por esta última aproximación.

En primer lugar podemos comprobar que esta población se encuentra en equilibrio Hardy-Weinberg

$$\text{Frec F } (p) = \frac{17 + 102/2}{17 + 102 + 153} = 0{,}25 \tag{8.148}$$

$$\text{Frec f } (q) = 1 - p = 1 - 0{,}25 = 0{,}75 \tag{8.149}$$

$$\text{Frec FF } = p^2 = 0{,}0625 \Rightarrow n_{esperado} = 0{,}0625 \times 272 = 17 \text{ individuos} \tag{8.150}$$

$$\text{Frec Ff } = 2pq = 0{,}375 \Rightarrow n_{esperado} = 0{,}375 \times 272 = 102 \text{ individuos} \tag{8.151}$$

$$\text{Frec ff } = q^2 = 0{,}5625 \Rightarrow n_{esperado} = 0{,}5625 \times 272 = 153 \text{ individuos} \tag{8.152}$$

Corroboramos por tanto que nos encontramos en equilibrio de Hardy-Weinberg.

Expresamos los genotipos como desvíos del punto m

$$m = \frac{20 + 8}{2} = 14 \text{ cm} \tag{8.153}$$

Resultado que podíamos intuir debido a que coincide con el valor del genotipo heterocigota dada a la ausencia de dominancia.

Pasamos a calcular a

$$a = 20 - 14 = 6 \text{ cm} \tag{8.154}$$

$$-a = 8 - 14 = -6 \text{ cm} \tag{8.155}$$

Por lo que podemos aplicar la ecuación 8.13 para expresar la media en función del punto m

$$M = a(p - q) + 2pqd \tag{8.156}$$

$$M = 6(0{,}25 - 0{,}75) + 2 \times 0{,}25 \times 0{,}75 \times 0 \tag{8.157}$$

$$M = -3 \text{ cm} \tag{8.158}$$

Para determinar la media en escala original basta con sumarle el valor del punto m, por lo que el largo promedio de cola en la población es de $14 - 3 = 11$ cm.

La segunda alternativa consiste en simplemente multiplicar la frecuencia de cada genotipo por su valor y luego sumarlas:

$$\text{frec (FF)} = 17/272 = 0{,}0625 \tag{8.159}$$

$$\text{frec (Ff)} = 102/272 = 0{,}375 \tag{8.160}$$

$$\text{frec (ff)} = 153/272 = 0{,}5625 \tag{8.161}$$

$$0{,}0625 \times 20 + 0{,}375 * 14 + 0{,}5625 \times 8 = 11 \text{ cm} \tag{8.162}$$

3. Como discutimos en la parte *a)* del ejercicio, al tratarse de una característica determinada por los alelos de un solo locus podemos descartar la existencia de epistasis (interacción entre alelos de diferentes loci). A su vez, como el valor del genotipo heterocigota coincide con el promedio de los homocigotas, podemos determinar que no hay efecto de dominancia. Por lo tanto, el único efecto que afecta esta característica es el aditivo.

b.

1. Al ahora existir dos loci que determinan la característica existirán más genotipos posibles y mayor número de interacciones. En primer lugar, podemos descartar los efectos epistáticos debido a que no existe interacción entre los alelos de los loci. Posteriormente podemos hallar el valor aditivo de cada genotipo calculando la suma del valor de los alelos que lo componen (como en el caso anterior). Por último, deberemos tener en cuenta el efecto de dominancia que existe entre los alelos de este nuevo locus, por lo que a los genotipos que presenten heterocigosis Mm deberemos sumarles el valor de la dominancia (+3 cm).

En la siguiente tabla se resume el largo de patas por genotipo calculado con estas consideraciones:

	FF	Ff	ff
MM	24	18	12
Mm	22	16	10
mm	14	8	2

2. Podemos calcular el largo de patas promedio ponderando la frecuencia de cada genotipo por su valor.

Tenemos así

$$\text{frec(FFMM)} \times 24 + \text{frec(FFMm)} \times 22 + +\text{frec(ffmm)} \times 2 = 14{,}8 \text{ cm} \tag{8.163}$$

3. Si, los efectos genéticos que afectan a la característica se modificaron. Ahora, además de tener en cuenta los efectos aditivos debemos considerar la dominancia que presenta M sobre m en el nuevo locus.

Ejercicio 8.5

Campbell, McGrath, and Paaby (2018)

Un grupo de investigación se dedica a estudiar la capacidad de resistencia de una cepa de *S. cerevisiae* frente a diferentes productos de la destilación (una característica poligénica de alta relevancia para la industria). De particular interés es la capacidad de metabolización del ácido acético, por lo cual el grupo trama un experimento para medir la tasa de metabolización en dos cepas para las que se sabe las frecuencias alélicas en algunos locus.

Es importante resaltar que, en las condiciones experimentales usadas por el grupo, la población de *S. cerevisiae* se compone únicamente de células haploides que se reproducen de forma asexual.

a. Mediante una serie de experimentos se obtiene una población compuesta por individuos donde se han fijado alelos de alto rendimiento en los loci codificantes para proteínas de la vía metabólica (en las condiciones especificadas anteriormente) la varianza fenotípica es de $V_P \approx 1{,}2 \times 10^{-3}\ g \times hora^{-1}$ (lo que se estima es la cantidad de masa de ácido acético metabolizada por hora). Teniendo en cuenta que la varianza asociada al error de medición se estima en $V_E\ approx 1{,}0 \times 10^{-4}\ g \times seg^{-1}$, ¿qué componentes genéticos aportan a la variabilidad de la característica fenotípica en la población? **Nota:** por simplicidad, no tengamos en cuenta los efectos de interacción genotipo-ambiente.

b. Teniendo en cuenta que el ciclo de vida de la levadura (*Saccharomyces cerevisiae*) alterna entre dos fases (una haploide y la otra diploide), discute si se puede utilizar el mismo abordaje en cualquier condición de estudio

Solución

a. Notemos que en organismos haploides, por definición, no puede existir la varianza de efecto de dominancia, en tanto cada organismo hereda un solo alelo para cada locus. A su

vez, los investigadores trabajan sobre una población donde se han fijado diferentes alelos involucrados directamente en la vía metabólica que produce el fenotipo; esto implica que no existirá varianza aditiva. Esto implica que la varianza fenotípica, según el modelo básico, queda reducida a

$$V_P = V_G + V_E + V_{G\times E} = (V_I) + V_E + V_{G\times E} \tag{8.164}$$

Lo que es más, si no tenemos en cuenta los efectos de interacción genotipo-ambiente, tenemos que

$$V_P = V_I + V_E \tag{8.165}$$

Es decir, los únicos factores que aportan a la varianza fenotípica son la varianza de interacción (epistasis) y la varianza asociada al ambiente (la cual, en una visión simplista, equivale al error de medición).

Tenemos así que, en este caso

$$1{,}2 \times 10^{-3}\ g \times hora^{-1} = V_I + 1{,}0 \times 10^{-4}\ g \times seg^{-1} \tag{8.166}$$

de lo que se deduce que la varianza asociada a la epistasis es de

$$\begin{aligned} V_I &\approx (1{,}2 \times 10^{-3} - 1{,}0 \times 10^{-4})\ g \times hora^{-1} \\ V_I &\approx 1{,}1 \times 10^{-4}\ g \times hora^{-1} \end{aligned} \tag{8.167}$$

Se puede estimar así el componente de varianza asociado a la epistasis, que es el fenómeno que más contribuye a la variación fenotípica observada en el modelo.

b. Notemos que durante las fases diploides ya no podremos descartar los efectos de desvío de dominancia atribuibles a la interacción intralocus en las cepas estudiadas. A su vez, si bien en principio no existirá varianza aditiva (ya que en los loci que se consideran hay alelos fijados), a medida que se generen variantes por efecto de la mutación este componente genético debe también ser tenido en cuenta. Teniendo en cuenta lo anterior, notamos que según el modelo genético básico

$$V_P = V_G + V_E + V_{G\times E} = (V_A + V_D + V_I) + V_E + V_{G\times E} \tag{8.168}$$

Aun desestimando la interacción genotipo-ambiente, se tienen tres componentes que aportan a la varianza sin posibilidad de discernir el aporte relativo (es decir, V_A, V_D y V_I).

Por lo tanto, el abordaje experimental realizado no es directamente extrapolable a las condiciones en las que *S. cerevisiae* alterne entre una fase haploide y una diploide.

Capítulo 9

Parentesco y semejanza entre parientes

I have an interesting case. I'm treating two sets of Siamese twins with split personalities.
I'm getting paid by eight people.[1]

Leonard Zelig

En nuestro modelo genético básico identificamos factores genéticos y ambientales (así como sus interacciones) que son, de acuerdo con el modelo, los causantes de la variación fenotípica que observamos. Desafortunadamente, en una gran parte de los casos estamos limitados a observar solamente el fenotipo de los individuos y pese a esto debemos poder estimar los diferentes componentes de la varianza del modelo, tanto para su utilización directa como para calibrar las relaciones entre las distintas fuentes de variación. Afortunadamente, la genética de la mayor parte de las especies de interés económico posee unas reglas de transmisión de la información que son bastante sencillas (al menos a grandes rasgos), lo cual nos ha ido permitiendo construir modelos para representar la expresión de las características en los fenotipos. Son, por lo tanto, estas mismas reglas las que gobiernan las similitudes que observamos entre los organismos que comparten una parte de su información genética, algo que todos reconocemos como parentesco.

En este capítulo se discutirá el concepto de parentesco en términos generales, antes de presentar el concepto de **parentesco aditivo**, estrechamente relacionado a conceptos que vimos previamente en el capítulo «Apareamientos no-aleatorios». Veremos cómo el parentesco aditivo se relaciona con nuestro modelo genético básico, así como diferentes formas de calcularlo y de representar las relaciones entre individuos. Discutiremos al mismo tiempo el concepto de **consanguinidad** y su relación con el parentesco aditivo. Luego pasaremos a discutir otro concepto relacionado con nuestro modelo genético básico que es el de **parentesco de dominancia** y veremos en qué relaciones de parentesco este es distinto de cero.

Durante la construcción del modelo genético básico en el capítulo anterior vimos que la descomposición de la varianza fenotípica en sus distintas fuentes (tanto genéticas como ambientales, y su relativa importancia) se constituían en un elemento clave para entender las posibilidades de trabajar sobre la genética de una característica. Sin embargo, dado que la única varianza observable es la fenotípica, hasta ahora no hemos visto ni soslayado ninguna manera razonable de estimar los distintos componentes. El elemento clave para poder estimar los distintos componentes genéticos de la varianza es la semejanza que existe entre parientes, semejanza que se debe a la genética y las condiciones que comparten. Por lo tanto, se desarrollará un modelo causal para estas semejanzas, para luego vincularlo con el modelo observacional y de esta forma alcanzar nuestro objetivo de estimar los componentes

[1] Leonard Zelig es el personaje central de la película *Zelig*, 1983, de Woody Allen. La traducción podría ser algo así como «Tengo un caso interesante. Estoy tratando dos parejas de hermanos siameses con múltiples personalidades. Me pagan ocho personas».

de varianza. Finalmente, discutiremos las diferencias entre los conceptos de parentesco estadístico (todos los que veremos hasta el final) y parentesco genómico, el derivado de conocer la información genómica de los individuos.

OBJETIVOS DEL CAPÍTULO

☐ Discutir el concepto de parentesco en términos generales.

☐ Ver cómo el parentesco aditivo se relaciona con nuestro modelo genético básico.

☐ Analizar formas de calcular el parentesco aditivo y de representar las relaciones entre individuos.

☐ Discutir el concepto de **consanguinidad** y su relación con el parentesco aditivo.

☐ Presentar el concepto de **parentesco de dominancia** y analizar en qué relaciones de parentesco se observa el mismo.

9.1 Parentesco

Todos tenemos una idea bastante desarrollada de lo que significa el término *parentesco*, en especial cuando nos referimos a nuestra especie. En este caso, de acuerdo con la acepción más común, se trata de un vínculo, tanto por «sangre» como por relaciones de afinidad estable entre personas. Esto pone sobre el tapete, al menos en lo que refiere al término *parentesco*, una distinción clara entre las construcciones sociales y lo que vamos a desarrollar como concepto genético de parentesco. Por ejemplo, a nadie se le ocurriría pensar que nuestro cónyuge no es nuestro pariente, aunque desde el punto de vista genético posiblemente tengamos lo mismo o menos en común que con cualquier miembro al azar de la población. En todo caso, pese a la distinción, lo que queda claro es que cuando hablamos de parentesco nos referimos a alguna clase de similitud o afinidad especial entre miembros de un grupo, mayor que con el resto de la población no-emparentada.

En nuestro caso, estamos interesados exclusivamente en el parentesco genético entre los individuos, ya que el mismo será un indicador de la constitución genética compartida entre los mismos. Esto, a su vez, nos permitirá determinar el «mérito» genético de los animales, predecir el mismo para sus parientes, así como estimar los parámetros genéticos más relevantes para determinar la base genética de una característica y las estrategias para mejorarla. Como vimos previamente, en nuestro modelo genético básico los componentes genéticos son varios (aditivos, de dominancia, epistáticos) y, por lo tanto, el parentesco debería reflejar estos distintos aportes a la base genética común entre dos individuos. Como veremos más adelante, la partición en componentes aditivos, de dominancia y epistáticos para la parte genética del fenotipo nos llevará a la necesidad de definir coeficientes de parentesco específicos a cada componente. Es decir, entre dos individuos tendremos un valor para el

parentesco aditivo, otro valor para el parentesco de dominancia y otro para el parentesco epistático.

La pregunta que nos podemos hacer entonces es qué consideramos como parentesco y qué característica tiene que tener el o los índices que lo representen. Claramente, cuando entre dos individuos no exista parentesco de ningún tipo sería ideal que los coeficientes de parentesco fuesen iguales a cero. Por otra parte, desde un punto de vista intuitivo, resulta natural esperar que el parentesco de un individuo consigo mismo sea igual al 100 %, es decir igual a 1. Dicho de otra forma, cuando un individuo comparte con otro (o con él mismo) todo su genoma, entonces esperaríamos que comparta el 100 % de su base genética. Por ejemplo, en el caso de los gemelos monocigóticos, despreciando el impacto de la epigenética (algo cada vez más discutible), los mismos comparten exactamente el mismo genoma y, por lo tanto, esperaríamos que nuestros índices o coeficientes de parentesco lo reflejaran indicándolo con un valor de 1.

Finalmente, hay un punto muy relevante a destacar antes de comenzar nuestra jornada con el tema de parentesco genético. Históricamente, desde el comienzo de la genética hasta hace relativamente poco tiempo, el genoma de los individuos era completamente desconocido e imposible de conocer. No es hasta mediados del siglo XX que se logra dilucidar la estructura del ADN y conocer el sistema de codificación de la información (código genético). Recién hacia el último cuarto del siglo pasado se logra entender la organización de los genes en procariotas y eucariotas, así como desarrollar los primeros métodos de secuenciación genética, notablemente reducido a unas pocas centenas de bases nucleotídicas. Sin embargo, desde el comienzo de la genética de poblaciones y cuantitativa quedaron en claro las bases de la transmisión de la información genética de progenitores a descendientes. En este capítulo nos referiremos, excepto mención específica, a individuos diploides en especies sexuales. Es decir, en estas especies un padre transmite la mitad de su información genética a sus hijos y a partir de esto podemos derivar las expectativas de compartir una fracción de su genoma que tenemos para cada par de individuos para los que conocemos su relación. Esto es, estamos refiriéndonos a una definición de parentesco que es estadística (o probabilística) en su base.

Por ejemplo, como veremos más adelante, el **parentesco aditivo** entre **hermanos enteros** (lo que normalmente llamamos hermanos en poblaciones humanas) es igual a $\frac{1}{2}$. Sin embargo, todos sabemos que dentro de cada grupo de hermanos existen algunos más parecidos entre sí que con el resto. Si bien el «parecido» es apenas una característica exterior que no podemos asociar directamente al genoma, es una clara indicación de que no necesariamente todos los hermanos comparten exactamente la misma fracción del genoma. De hecho, aunque totalmente improbable, teóricamente dos **hermanos enteros** podrían no compartir ningún alelo **IBD** de su genoma, y también con la misma implausibilidad podrían compartir todos, aun no siendo gemelos monocigóticos. Al tratarse el parentesco estadístico de una variable aleatoria, esto introduce una fuente de variabilidad en las estimaciones posteriores. Afortunadamente, desde fines del siglo XX es posible genotipar un gran número de loci en cada individuo, o aún secuenciar el genoma completo de ambos, por lo que la comparación de similaridad genética puede ser derivada directamente. Dicho de otra forma, el parentesco aditivo entre dos hermanos en particular no tendrá la necesidad de ser igual a $\frac{1}{2}$ y ahora será sustituido por la fracción realmente compartida del genoma, tema que veremos al final de este capítulo.

9.2 Parentesco aditivo

En el capítulo anterior discutimos que en nuestro modelo genético podíamos descomponer el aporte de la genética en distintos componentes ($G = A + D + I$), uno de los cuales era el que correspondía a los efectos aditivos, es decir, aquellos que podemos describir sumando los aportes de cada uno de los alelos presentes. Por lo tanto, desde el punto de vista intuitivo, una medida de similaridad entre individuos que refleje este aspecto sería la proporción de alelos idénticos por ascendencia (**IBD**) compartidos por los individuos. En la figura 9.1 podemos ver representadas un par de situaciones que reflejan distintos grados de parentesco entre individuos, así como las posibles configuraciones de los alelos heredados. A la izquierda tenemos la relación que existe entre un progenitor **A** y su descendencia, representada por el individuo **C**. El individuo **A** posee dos alelos, que llamaremos en forma arbitraria A_1 y A_2, por ahora independientemente de si son distintos, idénticos en estado o idénticos por ascendencia. Claramente, el descendiente de **A** (**B**) debe recibir de este uno de sus dos alelos, **sí o sí**. De hecho, como veremos más adelante, la relación entre progenitor y descendiente es la única en la que **la proporción de alelos** compartidos está totalmente determinada (es decir, no es aleatoria). Dicho de otra forma, para cada posición del genoma o loci, un hijo recibirá la mitad de la información de su padre y la mitad de su madre, por lo que compartirá con ellos $\frac{1}{2}$ de los alelos.

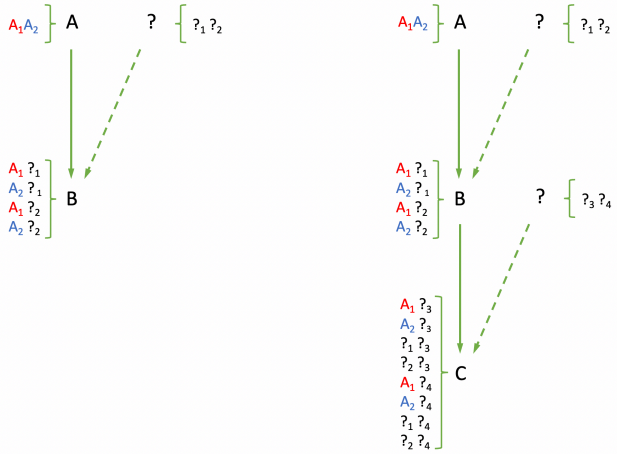


Figura 9.1. Representación de la relación entre alelos en el caso de parentesco aditivo directo. A la izquierda la relación entre padre e hijo, mientras que a la derecha agregamos una nueva generación, lo que nos permite ver la relación entre un individuo (**A**) y su nieto (**C**).

Esto lo podemos ver más explícitamente en el diagrama de la figura. El progenitor **A**, que para un locus determinado tiene los dos alelos A_1 y A_2 se aparea con un progenitor desconocido o irrelevante para nosotros y que por lo tanto marcamos con **?** y a sus alelos como $?_1$ y $?_2$. Al existir dos posibles gametos producidos por el progenitor de nuestro interés (A_1 y A_2) y dos gametos por el otro progenitor ($?_1$ y $?_2$), tenemos entonces **4** posibles genotipos que surgen de combinarlos: $A_1?_1$, $A_2?_1$, $A_1?_2$ y $A_2?_2$. Veamos ahora cuál sería la proporción promedio de alelos compartidos entre un progenitor y su progenie. El progenitor será siempre A_1A_2, pero la descendencia podrá ser de cualquiera de los **4** tipos mencionado previamente. Si ponemos esta información en un cuadro podemos entender mejor por qué

en cualquiera de los **4** posibles genotipos del descendiente la proporción de alelos **IBD** es $\frac{1}{2}$ y, por lo tanto, su promedio es también $\frac{1}{2}$.

	Alelos compartidos con progenitor (A_1A_2)	Proporción
$A_1?_1$	A_1	$1/2$
$A_2?_1$	A_2	$1/2$
$A_1?_2$	A_1	$1/2$
$A_2?_2$	A_2	$1/2$
Promedio		**1/2**

Por ejemplo, si el descendiente tiene el genotipo $A_1?_1$ entonces compartirá con el progenitor **A** el alelo A_1, es decir un alelo de los dos que tiene. En el caso de que sea $A_2?_1$ su genotipo, entonces será el alelo A_2 el que compartirá, pero también será uno de los dos alelos, es decir, una proporción de $\frac{1}{2}$. Como las **4** combinaciones (genotipos) del descendiente todas comparten $\frac{1}{2}$ de los alelos con su progenitor, entonces la proporción promedio será también igual a $\frac{1}{2}$.

Veamos lo que ocurre ahora en el caso de un individuo con su nieto (o nieta), como ocurre en la figura 9.1 a la derecha. En este caso, el progenitor **A** tiene una descendencia **B**, como ya vimos, con los **4** genotipos posibles $A_1?_1$, $A_2?_1$, $A_1?_2$ y $A_2?_2$. Este individuo **B** se apareará con otro individuo desconocido de la población, al que le asignamos por lo tanto los alelos $?_3$ y $?_4$, y el resultado será ahora de **8** genotipos posibles en la descendencia de **B**, es decir **C**. Para conocer el promedio de alelos idénticos por ascendencia entre **A** y **C** (es decir, su parentesco aditivo), los volvemos a colocar en un cuadro donde comparar cada una de las combinaciones posibles.

	Alelos compartidos con progenitor (A_1A_2)	Proporción
$A_1?_3$	A_1	$1/2$
$A_2?_3$	A_2	$1/2$
$?_1?_3$	$-$	$0/2 = 0$
$?_2?_3$	$-$	$0/2 = 0$
$A_1?_4$	A_1	$1/2$
$A_2?_4$	A_2	$1/2$
$?_1?_4$	$-$	$0/2 = 0$
$?_2?_4$	$-$	$0/2 = 0$
Promedio		**4/16=1/4**

De las **8** combinaciones posibles en **C**, es decir $8 \times 2 = 16$ alelos en consideración, hay **4** combinaciones que comparten 1 alelo y otras **4** que no comparten ninguno. En total, $\frac{4\times1+4\times0}{16} = \frac{4}{16} = \frac{1}{4}$ de los alelos son compartidos en la relación abuelo-nieto. Tanto en este caso como en el del padre-hijo, los alelos compartidos son claramente **IBD** y, por lo tanto, estamos en condiciones ahora de introducir una definición formal del parentesco aditivo.

Para dos individuos X e Y, llamamos **parentesco aditivo** entre ellos (que notamos como a_{XY}) a la **proporción esperada de alelos en común**, idénticos por ascendencia entre los genotipos de X e Y. De otra forma, como vimos en el capítulo «Apareamientos no-aleatorios» el parentesco aditivo es igual a el doble de la probabilidad de que al extraer al azar un alelo en cada individuo, estos sean **IBD**, es decir $f_{XY} = \frac{1}{2}a_{XY} \therefore a_{XY} = 2f_{XY}$. La relación entre ambas descripciones se debe a que cuando lo miramos desde la proporción en los genotipos hacemos 2 comparaciones en lugar de una.

Una de las cosas a remarcar que observamos en el parentesco **abuelo-nieto** y que no se observaba en el **padre-hijo** es el carácter aleatorio del número de alelos idénticos, de acuerdo con el genotipo del nieto (en nuestro caso). Es decir, ya no todas las combinaciones tienen la misma proporción de coincidencias como en padre-hijo, que eran $\frac{1}{2}$. Ahora, en abuelo-nieto tenemos 4 genotipos que coinciden en un alelo de dos ($\frac{1}{2}$), pero otros 4 genotipos que no tienen coincidencia, lo que resalta el hecho de que este enfoque de parentesco es puramente estadístico. Cuando el número de descendientes tiende a infinito, entonces observaremos la misma proporción de cada uno de los 8 genotipos y, por lo tanto, el promedio tenderá al valor teórico, $\frac{1}{4}$ en nuestro caso. El otro aspecto relacionado que podemos notar es la razón por la que requerimos la condición de idéntico por ascendencia, algo que ya hemos tratado en abundancia en el capítulo «Apareamientos no-aleatorios». Dado que en general asumimos que existe un número limitado de formas alternativas de secuencias (alelos) en la población, la presencia de dos secuencias idénticas en dos individuos no necesariamente habla de parentesco, al menos en el intervalo de tiempo relevante para nuestra población. Si recordamos del capítulo «Apareamientos no-aleatorios», a esas secuencias le llamábamos idénticas en estado y no eran informativas de parentesco. Por lo tanto, para que la identidad de secuencias sea informativa de parentesco debemos exigir que las mismas correspondan a **IBD**.

Hasta ahora, los dos casos que hemos examinado corresponden a un tipo de relacionamiento entre individuos que conocemos como parentesco directo, donde se puede trazar un camino de una sola dirección del flujo de información genética. Es decir, la información genética va de padres a hijos, pero no al revés. Estos hijos la transmitirán, a su vez, a sus propios hijos, por lo que el flujo sigue la dirección de **abuelo** $\Rightarrow$ **padre** $\Rightarrow$ **hijo**. Una situación diferente ocurre cuando queremos analizar el parentesco aditivo entre medios hermanos (comparten solo un progenitor) o hermanos enteros (comparten los dos progenitores). En cualquiera de estos dos casos, representados en la figura 9.2, la información fluye desde los ancestros (progenitores) a los descendientes, pero la información **fluye en paralelo** por dos ramas distintas, la que lleva a un hermano (o medio) y la que lleva al otro hermano (o medio).

Para analizar el parentesco aditivo entre **medios hermanos**, podemos hacer un cuadro con todas las combinaciones de genotipos en cada uno de los medios hermanos. Como los mismos tienen un progenitor en común (por ejemplo, el padre suele ser lo usual en especies domésticas donde un macho puede aparearse con varias hembras en el mismo período reproductivo) pero el otro es diferente, elegimos nuevamente marcar los progenitores que no pueden aportar alelos **IBD** con el signo de interrogación y sus alelos con subíndices correspondientes.

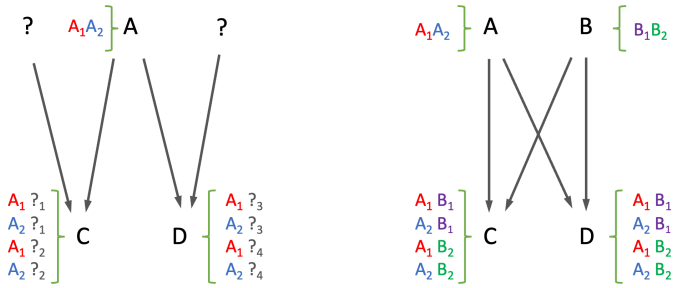


Figura 9.2. Representación de la relación entre alelos en el caso de parentesco aditivo colateral. A la izquierda la relación entre medios hermanos (MH), mientras que a la derecha observamos la relación de alelos entre hermanos enteros (HE).

De acuerdo con esto, si colocamos los genotipos posibles (de los individuos **C** y **D** en la figura 9.2 izquierda) en filas y columnas y contamos el número de coincidencias, tenemos

	$A_1?_1$	$A_2?_1$	$A_1?_2$	$A_2?_2$
$A_1?_3$	$A_1 = 1$	0	$A_1 = 1$	0
$A_2?_3$	0	$A_2 = 1$	0	$A_2 = 1$
$A_1?_4$	$A_1 = 1$	0	$A_1 = 1$	0
$A_2?_4$	0	$A_2 = 1$	0	$A_2 = 1$
Suma	2	2	2	2
Promedio	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{2}{8}$

Esto nos deja con 4 promedios, todos iguales a $\frac{2}{8} = \frac{1}{4}$, que debemos a su vez promediar, lo que al ser todos iguales nos lleva a que su promedio sea igual a $\frac{1}{4}$. Otra forma de verlo es a partir de la suma de todas las coincidencias, que son $4 \times 2 = 8$, en un total de $4 \times 4 \times 2 = 32$ (4 genotipos en cada medio hermano y con dos alelos cada uno). Por lo tanto, el **parentesco aditivo entre medios hermanos** será $\frac{8}{32} = \frac{1}{4}$, idéntico resultado al anterior.

En el caso de **hermanos enteros** (los individuos **C** y **D** en la figura 9.2 derecha), los hermanos comparten a los dos padres y, por lo tanto, las posibles configuraciones de los alelos serán iguales para ambos individuos. Esto nos deja con el siguiente cuadro

	A_1B_1	A_2B_1	A_1B_2	A_2B_2
A_1B_1	A_1,$B_1 = 2$	$B_1 = 1$	$A_1 = 1$	0
A_2B_1	$B_1 = 1$	A_2,$B_1 = 2$	0	$A_2 = 1$
A_1B_2	$A_1 = 1$	0	A_1,$B_2 = 2$	$B_2 = 1$
A_2B_2	0	$A_2 = 1$	$B_2 = 1$	A_2,$B_2 = 2$
Suma	4	2	2	2
Promedio	$\frac{4}{8}$	$\frac{4}{8}$	$\frac{4}{8}$	$\frac{4}{8}$

El promedio de cada columna (o fila) es de $\frac{4}{8} = \frac{1}{2}$, por lo que el promedio general será también igual a $\frac{1}{2}$ y, por lo tanto, **el parentesco aditivo entre hermanos enteros será** de $\frac{1}{2}$.

Parentesco directo y colateral. Llamamos **parentesco directo** al que se da entre un individuo y su descendencia, sin importar el número de generaciones entre ellos (por ejemplo, hijos, nietos, bisnietos, tataranietos, choznos, bichoznos, etcétera). Llamamos **parentesco colateral** al que no es directo, es decir, que se da entre individuos no directamente vinculados entre sí por una relación de ascendencia (por ejemplo, entre medios hermanos, entre hermanos, tío-sobrino, etcétera).

Consanguinidad

Hasta ahora hemos asumido, en forma implícita, que el ancestro de referencia (padre, abuelo, etcétera) **no era consanguíneo.** Como vimos en el capítulo «Apareamientos no-aleatorios», cuando un individuo es consanguíneo la probabilidad de que los alelos en cualquier loci sean **IBD** se ve incrementada, precisamente en una cantidad igual al coeficiente de consanguinidad F. Como nuestra definición de parentesco aditivo hace referencia directa a la probabilidad de que los alelos en dos individuos sean **IBD**, la consanguinidad del ancestro (y solo esta) incrementará el parentesco aditivo en la misma proporción. Por ejemplo, en el caso de medios hermanos, si el ancestro **A** (el padre, por ejemplo) posee una consanguinidad $F_A = \frac{1}{8}$, el parentesco aditivo entre **C** y **D** será ahora $a_{CD} = \frac{1}{4}(1 + F_A) = \frac{1}{4}(1 + \frac{1}{8}) = \frac{1}{4}(\frac{8}{8} + \frac{1}{8}) = \frac{1}{4}\frac{9}{8} = \frac{9}{32}$. Este proceso afectará la probabilidad de **IBD** para cada uno de los ancestros, por lo que debemos encontrar una forma de aplicar este esquema a situaciones que involucren más de un ancestro.

La **consanguinidad de un individuo** es **igual a la mitad del parentesco aditivo de sus padres.** Si **A** y **B** son los padres de **C**, entonces $F_C = \frac{1}{2}a_{AB}$. Esto refuerza el concepto de que un individuo es consanguíneo sí y solo sí sus padres son parientes (es decir, el parentesco aditivo entre ellos es mayor a 0). Otra forma de verlo, en conexión con lo visto en el capítulo «Apareamientos no-aleatorios» es que $F_C = \frac{1}{2}a_{AB} = f_{AB} = F_C$, es decir, **el coeficiente de consanguinidad de un individuo es igual al coeficiente de coancestría de sus padres.**

Veamos cómo llegamos al cálculo del coeficiente de consanguinidad en el caso de un hijo entre medios hermanos, como se aprecia en la figura 9.3. El coeficiente de consanguinidad de **E** es igual a la probabilidad de que ambos alelos en el mismo sean **IBD**, que al tener solo un progenitor en común (**A**) es igual a la probabilidad de que los dos alelos sean A_1 (es decir A_1A_1 el genotipo de **E**) o los dos sean A_2. Si la consanguinidad de **A** es cero, es decir $F_A = 0$, entonces la probabilidad de que A_1 llegue a **E** por la izquierda es $\frac{1}{2}\frac{1}{2} = \left(\frac{1}{2}\right)^2$ y la que llegue por la derecha también es igual a $\frac{1}{2}\frac{1}{2} = \left(\frac{1}{2}\right)^2$. Esto se debe a que, por la rama izquierda, la probabilidad de que A_1 pase de **A** a **C** es $\frac{1}{2}$ y la probabilidad de que de **C** pase a **E** también es un medio y como los dos eventos deben darse, entonces multiplicamos sus probabilidades. Idéntico razonamiento podemos hacer para la rama derecha, en este caso

pasando por **D** en lugar de **C**. Por lo tanto, para que los dos eventos se den a la vez, es decir llegar con A_1 por la izquierda y por la derecha, ambos deben también darse y, por lo tanto, tenemos que multiplicar las probabilidades de los dos eventos y entonces

$$Pr(E_1 = A_1, E_2 = A_1) = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^4 = \frac{1}{16} \tag{9.1}$$

Exactamente lo mismo podemos también razonar para el alelo A_2, por lo que

$$Pr(E_1 = A_2, E_2 = A_2) = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^4 = \frac{1}{16} \tag{9.2}$$

Como los dos eventos son mutuamente excluyentes podemos sumar las probabilidades de ambos y, por lo tanto,

$$F_E = Pr(E_1 = A_1, E_2 = A_1) + Pr(E_1 = A_2, E_2 = A_2) = \frac{1}{16} + \frac{1}{16} = \frac{1}{8} \tag{9.3}$$

que es igual además, como es de esperar, a la mitad del parentesco entre sus padres, ya que como vimos $a_{CD} = \frac{1}{4} \therefore F_E = \frac{1}{2} a_{CD} = \frac{1}{2}\frac{1}{4} = \frac{1}{8}$. En el caso de que $F_A > 0$, entonces se seguirá cumpliendo $F_E = \frac{1}{2} a_{CD}$, pero ahora a_{CD} será mayor a $\frac{1}{4}$ ya que la probabilidad de que los alelos sean **IDB** debe incrementarse en la proporción F_A.

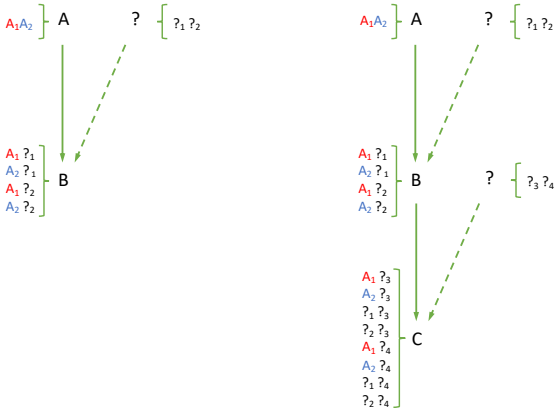


Figura 9.3. Derivación del coeficiente de consanguinidad para el hijo de medios hermanos. El coeficiente de consanguinidad de **E** es igual a la probabilidad de que ambos alelos en el mismo sean **IBD**, que al tener solo un progenitor en común (**A**) es igual a la probabilidad de que los dos alelos sean A_1 (es decir A_1A_1 el genotipo de **E**) o los dos sean A_2. Si la consanguinidad de **A** es cero, es decir $F_A = 0$, entonces la probabilidad de que A_1 llegue a **E** por la izquierda y también por la derecha es igual a $2 \times \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{1}{8}$

PARA RECORDAR

- Para dos individuos X e Y, llamamos **parentesco aditivo** (a_{XY}) entre ellos a la proporción esperada de alelos en común, idénticos por ascendencia (**IDB**) entre los genotipos de X e Y.
- La relación entre progenitor y descendiente es la única en la que la proporción de alelos compartidos está totalmente determinada.
- **Parentesco directo** es aquel que se da entre un individuo y su descendencia, sin importar el número de generaciones entre ellos. **Parentesco colateral** es aquel que no es directo, es decir, que se da entre individuos no directamente vinculados entre sí por una relación de ascendencia.
- La consanguinidad de un individuo es igual a la mitad del parentesco aditivo de sus padres, que es lo mismo que decir que es igual al coeficiente de coancestría de sus padres.

Diagramas de flechas

Una forma sencilla y práctica de calcular el parentesco aditivo a partir de **pedigrís** es considerar que cada pasaje de información de un individuo a su descendencia es de la mitad de su genoma, por lo que los alelos compartidos entre padre-hijo o madre-hijo es siempre de $\frac{1}{2}$. Es decir, cada flecha que une un individuo con su descendencia implica una reducción a la mitad de la información compartida. Veamos cómo esto se aplica en términos prácticos a las relaciones que ya hemos estudiado. Si nos fijamos en la figura 9.1 a la izquierda, tenemos la relación entre un progenitor **A** y su descendencia **B**. Entre ellos hay una sola flecha que marca el flujo de información de **A** hacia **B** y, por lo tanto, podemos representar esto como $A \to B$. Cada pasaje de información (flecha) la misma se reduce a la mitad (porque el individuo diploide sexual tiene dos progenitores y cada uno aporta la mitad del nuevo genoma), por lo que si llamamos n al número de flechas entre los dos individuos a los que les queremos calcular el parentesco aditivo ($n = 1$ en este caso), tenemos ahora

$$a_{AB} = \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^1 = \frac{1}{2} \tag{9.4}$$

Lo anterior presupone que el ancestro (**A**) no es consanguíneo, es decir que $F_A = 0$. Como ya vimos antes, si $F_A \neq 0$, entonces debemos incluir este incremento en la ecuación, la que quedaría

$$a_{AB} = \left(\frac{1}{2}\right)^n (1 + F_A) = \left(\frac{1}{2}\right)^1 (1 + F_A) = \frac{1}{2}(1 + F_A) \tag{9.5}$$

Claramente, la ecuación anterior también se aplica cuando $F_A = 0$, ya que en ese caso $a_{AB} = \left(\frac{1}{2}\right)^n (1 + F_A) = \left(\frac{1}{2}\right)^n (1 + 0) = \left(\frac{1}{2}\right)^n$.

En el caso de un individuo **A** y su nieto **C**, como aparece representado en la figura 9.1 derecha, tenemos ahora que el pasaje de información (genética) se puede representar como $A \to B \to C$, es decir, hay dos flechas en el camino de la información de **A** hacia **C**.

Por lo tanto, teniendo en cuenta que la consanguinidad de **A** es F_A, tenemos ahora

$$a_{AC} = \left(\frac{1}{2}\right)^n (1 + F_A) = \left(\frac{1}{2}\right)^2 (1 + F_A) = \frac{1}{4}(1 + F_A) \tag{9.6}$$

para el caso de la relación abuelo-nieto.

Veamos ahora qué ocurre en el caso de parentesco colateral. En la figura 9.2 izquierda podemos ver la relación entre dos **medios hermanos**, **C** y **D**, que podemos representar de acuerdo con el diagrama de flechas como $C \leftarrow A \rightarrow D$. Notemos que las flechas tienen distintas direcciones ahora, ya que la información fluye desde **A**, tanto hacia **C** como hacia **D**. Dicho de otra forma, en el parentesco colateral hay (al menos) un punto (ancestro) donde las flechas cambian de sentido. Sin embargo, algo que se mantiene es que en cada una de las **ramas**, cada transferencia de información (flecha) la misma se reduce a la mitad, por lo que si llamamos n_1 a una de las dos ramas (por ejemplo, la que lleva a **C**) y n_2 a la otra rama (la que lleva a **D**), entonces, teniendo en cuenta la consanguinidad del ancestro **A** podemos escribir el parentesco entre **C** y **D** como

$$a_{CD} = \left(\frac{1}{2}\right)^{n_1+n_2} (1 + F_A) = \left(\frac{1}{2}\right)^{1+1} (1 + F_A) = \left(\frac{1}{2}\right)^2 (1 + F_A) = \frac{1}{4}(1 + F_A) \tag{9.7}$$

La pregunta que sigue es cómo incorporamos la existencia de otros ancestros en común en el caso del parentesco colateral. La situación la vemos representada en el parentesco aditivo entre **hermanos enteros**, como aparece en la figura 9.2 derecha. Ahora ya no tenemos un solo ancestro en común entre **C** y **D**, como teníamos en medios hermanos, sino los dos progenitores de ambos que son los mismos. Claramente tenemos dos **fuentes** de información genética compartida, **A** y **B** y cada una con un **camino** propio. Es decir, tenemos los caminos independientes $C \leftarrow A \rightarrow D$ y $C \leftarrow B \rightarrow D$, ambos aportando alelos que podrían ser **IBD** entre **C** y **D**, por lo que debemos sumar los aportes de ambas trayectorias. Además, como vimos antes, tanto **A** como **B** pueden tener consanguinidades distintas de cero, por lo que debemos considerarlas. En este caso particular nuestro, el parentesco aditivo entre ambos hermanos enteros podemos escribirlo

$$\begin{aligned} a_{CD} = \left(\frac{1}{2}\right)^{n_{1A}+n_{2A}} (1 + F_A) + \left(\frac{1}{2}\right)^{n_{1B}+n_{2B}} (1 + F_B) = \\ = \left(\frac{1}{2}\right)^2 (1 + F_A) + \left(\frac{1}{2}\right)^2 (1 + F_B) \therefore \\ a_{CD} = \frac{1}{4}(2 + F_A + F_B) \end{aligned} \tag{9.8}$$

En el caso particular de que ni **A** ni **B** sean consanguíneos, es decir $F_A = F_B = 0$, entonces el parentesco entre hermanos enteros queda reducido a

$$a_{CD} = \frac{1}{4}(2 + F_A + F_B) = \frac{1}{4}(2 + 0 + 0) = \frac{1}{4} \times 2 = \frac{1}{2} \tag{9.9}$$

Este análisis que realizamos primero para padres-hijos, abuelos-nietos y luego para medios hermanos y hermanos enteros podemos ahora generalizarlo a cualquier tipo de relación de parentesco entre dos individuos **X** e **Y**.

Para ello debemos seguir una breve serie de pasos (es decir, un algoritmo):

1. Transformar el pedigrí en un diagrama de flechas, donde cada individuo aparecerá una sola vez. Del individuo pueden salir tantas flechas como descendientes tenga en consideración, pero podrán entrar solo 0, 1 o 2 flechas (el primer caso si no conocemos ningún progenitor, luego si conocemos 1, o ambos en el último caso). Las flechas irán dirigidas desde el progenitor al descendiente.
2. Identificar a todos los ancestros comunes a **X** e **Y**, es decir, que sean ancestros de ambos. Para cada ancestro, si existe la información necesaria (ancestros de los mismos) calcular la consanguinidad (que es igual a un medio del parentesco aditivo entre sus padres).
3. Establecer todos los caminos independientes que vayan desde un individuo de interés al otro. En el caso de **parentesco directo**, el camino irá de un individuo al otro directamente, con las flechas en la misma dirección. En el caso de **parentesco colateral**, **cada camino** de un individuo de interés al otro pasará por un ancestro en común entre ambos, donde ocurrirá el único cambio de dirección de las flechas. Un individuo ocurre solo una vez en cada camino.
4. En caso de existir varios caminos diferentes, los resultados de los mismos se suman. Solo se tiene en cuenta en cada caso la consanguinidad del ancestro, no la de los individuos intermedios en el camino.
5. Si llamamos A_i a los distintos ancestros en común entre los dos individuos, **X** e **Y**, a calcularles el parentesco aditivo, teniendo en cuenta los pasos anteriores aplicamos luego la fórmula

$$a_{XY} = \sum_{A_i} \left(\frac{1}{2}\right)^{n_{i1}+n_{i2}} (1 + F_{A_i}) \tag{9.10}$$

con n_{i1} y n_{i2} las flechas en cada una de las dos ramas en el caso del parentesco colateral o directamente n_i en el caso de parentesco directo.

Ejemplo 9.1

Un excelente ejemplo, involucrando tanto parentesco directo como parentesco colateral, es el que figura en el *Manual de prácticas del curso de Zootecnia* (Echeverría et al. 2012), página 31 y cuyo diagrama de flechas reproducimos en la figura 9.4.

La idea es calcular el parentesco aditivo entre los individuos **E** y **G**, que son entre ellos padre e hijo respectivamente. De lo que vimos antes sabemos que normalmente un padre transmite a su hijo la mitad de su información genética, por lo que en ausencia de otra información (consanguinidad o relación con el otro progenitor) el parentesco entre padre e hijo será de $\frac{1}{2}$.

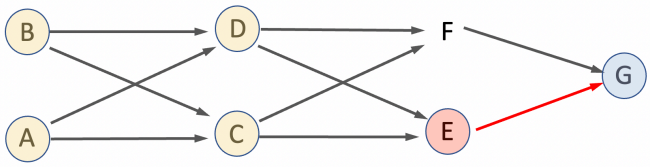


Figura 9.4. Diagrama de flechas con un cruzamiento recurrente entre hermanos. El objetivo es calcular a_{EG}. La flecha roja indica una relación de **parentesco directo** entre los dos individuos a los que les queremos calcular el parentesco aditivo (**E** y **G**), mientras que el círculo rojo indica el ancestro de esa relación. Los círculos amarillos indican ancestros comunes a ambos candidatos.

Sin embargo, solo de observar la figura podemos sospechar que en algunos individuos la consanguinidad no será cero, ya que son hijos de parientes. Por ejemplo, **C** y **D** son hermanos enteros, así como **E** y **F**, por lo que deberemos proceder a realizar los cálculos con cuidado.

El primer paso, de acuerdo con nuestro algoritmo, consiste en transformar el pedigrí en un diagrama de flechas y ya está realizado. El segundo paso consiste en identificar todos los ancestros comunes a ambos individuos. En la figura aparecen marcados con un círculo amarillo todos los ancestros comunes a ambos que forman parte de algún parentesco colateral, mientras que el círculo rojo identifica el ancestro en el único camino de parentesco directo (**E** es padre de **G**). Para calcular la consanguinidad de cada ancestro recurrimos a la ecuación que nos indica que la misma es igual a la mitad del parentesco aditivo de los padres. Como de **A** y **B** no poseemos información, **asumimos que no poseen consanguinidad y que además no son parientes entre ellos**. Esto nos deja con $F_A = F_B = 0$. Además, como supusimos que **A** y **B** no son parientes, entonces sus hijos (**C** y **D**) no serán consanguíneos, por lo que $F_A = F_B = F_C = F_D = 0$. Como el parentesco aditivo entre **C** y **D** es de hermanos enteros sin consanguinidad en sus ancestros, es decir $a_{CD} = \frac{1}{2}$, entonces $F_E = F_F = \frac{1}{2}a_{CD} = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$

El tercer paso consiste en identificar todos los caminos por los ancestros comunes. Estos son

$$\begin{aligned} E &\to G \\ E \leftarrow C &\to F \to G \\ E \leftarrow D &\to F \to G \\ E \leftarrow C \leftarrow A &\to D \to F \to G \\ G \leftarrow F \leftarrow C \leftarrow A &\to D \to E \\ E \leftarrow C \leftarrow B &\to D \to F \to G \\ G \leftarrow F \leftarrow C \leftarrow B &\to D \to E \end{aligned} \tag{9.11}$$

Contando el número de flechas en cada camino y aplicando la ecuación (9.10) tenemos

$$a_{EG} = \left(\frac{1}{2}\right)^{1}(1+F_E) + \left(\frac{1}{2}\right)^{1+2}(1+F_C) + \left(\frac{1}{2}\right)^{1+2}(1+F_D) + \left(\frac{1}{2}\right)^{2+3}(1+F_A) + \\ + \left(\frac{1}{2}\right)^{3+2}(1+F_A) + \left(\frac{1}{2}\right)^{2+3}(1+F_B) + \left(\frac{1}{2}\right)^{3+2}(1+F_B) \tag{9.12}$$

Como $F_E = \frac{1}{4}$ y $F_A = F_B = F_C = F_D = 0$, aplicando los pasos **4** y **5** del algoritmo y sustituyendo estos valores en la ecuación precedente (arriba) tenemos

$$a_{EG} = \left(\frac{1}{2}\right)^{1}(1+\frac{1}{4}) + \left(\frac{1}{2}\right)^{3} + \left(\frac{1}{2}\right)^{3} + \left(\frac{1}{2}\right)^{5} + \left(\frac{1}{2}\right)^{5} + \left(\frac{1}{2}\right)^{5} + \left(\frac{1}{2}\right)^{5} \therefore \tag{9.13}$$

$$a_{EG} = \left(\frac{1}{2}\right)^{1}(\frac{5}{4}) + 2\left(\frac{1}{2}\right)^{3} + 4\left(\frac{1}{2}\right)^{5} = \frac{1}{2}\frac{5}{4} + \left(\frac{1}{2}\right)^{2} + \left(\frac{1}{2}\right)^{3} \therefore \tag{9.14}$$

$$a_{EG} = \frac{5}{8} + \frac{1}{4} + \frac{1}{8} = \frac{5}{8} + \frac{2}{8} + \frac{1}{8} = \frac{5+2+1}{8} = 1 \tag{9.15}$$

Es decir, el parentesco aditivo entre **E** y **G** ya no es típico entre padre e hijo ($\frac{1}{2}$), y ahora es $a_{EG} = 1$, lo que se explica tanto por la consanguinidad de **E** como por la existencia de varios otros ancestros comunes entre padre e hijo (**A**, **B**, **C** y **D**).

PARA RECORDAR

- Podemos calcular el **parentesco aditivo** a partir de pedigrís considerando que cada pasaje de información de un individuo a su descendencia es de la mitad de su genoma.
- Para determinar el parentesco aditivo entre dos individuos X e Y debemos:

1) Transformar el pedigrí en un diagrama de flechas, donde cada individuo aparecerá una sola vez.
2) Identificar a todos los ancestros comunes a **X** e **Y**, es decir, que sean ancestros de ambos.
3) Establecer todos los caminos independientes que vayan desde un individuo de interés al otro.
4) En caso de existir varios caminos diferentes, los resultados de los mismos se suman.
5) Por último, si llamamos A_i a los distintos ancestros en común entre los dos individuos, **X** e **Y**, podemos aplicar la siguiente fórmula:// $a_{XY} = \sum_{A_i}\left(\frac{1}{2}\right)^{n_{i1}+n_{i2}}(1+F_{A_i})$

con n_{i1} y n_{i2} las flechas en cada una de las dos ramas en el caso del parentesco colateral o directamente n_i en el caso de parentesco directo.

El método tabular

Los métodos descritos antes dejan de ser prácticos cuando tenemos decenas, centenas o miles de animales, como es usual en las evaluaciones genética de reproductores o cuando queremos estimar parámetros como la consanguinidad promedio de una población; para estos casos debemos buscar una alternativa que sea escalable. Una alternativa conocida y práctica es lo que se conoce como el **método tabular** (porque se trata de completar una «tabla» o matriz), que se sustenta en un par de principios básicos del parentesco aditivo y que nos permite implementar un algoritmo muy sencillo.

Los dos principios en los que se basa son los siguientes:

1) Si dos animales son parientes, entonces uno o ambos progenitores de uno deben ser parientes del otro individuo. Dicho de otra forma, si un individuo **X** es pariente de un individuo **Y**, que tiene padres **A** y **B**, entonces **X** debe ser pariente de **A** o de **B** (o de ambos). Como se trata de parentesco aditivo y los padres transmiten a sus hijos la mitad del mismo, entonces

$$a_{XY} = \frac{1}{2}(a_{XA} + a_{XB}) \tag{9.16}$$

con **A** y **B** padres de **Y**.

2) Como vimos antes, el coeficiente de consanguinidad de un individuo es igual a la mitad del parentesco aditivo entre sus padres. Es decir, si **A** y **B** son los padres de **Y**, entonces

$$F_Y = \frac{1}{2}a_{AB} \tag{9.17}$$

A estos dos principios del parentesco aditivo debemos agregar la base algorítmica que hace funcionar al **método tabular**: el **método de programación dinámica**. Este método es ampliamente utilizado en informática y se basa en que para completar las celdas de una tabla en la que las mismas siguen alguna dependencia con otras celdas, debo organizar las filas y columnas de tal forma que al recorrer la tabla de forma ordenada siempre cuente con la información que precise para calcular el valor de cada celda que me va quedando.

Veamos a través de un ejemplo concreto y una serie de **pasos** o **reglas** que iremos describiendo cómo funciona el método tabular. El ejemplo que tomaremos será el del diagrama de flechas que aparece en el ejemplo 9.1.

PASO 1: Ordena todos los animales que participan de la genealogía o pedigrí de acuerdo con su fecha de nacimiento, los más viejos al principio, los más jóvenes al final. En caso de no tener las fechas de nacimiento (como en el ejemplo), determine qué animales son los más ancestrales y utilice las flechas para describir «generaciones». En el ejemplo, **A** y **B** son los

más viejos, seguidos de **C** y **D**, seguidos de **E** y **F**, para finalmente tener a **G**. No importa si **A** es más viejo que **B** o viceversa, así como tampoco entre **C** y **D**, o entre **E** y **F**.

PASO 2: En una planilla o tabla (puede ser una hoja de cálculo), coloca **en la primera columna** (a partir de la tercera celda) y en la **segunda fila** (a partir de la segunda celda) los nombres de los animales tal cual fueron ordenados en el **PASO 1**.

PASO 3: Coloca en la primera fila de la tabla, encima de cada animal, sus progenitores, como aparece en la figura 9.5. En caso de no conocer a alguno de los progenitores, coloca un signo de interrogación en su lugar. En nuestro ejemplo, los padres de **A** y **B** son desconocidos, por lo que encima de ambos colocamos **? ?**. En el caso de **C** y **D** sus progenitores son **A** y **B**, por lo que encima de ambos colocamos esta información. Para **E** y **F** sus progenitores son **C D**, mientras que para **G**, sus progenitores serán **E F**.

padres	? ?	? ?	A B	A B	C D	C D	E F
	A	B	C	D	E	F	G
A							
B							
C							
D							
E							
F							
G							

Figura 9.5. Forma de ordenar los animales y sus progenitores para el cálculo de parentescos aditivos y consanguinidades mediante el método tabular para el diagrama de flechas del ejemplo 9.1.

PASO 4: Coloca un 1 en cada celda de la diagonal. El valor de 1 será el parentesco aditivo del animal consigo mismo, excepto si el animal es consanguíneo, en cuyo caso debemos sumarle luego (en un paso siguiente) la consanguinidad (que es igual a la mitad del parentesco entre sus padres).

PASO 5: Calcula el parentesco entre animales a partir del **principio 1** y de la consanguinidad a partir del **principio 2**. Para que el algoritmo funcione debemos trabajar en forma ordenada. Como la matriz debe ser simétrica pues el parentesco de **A** con **B** debe ser igual al del **B** con **A** ($a_{AB} = a_{BA}$), nosotros vamos a completar solo la **matriz triangular superior** y luego, al finalizar todo el proceso, copiaremos los valores a la triangular inferior. Es decir, vamos a ir recorriendo desde la primera fila, la primera celda a la izquierda hasta la última a la derecha, antes de pasar a la segunda fila. Lo mismo para esta, de izquierda a derecha hasta el final, antes de pasar a la tercera fila, etcétera.

Cada fila por completar comienza con la celda de la diagonal correspondiente y utilizando el **principio 2** calculamos la consanguinidad que debemos sumarle al 1 de la celda como la mitad del parentesco entre sus padres. En general, si esa información existe (el parentesco entre los padres del individuo) ya la vamos a tener calculada en alguna celda y bastará con multiplicarla por $\frac{1}{2}$. En nuestro caso, el primer animal es **A** si sus padres son desconocidos, por lo que no vamos a poder saber si son parientes y, por lo tanto, debemos asumir que la consanguinidad de **A** es cero ($F_A = 0$).

La siguiente celda corresponde al parentesco aditivo entre **A** (fila) y **B** (columna). Como no conocemos tampoco los padres de **B**, entonces el **principio 1** debemos aplicarlo como $a_{AB} = \frac{1}{2}(a_{A?} + a_{A?}) = \frac{1}{2}(0 + 0) = 0$, como podemos ver en la figura 9.6.

La próxima celda a calcular es la del parentesco entre **A** (fila) y **C** (columna). En este caso si conocemos a los padres de **C**, que son **A** y **B**, por lo que aplicando el **principio 1**, tenemos que $a_{AC} = \frac{1}{2}(a_{AA} + a_{AB})$. El parentesco de **A** consigo mismo está en la intersección de la primera fila con la diagonal (es decir, la primera celda de esta fila) y vimos que era $a_{AA} = 1$. En el caso de **A** con **B** vimos que era $a_{AB} = 0$, por lo que $a_{AC} = \frac{1}{2}(a_{AA} + a_{AB}) = \frac{1}{2}(1+0) = \frac{1}{2}$. Lo mismo se aplica para el parentesco entre **A** y **D** ya que es hermano entero de **C**.

En el caso del parentesco de **A** con **E** (cuyos padres son **C** y **D**), aplicando nuevamente el **principio 1**, tenemos $a_{AE} = \frac{1}{2}(a_{AC} + a_{AD}) = \frac{1}{2}(\frac{1}{2} + \frac{1}{2}) = \frac{1}{2}(1) = \frac{1}{2}$. Obviamente, lo mismo se aplica para su hermano entero **F**, por lo que $a_{AF} = a_{AE} = \frac{1}{2}$.

Nos queda entonces la última celda de esta fila que es el parentesco aditivo de **A** con **G** (cuyos padres son **E** y **F**). Aplicando el **principio 1**, tenemos que $a_{AG} = \frac{1}{2}(a_{AE} + a_{AF}) = \frac{1}{2}(\frac{1}{2} + \frac{1}{2}) = \frac{1}{2}(1) = \frac{1}{2}$.

padres	? ?	? ?	A B	A B	C D	C D	E F
	A	B	C	D	E	F	G
A	1	0	1/2(1+0)=1/2	1/2(1+0)=1/2	1/2(1/2+1/2)=1/2	1/2(1/2+1/2)=1/2	1/2(1/2+1/2)=1/2
B		1	1/2(0+1)=1/2	1/2(0+1)=1/2	1/2(1/2+1/2)=1/2	1/2(1/2+1/2)=1/2	1/2(1/2+1/2)=1/2
C			1	1/2(1/2+1/2)=1/2	1/2(1+1/2)=3/4	1/2(1+1/2)=3/4	1/2(3/4+3/4)=3/4
D				1	1/2(1/2+1)=3/4	1/2(1/2+1)=3/4	1/2(3/4+3/4)=3/4
E					1+1/2(1/2)=5/4	1/2(3/4+3/4)=3/4	1/2(5/4+3/4)=1
F						1+1/2(1/2)=5/4	1/2(3/4+5/4)=1
G							1+1/2(3/4)=11/8

Figura 9.6. Cálculo de parentescos aditivos y consanguinidades mediante el método tabular para el diagrama de flechas del ejemplo 9.1. Cada celda aparece con el cálculo correspondiente usando los principios 1 y 2.

Podemos pasar ahora a la segunda fila, donde aplicamos el mismo procedimiento que para la primera y no encontraremos ninguna novedad (cosa que es razonable porque las etiquetas de **A** y **B** son intercambiables). Tampoco aparecerán novedades hasta la fila correspondiente a **E**, en la que aparece la primera consanguinidad distinta de cero. Aplicando el **principio 2**, debemos sumarle al 1 del **PASO 4** la consanguinidad del individuo, que es igual a la mitad del parentesco aditivo de sus padres. Como los padres de **E** son **C** y **D**, vamos hasta la celda intersección de la fila correspondiente a **C** con la columna correspondiente a **D** y observamos que su valor es $a_{CD} = \frac{1}{2}$, por lo que $F_E = \frac{1}{2}a_{CD} = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$, valor que sumamos a 1 para obtener $a_{EE} = 1 + \frac{1}{4} = \frac{5}{4}$.

Si seguimos con estos procedimientos iremos completando la matriz triangular superior.

PASO 6: Completa los valores de la triangular inferior con los valores correspondientes de la triangular superior, es decir

$$a_{ji} = a_{ij} \quad \forall j > i \tag{9.18}$$

Esto nos permite llegar al resultado final (Figura 9.7), que es la matriz de parentesco entre los individuos del ejemplo 9.1. Podemos verificar con lo calculado explícitamente en dicho ejercicio que el parentesco aditivo entre **E** y **G** es $a_{EG} = 1$, lo que coincide con lo obtenido por el método tabular.

Como veremos más adelante y en otros cursos más avanzados, la matriz de parentesco aditivo, de tamaños $n \times n$ (con n el número de individuos) y usualmente notada con **A**, juega un

	A	B	C	D	E	F	G
A	1	0	1/2	1/2	1/2	1/2	1/2
B	0	1	1/2	1/2	1/2	1/2	1/2
C	1/2	1/2	1	1/2	3/4	3/4	3/4
D	1/2	1/2	1/2	1	3/4	3/4	3/4
E	1/2	1/2	3/4	3/4	5/4	3/4	1
F	1/2	1/2	3/4	3/4	3/4	5/4	1
G	1/2	1/2	3/4	3/4	1	1	11/8

Figura 9.7. Resultado final de la matriz de parentesco aditivo entre los individuos del ejemplo 9.1, usando el método tabular para calcularla.

papel fundamental en la estimación de los valores de cría de animales en las evaluaciones genéticas. Por ejemplo, tanto en la metodología en uso desde los años setenta del siglo XX, el llamado **BLUP** (Best Linear Unbiased Prediction), como en las metodologías más modernas que hacen uso del parentesco genómico (que discutiremos más adelante), la matriz de parentesco aditivo «estadística» es un elemento clave para dar cuenta de la dependencia entre observaciones y obtener estimados con buenas propiedades estadísticas.

Ejemplo 9.2

Verificar que el resultado del parentesco entre **D** y **F** obtenido de la matriz de parentesco es idéntico al correspondiente del diagrama de flechas del ejemplo 9.1.

De acuerdo con el diagrama de flechas, los ancestros comunes son **D** (parentesco directo), **A** y **B** (parentesco colateral). Los caminos posibles por los ancestros son los siguientes

$$\begin{aligned} D &\rightarrow F \\ D \leftarrow A \rightarrow C &\rightarrow F \\ D \leftarrow B \rightarrow C &\rightarrow F \end{aligned} \tag{9.19}$$

Por lo tanto, podemos escribir el parentesco aditivo entre **D** y **F** como

$$a_{DF} = (\frac{1}{2})^1(1 + F_D) + (\frac{1}{2})^3(1 + F_A) + (\frac{1}{2})^3(1 + F_B) \tag{9.20}$$

Como $F_A = F_B = F_D = 0$ en nuestro caso, los dos primeros porque no tenemos información de los padres ni de ellos mientras que para F_D los padres no son parientes, la ecuación anterior se reduce a

$$a_{DF} = (\frac{1}{2})^1 + (\frac{1}{2})^3 + (\frac{1}{2})^3 = \frac{1}{2} + 2(\frac{1}{2})^3 = \frac{1}{2} + \frac{1}{4} \therefore \\ a_{DF} = \frac{3}{4} \tag{9.21}$$

que es el mismo resultado que aparece en la matriz de parentesco calculada por el método tabular (cuarta fila, sexta columna o al revés).

PARA RECORDAR

- El **método tabular** constituye una alternativa escalable a los diagramas de flechas para calcular el parentesco. El mismo se basa en dos principios:

1) Si un individuo **X** es pariente de un individuo **Y** (que tiene padres **A** y **B**), entonces X debe ser pariente de **A** o de **B** (o de ambos). Como se trata de parentesco aditivo y los padres transmiten a sus hijos la mitad del mismo, entonces $a_{XY} = \frac{1}{2}(a_{XA} + a_{XB})$
2) El **coeficiente de consanguinidad** de un individuo es igual a la mitad del parentesco aditivo entre sus padres. Es decir, si **A** y **B** son los padres de **Y**, entonces $F_Y = \frac{1}{2}a_{AB}$

- Los siguientes pasos constituyen la guía para utilizar el método tabular:

PASO 1: Ordenar todos los animales que participan de la genealogía o pedigrí de acuerdo con su fecha de nacimiento: los más viejos al principio y los más jóvenes al final.
PASO 2: En una planilla o tabla, colocar en la primera columna (a partir de la tercera celda) y en la segunda fila (a partir de la segunda celda) los nombres de los animales tal cual fueron ordenados en el **PASO 1**.
PASO 3: Colocar en la primera fila de la tabla, encima de cada animal, los progenitores del mismo. En caso de no conocer a alguno de los progenitores coloque un signo de interrogación en su lugar.
PASO 4: Colocar un «1» en cada celda de la diagonal. El valor de 1 será el parentesco aditivo del animal consigo mismo, excepto si el animal es consanguíneo, en cuyo caso debemos sumarle luego (en un paso siguiente) la consanguinidad.
PASO 5: Calcular el parentesco entre animales a partir del **principio 1** y de la consanguinidad a partir del **principio 2**.
PASO 6: Completar los valores de la triangular inferior con los valores correspondientes de la triangular superior.

9.3 Parentesco de dominancia

En nuestra descomposición de los efectos genéticos del modelo genético básico teníamos que $G = A + D + I$. Para representar la relación de los efectos aditivos entre individuos construimos el **parentesco aditivo**, que definimos como la **proporción de alelos IBD** en la comparación entre los genotipos de dos individuos. Los efectos de dominancia los habíamos manejado como las desviaciones de la aditividad intralocus y por lo tanto deben corresponderse a los **genotipos compartidos** de alelos **IBD**, en lugar de a los alelos compartidos entre esos individuos.

Si recapitulamos las relaciones entre individuos que aparecen en la figura 9.1, vemos que en ninguno de los casos, padre-hijo y abuelo-nieto, podemos compartir el genotipo entre dos individuos ya que los otros progenitores son desconocidos y aparentemente sin relación. Es decir, si nos fijamos en la relación padre-hijo, tenemos que el genotipo del padre es A_1A_2 y el del hijo puede ser una cualquiera de las combinaciones $A_1?_1$, $A_1?_2$, $A_2?_1$, $A_2?_2$. Claramente, ninguna de estas es A_1A_2, por lo que el parentesco de dominancia entre **A** y **B** será $d_{AB} = \frac{0}{4} = 0$. Lo mismo para la relación abuelo-nieto, donde para el abuelo A_1A_2 tenemos ahora 8 posibles combinaciones en el nieto, que serán $A_1?_3$, $A_1?_4$, $A_2?_3$, $A_2?_4$, $?_1?_3$, $?_1?_4$, $?_2?_3$, $?_2?_4$. Como ninguna de ellas es igual a A_1A_2, entonces $d_{AC} = \frac{0}{8} = 0$.

Obviamente, lo mismo se puede decir de cualquier otro parentesco directo donde los apareamientos provengan de individuos siempre diferentes y no emparentados entre ellos.

La situación en la figura 9.2 izquierda representa la relación entre medios hermanos. En este caso, uno de los medios hermanos pertenece a alguna de las combinaciones $A_1?_1$, $A_1?_2$, $A_2?_1$, $A_2?_2$, mientras que el otro pertenece a las combinaciones $A_1?_3$, $A_1?_4$, $A_2?_3$, $A_2?_4$. Como en el primer hermano uno de los alelos es siempre $?_1$ o $?_2$ y en el otro hermano siempre uno de los alelos es $?_3$ o $?_4$, entonces nunca coincidirán plenamente los genotipos, entre las 16 combinaciones posibles de 4×4. Por lo tanto, también en este caso $d_{BC} = \frac{0}{16} = 0$

La situación va a cambiar para el caso de **hermanos enteros**, como vemos en la figura 9.8. Tenemos ahora también 16 pares de genotipos posibles, producto de 4 combinaciones para cada uno de los hermanos. En este caso, sin embargo, si nos fijamos con atención las 4 posibilidades de genotipo para cada hermano son las mismas, es decir A_1B_1, A_1B_2, A_2B_1 y A_2B_2. Por lo tanto, para cualquiera de las combinaciones del «primer» hermano que elijamos, siempre habrá una de las cuatro del segundo hermano que sea idéntica, como se puede apreciar en la tabla de la derecha de la figura. Es decir, de los 16 pares de genotipos correspondientes a tomar uno en un hermano y el otro en el otro hermano, en 4 de esos pares los genotipos son idénticos, por lo que ahora $d_{CD} = \frac{4}{16} = \frac{1}{4}$

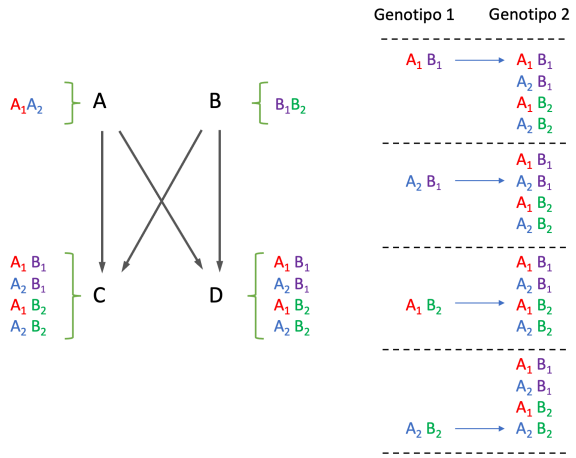


Figura 9.8. Representación de la relación entre hermanos enteros que lleva a un **parentesco de dominancia** distinto de cero. Dada la elección de cualquiera de los cuatro genotipos posibles en uno de los hermanos, la probabilidad de que el segundo hermano posea el mismo genotipo es de 1 en 4, es decir $d_{CD} = \frac{1}{4}$

En estructuras de parentesco más complejas es en general más difícil realizar el desarrollo de todas las combinaciones posibles y es necesario recurrir a otras formas de calcular el **parentesco de dominancia**, pero a los fines de este curso introductorio nos alcanzará con estas relaciones que vimos. Finalmente, los conceptos que acá hemos desarrollado para un locus pueden extenderse con facilidad conceptual (aunque no analítica) a las relaciones entre loci, definiendo un parentesco epistático o de interacción, que apenas mencionaremos llegado el caso.

PARA RECORDAR

- Los efectos de dominancia los habíamos manejado como las desviaciones de la aditividad intralocus y, por lo tanto, deben corresponderse a los genotipos compartidos de **alelos IBD**, en lugar de a los alelos compartidos entre esos individuos.
- Cualquier parentesco directo donde los apareamientos provengan de individuos diferentes y no emparentados entre ellos tendrán un $d_{XY} = 0$ debido a que los padres **no transmiten** a sus hijos combinaciones de alelos.
- Tomando como ejemplo de parentesco colateral los casos de medios hermanos y hermanos enteros, el parentesco de dominancia será $d_{XY} = 0$ y $d_{XY} = \frac{1}{4}$.

9.4 Semejanza entre parientes

Como vimos antes, a partir de nuestro modelo genético básico podemos particionar la varianza fenotípica en sus componentes genético y ambiental, y a su vez los primeros en sus componentes aditivos, de dominancia y de epistasis. Sin embargo, al no tratarse cada uno de estos componentes de entidades observables, no tenemos forma de calcular su valor directamente y debemos recurrir a diferentes estrategias para poder estimar su valor. Dicho de otra forma, lo único que nosotros observamos directamente es el fenotipo y, por lo tanto, lo único que podemos calcular directamente es la varianza fenotípica, por lo que si queremos conocer la varianza genética o aún sus componentes vamos a tener que contar con otra información que nos permita separarla. En este proceso, resulta fundamental entender las diferencias entre componentes causales de la varianza y componentes observables de la misma, así como la relación entre ellos.

Para comenzar a pensar en términos de componentes observables de la varianza veamos un ejemplo sencillo, que aparece representado en la figura 9.9.

En dicha figura tenemos representadas tres situaciones distintas, identificadas por los colores de los trazos continuos que las envuelven (gris, violeta y marrón). En cada una de las situaciones tenemos a su vez 3 potreros, en los que pastan 3 ovejas. Las ovejas pueden ser de 3 tamaños diferentes, 1, 2 o 3. En cada una de las situaciones hay el mismo número de ovejas de cada tamaño, 3 de cada una, pero varía como se distribuyen en cada potrero. En la situación de la izquierda, cada uno de los potreros tiene una oveja de cada tamaño, por lo que el contenido de los potreros es idéntico. Obviamente, la media de tamaño de cada potrero, $\bar{x}$, es igual en los tres (ya que son idénticos), por lo que toda la variación observada en el tamaño de las ovejas corresponde a variación **dentro** del potrero.

Por otro lado, la situación de la derecha es, de alguna manera, la opuesta. El potrero de la izquierda tiene todas las ovejas de tamaño 1, el del centro todas de tamaño 2 y el de la derecha todas de tamaño 3. Por lo tanto, **dentro** de cada potrero los tamaños serán uniformes y no existirá variación (lo que se refleja en que las medias, $\bar{x}$, son idénticas a los valores de las ovejas del potrero correspondiente). Esto nos deja con el hecho de que toda la variación de tamaños observada se corresponde a variación **entre** potreros.

Finalmente, en la situación del medio, las ovejas fueron asignadas al azar entre los 3 potreros. Mientras que en el potrero de la izquierda tenemos una oveja de cada tamaño, en el potrero del centro hay 2 ovejas de tamaño 2 y falta la de tamaño 3, y en el potrero de la derecha

hay 2 ovejas de tamaño 3, faltando la oveja de tamaño 2. Por un lado, ahora los potreros son homogéneos internamente (en todos hay variación **dentro**), pero tampoco son idénticos entre ellos, por lo que también tenemos variación **entre** potreros, lo que se ve reflejado en las diferencias de las medias.

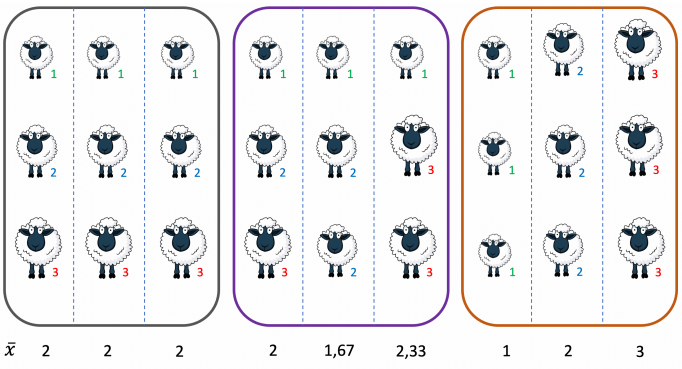


Figura 9.9. Distribución de la variabilidad dentro de grupos y entre grupos. En cada una de las tres situaciones hay un mismo número de ovejas de cada uno de los 3 valores (o sea, 3 ovejas de valor 1, 3 de valor 2 y 3 de valor 3). Sin embargo, en la situación de la izquierda, cada corral tiene una oveja de cada tipo, por lo que los 3 corrales son idénticos y no existe por lo tanto variabilidad **entre** grupos (corrales), concentrándose toda la variabilidad **dentro** de los grupos. En la situación de la derecha nos encontramos con que no existe variación **dentro** de los corrales (ya que el primero tiene todas las ovejas con valor 1, el del medio todas con valor 2 y el último todas con valor 3), por lo que toda la variación está concentrada **entre** los grupos. La situación del medio es intermedia, las ovejas asignadas al azar, con variabilidad **dentro** y variabilidad **entre**. En la línea de abajo, la media de cada grupo (diseño de las ovejas de http://cliparts.co/clipart/2377087 Fuente: Cliparts.co).

Claramente, la distinción de la fuente de variación en componentes **entre** y **dentro** de grupos (potreros) es observable, ya que podemos medir, contar y calcular directamente la varianza (por ejemplo) que se corresponde a cada una de las fuentes. Si llamamos σ^2_B a la varianza **entre** grupos y σ^2_W a la varianza **dentro** de grupos, una forma de distinguir la importancia relativa de las dos fuentes de variación es el coeficiente de correlación intraclase, que definimos como

$$t = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W} \tag{9.22}$$

Al tratarse de una relación entre una varianza y la suma de varianzas (la varianza total), el coeficiente de correlación intraclase podrá variar entre 0 y 1. En la situación de la izquierda de la figura 9.9 tenemos que no existe variación **entre** grupos (ya que son idénticos entre ellos), por lo que $\sigma^2_B = 0$ y, por lo tanto,$t = \frac{\sigma^2_B}{\sigma^2_B+\sigma^2_W} = \frac{0}{0+\sigma^2_W} = 0$. Por otra parte, en la situación de la derecha no existe variación **dentro** de grupos, por lo que $\sigma^2_W = 0$ y por lo tanto $t = \frac{\sigma^2_B}{\sigma^2_B+\sigma^2_W} = \frac{\sigma^2_B}{\sigma^2_B+0} = 1$

Un concepto adicional a entender, es la relación que existe entre las fuentes de variación observables y la similaridad de los individuos dentro de los grupos. La similaridad entre los miembros de cada uno de los grupos, para características métricas, la podemos identificar con la covariación, o en términos estadísticos, con la covarianza entre los miembros del grupo. A medida que los miembros de cada grupo son más similares entre sí, disminuye el efecto de variación **dentro** del grupo, aumentando en consiguiente la proporción que corresponde a **entre** grupos. Por lo tanto, podemos asimilar la covarianza **dentro** de grupos a la varianza **entre** grupos, algo que será fundamental en breve.

La razón por la que la covarianza entre individuos del mismo grupo es igual a la varianza entre grupos se puede demostrar fácilmente. Por ejemplo, en el caso de **hermanos enteros**, si asumimos un modelo en el que el fenotipo de un individuo j, perteneciente a una familia i se corresponde con el siguiente modelo

$$z_{ij} = \mu + f_i + w_{ij} \tag{9.23}$$

entonces la covarianza entre dos individuos j y k pertenecientes a la misma familia (i), es decir **hermanos enteros**, estará dada por

$$Cov_{HE} = Cov(z_{ij},z_{ik}) = Cov[(\mu + f_i + w_{ij}),(\mu + f_i + w_{ik})] = \tag{9.24}$$

$$Cov(f_i,f_i) + Cov(f_i,w_{ik}) + Cov(w_{ij},f_i) + Cov(w_{ij},w_{ik}) = Cov(f_i,f_i) \therefore \tag{9.25}$$

$$Cov_{HE} = \sigma_f^2 \tag{9.26}$$

Por lo tanto, **la covarianza entre individuos del mismo grupo** (familia) **es igual a la varianza entre grupos** (familias), que es lo que queríamos demostrar.

En la situación de la izquierda, los individuos son todos diferentes dentro de cada grupo, por lo que la similaridad (covarianza) dentro de grupos será nula y la proporción de varianza que corresponde a **entre** grupos también. Por el contrario, en la situación de la derecha todos los individuos de cada grupo son idénticos entre sí (máxima covarianza dentro) y, por lo tanto, la varianza **entre** grupos será igual a la total. En el ejemplo de la figura, aún no aportamos la información de cómo se formaron cada uno de los grupos, pero pronto vamos a comprender que si estos se constituyen con base en determinadas relaciones de parentesco entre los integrantes, la **causa de la similaridad** tendrá que ver con la base **genética común** entre ellos.

Pasemos ahora a intentar entender la base **causal** del parecido fenotípico entre parientes, ya que esto nos dará los elementos que nos hacen falta para poder comprender el rol que juega el parentesco en la estimación de los componentes de varianza del modelo genético básico. Asumamos aquí que X e Y son pares de individuos entre los cuales existe un parentesco determinado, el mismo para cada par (por ejemplo, un padre y un hijo, o dos medios hermanos, o dos hermanos enteros, etcétera). Como tenemos datos fenotípicos de los dos, el cálculo de la covarianza fenotípica es directo (aplicando la fórmula para la covarianza, $\mathrm{Cov_{XY}} = \frac{1}{(n-1)} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$). O sea, esta covarianza tendrá un número, en las mismas unidades que la varianza de la característica (las unidades al cuadrado) ya que se trata de la suma del producto de los desvíos de X e Y respecto a sus medias, ambas medidas para la misma característica.

Como, de acuerdo con nuestro modelo genético, $P_X = A_X + D_X + I_X + E_X$ y $P_Y = A_Y + D_Y + I_Y + E_Y$, la covarianza entre medidas en cada par de individuos será igual a

$$\mathrm{Cov}(P_X,P_Y) = \mathrm{Cov}([A_X + D_X + I_X + E_X],[A_Y + D_Y + I_Y + E_Y]) \tag{9.27}$$

$$\mathrm{Cov}(P_X,P_Y) = \mathrm{Cov}(A_X,A_Y) + \mathrm{Cov}(A_X,D_Y) + ... + \mathrm{Cov}(D_X,A_Y) + \mathrm{Cov}(D_X,D_Y) + ... \tag{9.28}$$

En general, en nuestro modelo genético, los términos de covarianza entre componentes de un tipo en un individuo y de otro tipo en otro individuo son cero. Por ejemplo, como vimos antes, por construcción $\mathrm{Cov}(A_X,D_X) = 0$, $\mathrm{Cov}(A_X,I_X) = 0$ y por las mismas razones lo podemos extender a $\mathrm{Cov}(A_X,D_Y) = 0$, $\mathrm{Cov}(A_X,I_Y) = 0$ (no es de esperar que el valor aditivo en un individuo se correlacione o covaríe con desvíos en el otro individuo).

De esta manera,

$$\mathrm{Cov}(P_X,P_Y) = \mathrm{Cov}(A_X,A_Y) + \mathrm{Cov}(D_X,D_Y) + \mathrm{Cov}(I_X,I_Y) + \mathrm{Cov}(E_X,E_Y) \tag{9.29}$$

Por otro lado, vimos que las definiciones de parentesco aditivo que corresponden a Gustave Malecot y Sewall Wright son equivalentes cuando asumimos que ambos individuos no poseen consanguinidad ($F_X = 0$, $F_Y = 0$). Es decir, de acuerdo con Sewall Wright (1922), el parentesco aditivo es igual

$$R_{XY} = \frac{Cov(A_X,A_Y)}{\sqrt{Var(A_X)Var(A_Y)}} = \frac{a_{XY}}{\sqrt{a_{XX}.a_{YY}}} = \frac{a_{XY}}{\sqrt{(1+F_X)(1+F_Y)}} \tag{9.30}$$

por lo que si $F_X = 0$ y $F_Y = 0$ ambas definiciones son equivalentes ($a_{XY} = R_{XY}$). Entonces, de acuerdo con esto, si además asumimos que $Var(A_X) = Var(A_Y)$ (atención, claramente puede no ser cierto en algunas relaciones de parentesco), entonces

$$a_{XY} = \frac{Cov(A_X,A_Y)}{\sqrt{Var(A_X)Var(A_Y)}} = \frac{Cov(A_X,A_Y)}{Var(A)} \iff Cov(A_X,A_Y) = a_{XY}.Var(A) \tag{9.31}$$

De la misma manera, para el parentesco de dominancia (probabilidad de compartir el genotipo en un locus dado entre dos individuos, correlación de los desvíos de dominancia), asumiendo ahora que $\mathrm{Var}(D_X) = \mathrm{Var}(D_Y)$, tenemos

$$d_{XY} = \frac{Cov(D_X,D_Y)}{\sqrt{Var(D_X)Var(D_Y)}} = \frac{Cov(D_X,D_Y)}{Var(D)} \iff Cov(D_X,D_Y) = d_{XY}.Var(D) \tag{9.32}$$

O sea, de ambas deducciones podemos observar que en general el tipo de covarianza entre individuos representa una determinada fracción de la correspondiente varianza y que esa fracción, el coeficiente, es un número que indica la correlación esperada para ese componente del modelo genético entre los dos individuos considerados. De esta forma, podemos extender esa relación a los otros componentes, como epistasis y la parte no genética (ambiental). Sin pérdida de generalidad, asumimos también que la epistasis es usualmente despreciable en las características de interés para nosotros, por lo que, si consideramos que la fracción de la varianza ambiental compartida entre los dos individuos es la varianza del ambiente en común ($e_{XY}\mathrm{Var(E)} = \mathrm{Var(EC)}$, con Var(EC) varianza del ambiente en común, por las siglas en inglés), sustituyendo llegamos a la ecuación general

$$\mathrm{Cov(P_X,P_Y)} = \mathrm{Cov(A_X,A_Y)} + \mathrm{Cov(D_X,D_Y)} + \mathrm{Cov(I_X,I_Y)} + \mathrm{Cov(E_X,E_Y)} \tag{9.33}$$

$$\mathrm{Cov(P_X,P_Y)} = a_{XY}Var(A) + d_{XY}Var(D) + Var(EC) \tag{9.34}$$

Por ejemplo, para padres e hijos, utilizando los subíndices $_P$ para **padre** y $_H$ para **hijo**, podríamos haber derivado explícitamente la relación entre la covarianza fenotípica y los componentes de varianza

$$\mathrm{Cov(P_P,P_H)} = \mathrm{Cov(A_P + D_P + I_P + E_P, A_H + D_H + I_H + E_H)} \tag{9.35}$$

$$\mathrm{Cov(P_P,P_H)} = \mathrm{Cov(A_P + D_P + I_P + E_P, }\frac{1}{2}A_P + D_H + I_H + E_H) \tag{9.36}$$

$$\mathrm{Cov(P_P,P_H)} = Cov(A_P, \frac{1}{2}A_P) = \frac{1}{2}Cov(A_P,A_P) = \frac{1}{2}Var(A) \tag{9.37}$$

$$\mathrm{Cov(P_P,P_H)} = \frac{1}{2}Var(A) \tag{9.38}$$

Sin embargo, recordando que el parentesco aditivo entre un padre y su hijo es de $\frac{1}{2}$, mientras que el parentesco de dominancia es de 0 (los hijos no comparten ninguna de las combinaciones alélicas del padre por vía de descendencia), mientras que en general padres e hijos **no comparten ambiente** (en un sentido amplio), sustituyendo en la ecuación general (ecuación (9.34)) arribamos al mismo resultado.

Veamos que ocurre en el caso de hermanos enteros, ahora utilizando directamente la ecuación general. Si recordamos que el parentesco aditivo entre hermanos enteros es $a_{XY} = \frac{1}{2}$ y el parentesco de dominancia $d_{XY} = \frac{1}{4}$, mientras que por definición comparten **todo el ambiente común**, sustituyendo estos valores en la ecuación (9.34) tenemos

$$\begin{aligned} Cov(P_X,P_Y) &= a_{XY}Var(A) + d_{XY}Var(D) + Var(EC) \\ Cov_{P_{HE}} &= \frac{1}{2}Var(A) + \frac{1}{4}Var(D) + Var(EC) \end{aligned} \tag{9.39}$$

Ejemplo 9.3

Para nuestras amigables ovejas verdes la varianza aditiva del color (verde) fue de 3,2 unidades^2, la varianza de dominancia de 0,4 unidades^2 y la varianza del ambiente común de 1,2 unidades^2. Determinar la covarianza esperada para medios hermanos y para hermanos enteros.

De acuerdo con la ecuación (9.34), sustituyendo el valor de los coeficientes de parentesco aditivo, de dominancia y del ambiente común para cada caso tenemos que para **medios hermanos**

$$Cov_{P_{MH}} = \frac{1}{4}\mathbb{V}_A = \frac{1}{4}3{,}2 = 0{,}8 \text{ unidades}^2 \tag{9.40}$$

mientras que en el caso de **hermanos enteros**

$$Cov_{P_{HE}} = \frac{1}{2}\mathbb{V}_A + \frac{1}{4}\mathbb{V}_D + \mathbb{V}_{EC} = \frac{1}{2}3{,}2 + \frac{1}{4}0{,}4 + 1{,}2 = 2{,}9 \text{ unidades}^2 \tag{9.41}$$

PARA RECORDAR

- La distinción de la fuente de variación en componentes **entre** y **dentro** de grupos es observable. Si llamamos σ^2_B a la varianza **entre** grupos y σ^2_W a la varianza **dentro** de grupos, una forma de distinguir la importancia relativa de las dos fuentes de variación es el coeficiente de correlación intraclase, que definimos como $t = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W}$
- Al tratarse de una relación entre una varianza y la suma de varianzas (la varianza total), el **coeficiente de correlación intraclase** podrá variar entre 0 y 1.
- La similaridad entre los miembros de cada uno de los grupos, para características métricas, la podemos identificar con la covarianza entre los miembros del grupo. A medida que los miembros de cada grupo son más similares entre sí, disminuye el efecto de variación **dentro del grupo**, aumentando en consiguiente la proporción que corresponde a **entre grupos**. Por lo tanto, podemos asimilar la covarianza dentro de grupos a la varianza entre grupos.

9.5 Estimación de las varianzas aditiva y de dominancia

Vimos antes que es posible relacionar la covarianza fenotípica entre individuos con un grado de parentesco determinado y los distintos componentes de varianza de nuestro modelo genético básico. En particular, usualmente es de nuestro interés poder estimar algún componente de varianza en particular, por ejemplo, la varianza aditiva. Como veremos más adelante, la varianza aditiva es el numerador de ecuación que define la heredabilidad como relación de varianzas, $h^2 = \hat{\sigma}^2_A/\hat{\sigma}^2_P$, posiblemente uno de los parámetros más relevantes a la hora de decidir una estrategia de mejora, además de ser la fuente de variación sobre la que actúa normalmente la selección.

De acuerdo con la ecuación general (9.34) vista antes,

$$Cov(P_X,P_Y) = a_{XY}\ \sigma^2_A + d_{XY}\ \sigma^2_D + \sigma^2_{EC} \tag{9.42}$$

Si prestamos atención, en la ecuación (9.42) hicimos algún cambio de notación respecto a la versión original para remarcar algunos aspectos importantes de esta última fórmula. Por un lado, mezclamos letras griegas (sigma, σ) con nuestra notación más tradicional hasta ahora ($Cov_{P_{XY}}$ en lugar de σ_{XY}, mucho más sencilla). El objetivo de esta doble notación es remarcar la existencia de una única cantidad directamente estimable de los datos, que es la covarianza fenotípica. Además, introdujimos el acento circunflejo (^) como símbolo de algo que queremos estimar (a lo que le llamamos estimador del «verdadero valor»).

A partir del resultado de la ecuación (9.42) resulta sencillo derivar un estimador de la varianza aditiva, «despejando» $\hat{\sigma}^2_A$ de dicha ecuación

$$\begin{aligned} Cov(P_X,P_Y) = a_{XY}\ \sigma^2_A + d_{XY}\ \sigma^2_D + \sigma^2_{EC} \therefore \\ a_{XY}\ \sigma^2_A = Cov(P_X,P_Y) - [d_{XY}\ \sigma^2_D + \sigma^2_{EC}] \therefore \\ \hat{\sigma}^2_A = \frac{Cov_{P_{XY}} - d_{XY}\ \hat{\sigma}^2_D - \hat{\sigma}^2_{EC}}{a_{XY}} \end{aligned} \tag{9.43}$$

Básicamente, nuestro estimador general de la varianza aditiva se construye restando la fracción del estimado de varianza de dominancia y de la varianza del ambiente común que correspondan, al valor obtenido de estimar la **covarianza entre observaciones en los pares de individuos** (es decir, la covarianza fenotípica), y luego dividiendo entre el parentesco aditivo entre esos pares. De todos estos números, los parentescos aditivos y de dominancia son usualmente conocidos, así como la presencia o ausencia de varianza de ambiente en común, aunque no, obviamente, su valor. Claramente, se trata de muchas incógnitas para una sola ecuación. Sin embargo, a partir de los datos fenotípicos, es bastante obvio que podemos utilizar el estimador clásico de la covarianza a partir de una muestra ($\hat{\sigma}_{XY} = [\sum(x-\bar{x})(y-\bar{y})]/(n-1)$) para obtener un estimado de la covarianza fenotípica (que antes notamos como $\mathrm{Cov}_{\mathrm{P_{XY}}}$ especialmente para que se entienda que es la única cantidad que podemos estimar directamente de los datos).

Esto nos deja con una herramienta de estimación muy interesante, pero a la que generalmente le faltarán insumos (información) para explotar todo su potencial. Veamos lo que ocurre con tres estructuras de parentesco que son usuales en diseños experimentales relacionados a la genética cuantitativa: a) padre-hijo (PH), b) medios-hermanos (MH) y c) hermanos enteros (HE). En el primer caso, tenemos que el parentesco aditivo vale $1/2$ (un padre comparte exactamente la mitad de su información genético-aditiva con su hijo, es el único parentesco que es determinístico en este sentido) y nada de parentesco de dominancia (no es posible que comparta ninguna combinación de sus alelos), así como nada del ambiente es común (en general, para mediciones realizadas en la misma edad en dos generaciones el ambiente ha cambiado lo suficiente como para no estar correlacionado).

Por lo tanto, si sustituimos esos coeficientes en la fórmula general, tenemos

$$\hat{\sigma}_A^2 = \frac{[\mathrm{Cov}_{\mathrm{PH}} - 0\ \hat{\sigma}_D^2 - 0\ \hat{\sigma}_{EC}^2]}{\frac{1}{2}} = 2\ \mathrm{Cov}_{\mathrm{PH}} \tag{9.44}$$

o lo que es lo mismo, 2 veces la covarianza fenotípica entre parejas de un padre y su hijo (uno) es un estimador insesgado de la varianza aditiva (ya veremos un poquito más adelante cuáles son algunas propiedades deseables de los estimadores).

Repitiendo el procedimiento para el caso de medios-hermanos, el coeficiente de parentesco aditivo es ahora de $1/4$, mientras que el de dominancia es 0 y no hay un ambiente en común (porque, por ejemplo, son criados por distintas madres), lo que nos lleva a

$$\hat{\sigma}_A^2 = \frac{[\mathrm{Cov}_{\mathrm{PMH}} - 0\ \hat{\sigma}_D^2 - 0\ \hat{\sigma}_{EC}^2]}{\frac{1}{4}} = 4\ \mathrm{Cov}_{\mathrm{PMH}} \tag{9.45}$$

En este caso, 4 veces la covarianza fenotípica entre medio-hermanos es un estimador insesgado de la varianza aditiva (en la medida en que se cumpla lo que asumimos, por ejemplo, que no comparten un ambiente en común).

Finalmente, en el caso de **hermanos enteros** (aquellos que comparten padre y madre), la situación se vuelve un poco más complicada. Ahora el parentesco aditivo es de $1/2$, pero el de dominancia ya no es de 0 sino de $1/4$ y en las características que tienen relación con la etapa de crecimiento el ambiente en común es considerable (pensemos, por ejemplo, en los lechones de la misma camada, donde además de compartir el ambiente uterino, los lechones comparten la madre durante la etapa de crecimiento, así como el chiquero). Esto nos lleva a que ahora

$$\hat{\sigma}_A^2 = \frac{[\mathrm{Cov}_{\mathrm{PHE}} - \frac{1}{4}\ \hat{\sigma}_D^2 - \hat{\sigma}_{EC}^2]}{\frac{1}{2}} = 2\ \mathrm{Cov}_{\mathrm{PHE}} - \frac{1}{2}\ \hat{\sigma}_D^2 - 2\ \hat{\sigma}_{EC}^2 \leq 2\ \mathrm{Cov}_{\mathrm{PHE}} \tag{9.46}$$

ya que las varianzas deben ser mayores o iguales a 0. Dicho de otra forma, 2 veces la covarianza fenotípica entre hermanos enteros sobrestima consistentemente el «verdadero valor» de la varianza aditiva, excepto en el caso de que se trate de una característica en la que no exista varianza de dominancia y no exista ambiente en común. Como no tenemos forma de separar los aportes de los distintos componentes causales de varianza que quedan, $2\ \mathrm{Cov}_{\mathrm{PHE}}$ es un **estimador sesgado** (hacia arriba) de la varianza aditiva.

Ejemplo 9.4

En un estudio realizado para obtener estimados de componentes genéticos de la varianza de diámetro de fibras en ovejas verdes se realizaron mediciones bajo dos diseños experimentales: 1) un diseño de **medios hermanos** y 2) un diseño de **hermanos enteros**. En el primer diseño se obtuvo un estimado de la covarianza fenotípica de $\mathrm{Cov}_{\mathrm{PMH}} = 0{,}45\ \mu\mathrm{m}^2$, mientras que en el segundo se obtuvo un estimado de la covarianza fenotípica de $\mathrm{Cov}_{\mathrm{PHE}} = 1{,}25\ \mu\mathrm{m}^2$.

En los dos diseños los hermanos fueron separados de sus madres al nacer y alimentados y criados todos juntos (por lo tanto, no compartieron un ambiente especial a las parejas de hermanos). Obtener un estimado de la varianza aditiva, así como un estimado de la varianza de dominancia a partir de los datos anteriores.

Usando la ecuación (9.34), en el caso de **hermanos enteros** tenemos que la covarianza fenotípica, en nuestro caso ($\hat{\sigma}^2_{EC} = 0$) es igual a

$$\mathrm{Cov}_{\mathrm{P_{HE}}} = \frac{1}{2}\hat{\sigma}^2_A + \frac{1}{4}\,\hat{\sigma}^2_D + \hat{\sigma}^2_{EC} = \frac{1}{2}\hat{\sigma}^2_A + \frac{1}{4}\,\hat{\sigma}^2_D = 1{,}25\ \mu\mathrm{m}^2 \tag{9.47}$$

Multiplicando por 2, tenemos entonces que

$$2\ \mathrm{Cov}_{\mathrm{P_{HE}}} = \hat{\sigma}^2_A + \frac{1}{2}\,\hat{\sigma}^2_D = 2 \times 1{,}25 = 2{,}50\ \mu\mathrm{m}^2 \tag{9.48}$$

Para el caso de **medios hermanos**, la covarianza fenotípica es igual a

$$\begin{aligned} \mathrm{Cov}_{\mathrm{P_{MH}}} = \frac{1}{4}\hat{\sigma}^2_A = 0{,}45\ \mu\mathrm{m}^2 \therefore \\ 4\ \mathrm{Cov}_{\mathrm{P_{MH}}} = \hat{\sigma}^2_A = 4 \times 0{,}45 = 1{,}80\ \mu\mathrm{m}^2 \end{aligned} \tag{9.49}$$

Esto último nos da un estimado directo e insesgado de la varianza aditiva, que es por lo tanto $\hat{\sigma}^2_A = 1{,}80\ \mu\mathrm{m}^2$.

Por otro lado, restando las dos ecuaciones de covarianzas multiplicadas por sus respectivos coeficientes para que en ambas el coeficiente que multiplica $\hat{\sigma}^2_A$ sea 1 (y así se eliminen al restarlos), tenemos ahora

$$\begin{aligned} 2\ \mathrm{Cov}_{\mathrm{P_{HE}}} - 4\ \mathrm{Cov}_{\mathrm{P_{MH}}} = \hat{\sigma}^2_A + \frac{1}{2}\,\hat{\sigma}^2_D - \hat{\sigma}^2_A = \frac{1}{2}\,\hat{\sigma}^2_D = (2{,}50 - 1{,}80)\ \mu\mathrm{m}^2 \therefore \\ \frac{1}{2}\,\hat{\sigma}^2_D = 0{,}70\ \mu\mathrm{m}^2 \therefore \hat{\sigma}^2_D = 2 \times 0{,}70\ \mu\mathrm{m}^2 = 1{,}40\ \mu\mathrm{m}^2 \end{aligned} \tag{9.50}$$

Por lo tanto, para el diámetro de fibras en ovejas verdes, la varianza de dominancia es $\hat{\sigma}^2_D = 1{,}40\ \mu\mathrm{m}^2$

Propiedades deseables de los estimadores

En varias partes del libro hemos hablado de estimadores y en algunos casos hemos apelado a alguna propiedad deseable de los mismos, pero sin explicar demasiado nuestras pretensiones o siquiera qué es un estimador. Para entender mejor de lo que estamos hablando podemos poner un ejemplo sencillo: deseamos estimar la proporción de varones en la población uruguaya, obviamente sin acceder a los datos censales. En un instante dado del tiempo esta proporción tendrá un valor fijo, por ejemplo, si hay 1.707.176 mujeres y 1.579.138 varones,

la proporción de varones en la población total es igual a $\frac{1.579.138}{1.579.138+1.707.176} = 0{,}4805195$. Claramente esta proporción variará en forma ínfima en los lapsos de tiempo de días o aún meses, por lo que podemos asumirla fija. Desafortunadamente, censar a toda la población es un trabajo enorme, que se realiza cada varios años y nosotros queremos obtener un estimado actual de la proporción de varones.

Para acometer esta tarea se nos ocurre un procedimiento sencillo (entre muchos posibles): contar el número de peatones de cada sexo que pasan por una esquina concurrida de Montevideo durante un intervalo de tiempo predefinido (por ejemplo, una hora). Al cabo de esa hora, tendremos una tabla con conteos para varones y mujeres y, por lo tanto, podemos sacar de la misma la fracción correspondiente de varones. Claramente, en una hora nunca va a pasar toda la población de Uruguay por esa esquina, por lo que la fracción que salga de nuestro conteo no será otra cosa que un **estimador** del «verdadero» valor poblacional. Un estimador es por lo tanto un estadístico (un procedimiento de cálculo a realizar con los datos) que pretende **estimar** el verdadero valor del parámetro.

Los estimadores son en sí mismo variables aleatorias ya que su valor dependerá, en general, de la muestra extraída. Al tratarse de variables aleatorias, si realizamos el experimento de repetir el muestreo muchas veces, en nuestro caso podría ser realizar el conteo todos los días durante un mes, por ejemplo, los valores del estimador variarán consecuentemente. Una forma sencilla de imaginarnos el impacto de la variación aleatoria en la repetición del muestreo es hacer un histograma con los valores del estimador obtenido cada día. De hecho, si pensamos un poco, es posible que nuestro estimador previo se encuentre afectado por el día de la semana (los domingos casi nadie pasa por esa esquina y luego de una hora apenas habremos registrado de 10 a 15 personas), por lo que podemos pensar en el mismo estimador (fracción de varones en la muestra) pero basado en un procedimiento de muestreo diferente: contar los sexos en las primeras N personas que pasen a partir de las 8 de la mañana. Este estimador nos lleva a preguntarnos cómo afectará el tamaño de la muestra el comportamiento del estimador. Si $N = 1$ el estimador solamente podrá decir que la población está constituida 100 % por mujeres o 100 % por varones. Si $N = 2$, los valores posibles del estimador serán 0, 50 %, o 100 %. Con $N = 3$, los valores posibles serán 0, 33 %, 66 % o 100 %. ¿Qué ocurriría si en realidad el verdadero valor en la población es de 50 % y nuestro $N = 3$? En este caso, el estimador siempre le erraría al verdadero valor al menos en 17 %, por arriba o por abajo. Claramente, no parece un estimador que funcione muy bien con este tamaño muestral.

Lo anterior nos lleva a preguntarnos qué características serían deseables en un estimador, para luego ver si nuestro estimador cumple con las mismas. Existen varias características que podrían ser deseables en determinados casos, pero en general las siguientes son siempre deseables:

1. **Insesgado:** Un estimador es insesgado cuando el valor medio de las estimaciones obtenidas al repetir el experimento sucesivamente para diferentes muestras es igual al valor del parámetro. De otra forma, si θ es el parámetro que queremos estimar, entonces $\mathbb{E}(\hat{\theta}) = \theta$.
2. **Consistente:** Un estimador es consistente cuando el valor estimado se aproxima al del parámetro desconocido al aumentar el tamaño de la muestra arbitrariamente. En términos matemático $\hat{\theta}_{n\to\infty} \to \theta$.

3. **Eficiente:** Un estimador es eficiente cuando su varianza es tan pequeña como sea posible.
4. **Suficiente:** Un estimador es suficiente cuando aprovecha toda la información existente en la muestra.

Claramente, varias de estas propiedades es necesario analizarlas y demostrarlas en forma matemática para los estimadores de interés, pero solamente como ejemplo, podemos ver que nuestro estimador de la proporción de varones en la población uruguaya es (al menos) **insesgado** y **consistente**. Por ejemplo, aún con tamaños muestrales pequeños, si bien el valor derivado de cada muestra se puede apartar en forma sustantiva del verdadero valor, el valor promedio obtenido de repetir el muestreo en forma consistente coincidirá con el valor del parámetro poblacional. Por otro lado, si aumentamos el tamaño muestral de $N = 3$ a $N = 100$, los resultados podrán aproximarse al verdadero valor con una diferencia de 0,5 %, y si lo aumentamos aún más, a $N = 1000$, se podrán aproximar más aún (a 0,05 %), por lo que el estimador es consistente.

PARA RECORDAR

- Podemos describir la covarianza fenotípica mediante: $Cov(P_X,P_Y) = a_{XY}\ \sigma^2_A + d_{XY}\ \sigma^2_D + \sigma^2_{EC}$
- El estimador general de la varianza aditiva se construye restando la fracción del estimado de varianza de dominancia y de la varianza del ambiente común que correspondan, al valor obtenido de estimar la covarianza entre observaciones en los pares de individuos (es decir, la covarianza fenotípica), y luego dividiendo entre el parentesco aditivo entre esos pares.
- Los **estimadores** son estadísticos que pretenden estimar el verdadero valor del parámetro, son en sí mismo variables aleatorias ya que su valor dependerá, en general, de la muestra extraída.
- Las propiedades deseables en un estimador es que el mismo sea: **insesgado**, **consistente**, **eficiente** y **suficiente**.

Estimación de la varianza aditiva a partir del análisis de varianza (ANOVA)

El análisis de varianza, desarrollado por Ronald Fisher, es un método que nos permite estimar y comparar diferentes fuentes de variación. En una de sus versiones usuales, que seguramente hayas estudiado, es una herramienta que nos permite comparar las medias de diferentes tratamientos, así como la significación estadística de estas diferencias (Modelo I, o de efectos fijos). En otra versión (Modelo II, o de efectos aleatorios), los niveles del factor relevante se consideran una muestra aleatoria de todos los niveles posibles y, por lo tanto, los resultados se pueden extrapolar al resto de la población. Esto va a resultar particularmente atractivo para nosotros porque a partir de un diseño experimental con algunos animales vamos a poder estimar la varianza aditiva en toda la población y no solo restringiéndonos a los animales incluidos en el experimento.

Consideremos un diseño experimental de **medios hermanos**, en donde cada macho de los incluidos en el experimento se aparea con varias hembras, que dejan cada una cría, como se puede apreciar en la figura 9.10.

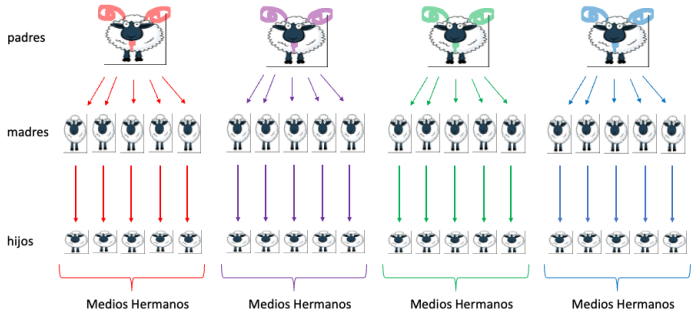


Figura 9.10. Representación de un experimento de estimación de la varianza aditiva mediante un diseño de medios hermanos. Cada carnero (padres, línea superior) se aparea con varias ovejas (madres, línea del medio), formando cada padre un grupo de **medios hermanos** entre sí (línea de abajo). Fuente del diseño de las ovejas: cliparts.co, modificado para este trabajo.

En la misma, cada uno de los 4 carneros (padres) se aparea con 5 ovejas (madres) y el resultado es, a partir de una cría por madre, grupos de 5 medios hermanos. En estos animales será donde realizaremos las mediciones fenotípicas de nuestro interés, por ejemplo, peso al destete, peso del primer vellón o la característica de interés. Como los grupos se forman a partir del padre en común, las madres se eligen al azar, a este modelo también se le conoce como **modelo padre**.

Desde el punto de vista estadístico podemos escribir el modelo como

$$Y_{ij} = \mu + s_i + e_{ij} \tag{9.51}$$

donde Y_{ij} representa el fenotipo del animal j dentro del padre i, μ es la media de la característica en la población, s_i representa el desvío de la media correspondiente al padre i (s por «sire») y e_{ij} representa el «residuo» del animal j dentro del padre i. Si suponemos que el diseño del experimento es balanceado, entonces el número de hijos por padre será igual (k medios hermanos) en todos los s grupos. En este diseño es muy fácil identificar las dos fuentes de variación: por un lado la variación producida por diferencias **entre padres** y por el otro lado las diferencias de los animales **dentro del grupo de medios hermanos**. Como no tenemos ninguna hipótesis particular ni factor asociado a ella que explique los desvíos de los hijos respecto al **valor del padre**, tratamos a estos desvíos como el residuo del modelo (que puede contener tantos errores de medición, efectos no explicados por el padre y otras fuentes de variación no relevante para nosotros). Dado que una fuente de variación es entre grupos y la otra dentro de grupos, llamamos a sus varianzas respectivamente $\sigma^2_B = \sigma^2_S$ (la $_B$ es de *between*, entre, mientras que la $_S$ es por *sire*) y σ^2_W (por *within*, dentro). Cuando

vimos antes el **coeficiente de correlación intraclase** le llamamos σ^2_B a la varianza **entre grupos**, mientras que acá vamos a notarla con σ^2_S para reforzar la idea de que es el efecto de los padres. Teniendo en cuenta todo lo anterior, podemos organizar la información en una **tabla de ANOVA** como la siguiente:

Fuente de variación	G. L.	Cuadrado medio	E(MS)
Entre padres (grupos de medios hermanos)	$s-1$	$\mathrm{MS_S}$	$\sigma^2_W + k\sigma^2_S$
Dentro de grupos (medios hermanos)	$s(k-1)$	$\mathrm{MS_W}$	σ^2_W
Total	$sk-1$		

Los cuadrados medios se calculan directamente de los datos fenotípicos, de acuerdo a

$$\mathrm{MS_S} = \frac{\sum_{i=1}^{s} n_i(\bar{x}_i - \bar{x})^2}{s-1} \tag{9.52}$$

y

$$\mathrm{MS_W} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{sk-s} \tag{9.53}$$

Si observamos la última columna de la tabla de ANOVA, la misma contiene la esperanza de los cuadrados medios, es decir, su valor teórico. Si igualamos el valor teórico esperado al valor del cuadrado medio correspondiente podemos despejar estimadores para las varianzas **entre** y **dentro**. Por ejemplo, si operamos con la línea correspondiente a la fuente de variación **dentro de grupo**, tenemos que

$$\sigma^2_W = MS_W \tag{9.54}$$

Es decir, el cuadrado medio $\mathrm{MS_W}$ es un estimador insesgado de la varianza **dentro** (σ^2_W). Igualando el cuadrado medio **entre grupos** a su esperanza y sustituyendo σ^2_W por $\mathrm{MS_W}$ tenemos

$$\begin{aligned} \mathrm{MS_S} = \sigma^2_W + k\sigma^2_S = \mathrm{MS_W} + k\sigma^2_S \therefore \\ k\sigma^2_S = \mathrm{MS_S} - \mathrm{MS_W} \therefore \\ \sigma^2_S = \frac{\mathrm{MS_S} - \mathrm{MS_W}}{k} \end{aligned} \tag{9.55}$$

En el caso de que el diseño sea desbalanceado, es decir que el número de hijos por padre (n_i) no sea el mismo, si llamamos $N = \sum n_i$ al número total de hijos de todos los toros, entonces el factor k corregido se calcula como

$$k = \frac{1}{s-1}\left(N - \frac{\sum n_i^2}{N}\right) \tag{9.56}$$

Pero en nuestro modelo, las diferencias entre grupos de medios hermanos (varianza σ^2_S) se corresponden con la similaridad dentro de cada grupo de medios hermanos, es decir su covarianza ($\mathrm{Cov_{P_{MH}}}$), es decir

$$\sigma^2_S = Cov_{P_{MH}} \tag{9.57}$$

Pero, de acuerdo con la ecuación (9.45), **en el caso de medios hermanos** la relación entre la covarianza (observable) de **medios hermanos** y la varianza aditiva (a estimar) está dada por

$$\hat{\sigma}^2_A = 4\, Cov_{P_{MH}} \tag{9.58}$$

Por lo tanto, uniendo el resultado de esta última ecuación con la ecuación (9.57), tenemos que nuestro estimador de la varianza aditiva a partir del análisis de varianza está dado por

$$\hat{\sigma}^2_A = 4\, Cov_{P_{MH}} = 4\sigma^2_S \tag{9.59}$$

Ejemplo 9.5

En la población de ovejas verdes que han estado presentes en varias partes del libro se realizó un experimento para estimar la varianza aditiva del **peso del vellón verde limpio** (PVVL). Teniendo en cuenta las condiciones de la población se trabajó con un diseño de **medios hermanos** balanceado, donde los segundos vellones de 4 medios hermanos por carnero (con 4 carneros) fueron medidos. Los resultados de la medición se encuentran en la siguiente tabla:

Padre	MH	Peso vellón limpio (kg)
p1	mh1.1	3,5
p1	mh1.2	3,4
p1	mh1.3	3,9
p1	mh1.4	3,7
p2	mh2.1	4,0
p2	mh2.2	4,1
p2	mh2.3	4,4
p2	mh2.4	3,7
p3	mh3.1	2,9
p3	mh3.2	2,8
p3	mh3.3	4,1
p3	mh3.4	3,5
p4	mh4.1	4,3
p4	mh4.2	3,6
p4	mh4.3	3,1
p4	mh4.4	3,4

Estimar la varianza aditiva de peso de vellón verde limpio a partir de los datos.

De acuerdo con lo que vimos previamente, podemos organizar una tabla de ANOVA con los estimadores derivados de los cuadrados medios y luego igualar estos resultados a las esperanzas. El cuadrado medio de los padres se calcula como

$$MS_S = \frac{\sum_{i=1}^{s} n_i(\bar{x}_i - \bar{x})^2}{s-1} \tag{9.60}$$

por lo que debemos calcular antes la media general y las medias para cada padre. Estas son

Padre	Media PVVL (kg)	Media general (kg)	Desvío padre (kg)	Desvío cuadrado (kg^2)
p1	3,625	3,65	−0,025	0,000625
p2	4,050	3,65	0,400	0,160000
p3	3,325	3,65	−0,325	0,105625
p4	3,600	3,65	−0,050	0,002500
Suma			0	0,268750

Por lo tanto, tomado los valores de la tabla anterior, tenemos que el cuadrado medio de padres es

$$MS_S = \frac{\sum_{i=1}^{s} n_i(\bar{x}_i - \bar{x})^2}{s-1} = \frac{4 \times 0{,}268750}{4-1} \therefore \\ MS_S = 0{,}3583333 \text{ kg}^2 \tag{9.61}$$

Por otro lado, para el cuadrado medio del error necesitamos los desvíos de cada observación respecto a la media de su padre y elevar esa diferencia al cuadrado, es decir

Padre	MH	PVLL (kg)	Desvío (kg)	Desvío cuadrado (kg^2)
p1	mh1.1	3,5	−0,125	0,015625
p1	mh1.2	3,4	−0,225	0,050625
p1	mh1.3	3,9	0,275	0,075625
p1	mh1.4	3,7	0,075	0,005625
p2	mh2.1	4,0	−0,050	0,002500
p2	mh2.2	4,1	0,050	0,002500
p2	mh2.3	4,4	0,350	0,122500
p2	mh2.4	3,7	−0,350	0,122500
p3	mh3.1	2,9	−0,425	0,180625
p3	mh3.2	2,8	−0,525	0,275625
p3	mh3.3	4,1	0,775	0,600625
p3	mh3.4	3,5	0,175	0,030625
p4	mh4.1	4,3	0,700	0,490000
p4	mh4.2	3,6	0,000	0,000000
p4	mh4.3	3,1	−0,500	0,250000
p4	mh4.4	3,4	−0,200	0,040000
Suma		58,4	0	2,265

Por lo tanto, poniendo estos valores en la fórmula y con $s = 4$, $k = 4$ y $sk - s = 12$ tenemos

$$MS_W = \frac{\sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{sk - s} = \frac{2{,}265}{12} = 0{,}18875 \text{ kg}^2 \tag{9.62}$$

Ahora, usando los resultados de las ecuaciones (9.55) y (9.59),

$$\begin{aligned} \sigma_S^2 = \frac{MS_S - MS_W}{k} = \frac{0{,}3583333 - 0{,}18875}{4} = 0{,}04239583 \text{ kg}^2 \\ \sigma_A^2 = 4\,\sigma_S^2 = 4 \times 0{,}04239583 = 0{,}1695833 \text{ kg}^2 \end{aligned} \tag{9.63}$$

que es el resultado que buscábamos.

Estimaciones basadas en modelos mixtos lineales

Los procedimientos de estimación de los componentes de varianza que vimos hasta ahora se basan en diseños de estructuras de parentesco determinadas, algo que puede ser relativamente sencillo o práctico de hacer en determinadas especies y casi imposible en otras. Más aún, el plantear este tipo de experimentos para la determinación de componentes de varianza suele traer aparejado costos y complejidades importantes, lo que limita en general el número de individuos que participan del estudio. A su vez, esto implica errores importantes en las estimaciones de los componentes de varianza que muchas veces reducen en forma sustantiva la utilidad de los resultados obtenidos.

Como veremos más adelante en el curso y en cursos específicos más avanzados, el procedimiento normalmente utilizado en la estimación de varianzas para las evaluaciones genéticas se basa en procedimientos iterativos de ajuste a los datos, «frecuentistas» o «bayesianos», a partir de las ecuaciones del modelo mixto lineal, que usualmente se conocen como **BLUP** (por mejor predictor lineal insesgado, en inglés). Para que este procedimiento funcione se precisa, aparte de los datos fenotípicos de la característica a la que le queremos estimar los componentes de varianza, la matriz de parentesco aditivo entre los animales que participan de la evaluación. La idea básica en el procedimiento «frecuentista» es que, si bien las varianzas (por ejemplo, genético-aditiva y ambiental) son desconocidas, es posible introducir valores arbitrarios para las mismas y ver el ajuste del modelo con esa prueba (por ejemplo, a través de la función de **verosimilitud**). El paso siguiente será cambiar ligeramente esos valores y recalcular la **verosimilitud** a cada vez que cambiamos los valores de las varianzas. Al repetir este proceso, con suerte se encontrará una tendencia a que crezca la **verosimilitud** hasta determinado punto en que comienza a bajar nuevamente: en este caso, el último punto de crecimiento sería **el máximo de la función de verosimilitud** y, por lo tanto, el valor de las varianzas que lo produjeron el mejor **estimador puntual** de las mismas. En algunos casos, es posible derivar formas explícitas para la varianza asintótica de los estimadores de varianza, lo que servirá para determinar los intervalos de confianza asintóticos de los mismos.

Los procedimientos bayesianos MCMC (Cadenas de Markov Monte Carlo) difieren del anterior en que hacen uso de información *a priori* sobre la distribución de los parámetros a estimar (las varianzas, por ejemplo) y que el resultado es la distribución *a posteriori* de

los parámetros (que son también considerados variables aleatorias), por lo que se puede realizar inferencia directamente sobre estas distribuciones (sin caer en la consideración de «asintótico»).

En cualquier caso, es de destacar que, si bien no existe un experimento diseñado para estimar los componentes y que por lo tanto las relaciones entre individuos son variadas, la clave de la estimación sigue siendo la relación esperada (parentesco) entre individuos. Esta información se encuentra ahora especificada en la matriz de parentesco aditivo **A**, que también se conoce como matriz del numerador del parentesco. La razón de este nombre es que si recordamos la ecuación (9.30) con la relación entre la definición del coeficiente de parentesco de Sewall Wright (1922) y el parentesco aditivo, para los individuos i y j la misma era

$$R_{ij} = \frac{a_{ij}}{\sqrt{(1+F_i)(1+F_j)}} \tag{9.64}$$

y, por lo tanto, los elementos de **A**, que para los individuos i y j son $A_{i,j} = A_{j,i} = a_{ij}$, son los elementos del numerador del coeficiente de parentesco de Sewall Wright (1922). Más adelante, en el ejemplo 9.6 una de las preguntas nos permitirá vincular más claramente estas dos formas de ver el parentesco.

PARA RECORDAR

- Estadísticamente, podemos definir el modelo padre como $Y_{ij} = \mu + s_i + e_{ij}$ dónde Y_{ij} representa el fenotipo del animal j dentro del padre i, μ es la media de la característica en la población, s_i representa el desvío de la media correspondiente al padre i y e_{ij} representa el «residuo» del animal j dentro del padre i.
- En este tipo de diseño podemos identificar dos fuentes de variación: la producida por diferencias **entre padres** y por el otro lado las diferencias de los animales **dentro del grupo de medios hermanos.** Determinadas estas fuentes, llamamos a sus varianzas respectivamente $\sigma^2_B = \sigma^2_S$ (la $_B$ es de *between*, entre, mientras que la $_S$ es por *sire*) y σ^2_W (por *within*, dentro).
- En el caso de medios hermanos la relación entre la covarianza (observable) de **medios hermanos** y la **varianza aditiva** (a estimar) está dada por $\hat{\sigma}^2_A = 4\,\mathrm{Cov}_{\mathrm{P_{MH}}}$
- Podemos realizar estimaciones a través de modelos mixtos lineales mediante el **BLUP**, para este procedimiento necesitamos, además de los datos fenotípicos de la característica a la que le queremos estimar los componentes de varianza, la matriz de parentesco aditivo entre los animales que participan de la evaluación.

9.6 Parentesco genómico

Hasta el momento nos hemos manejado con los parentescos esperados a partir de las reglas básicas de transmisión de la información de progenitores a descendencia, pero sin posibilidad de atender en alguna medida la verdadera proporción del genoma compartido por dos individuos. Es por esta razón que hemos llamado de parentesco estadístico (o probabilístico) al tratamiento basado en la esperanza matemática de la fracción del componente del genoma de nuestro interés (aditivo, dominancia, etcétera) compartido entre dos individuos. Como vimos antes, la única relación de parentesco que comparte siempre exactamente la misma fracción con otro es el de progenitor-progenie, en cuyo caso la mitad del genoma es compartido y, por lo tanto, el parentesco aditivo es igual a $\frac{1}{2}$. En el resto de las relaciones entre individuos, la fracción del genoma compartido será una variable aleatoria, con una distribución determinada, con su media y con su varianza.

Sin embargo, los avances en la genómica durante los últimos veinte años han permitido que el **genotipado** de animales, plantas, etcétera, sea una práctica rutinaria en diferentes programas de mejoramiento genético y aún una poderosa herramienta para el estudio de las poblaciones humanas. En el caso del mejoramiento genético, conocer la información del genotipo de los individuos mejora sustancialmente la predicción de las evaluaciones genéticas, aportando información adicional tanto para los individuos genotipados como para sus parientes no genotipados. La información aportada por el genotipado puede ayudar, por ejemplo, a entender las diferencias en *performance* entre dos hermanos, que a priori tendrían el mismo parentesco con sus padres (que son los que aportan el material genético en la misma proporción para ambos). Una forma de ver la relación entre el parentesco estadístico que vimos previamente y el genómico es la de considerar al primero como el **parentesco esperado** y al segundo como el **parentesco realizado** (u observado).

Veamos entonces de qué se trata el parentesco genético. Lo primero es definir de qué se trata el genotipado y qué es lo que usualmente obtenemos de un **chip de genotipado**. El genotipado consiste básicamente en determinar el genotipo, es decir el par de alelos concretos, presentes en cada uno de los loci de interés. Normalmente se trata de **SNP**, es decir, variantes de un solo nucleótido. Si bien en cada posición del genoma usualmente podemos encontrar cualquiera de las 4 bases (**A**, **C**, **G** o **T**) y, por lo tanto, $\frac{k(k+1)}{2} = \frac{4\times5}{2} = 10$ pares posibles, en general, los chips de genotipado se construyen sobre loci que son mayoritariamente bialélicos, es decir, loci para los cuales solo vamos a encontrar una de dos posibilidades. Esto reduce el número de pares posibles a $\frac{k(k+1)}{2} = \frac{2\times3}{2} = 3$. Por ejemplo, si se trata de un locus en el que los dos alelos presentes son **A** y **G**, podremos tener cualquiera de los siguientes tres genotipos: **AA**, **AG** o **GG**. Como en cada locus tenemos 3 posibilidades (con marcadores bialélicos), en principio, si los marcadores se encuentran a suficiente distancia entre ellos, tendremos 3^m combinaciones distintas posibles, con m el número de marcadores. Pese a que no parezca muy llamativo, este número de combinaciones posibles es impactante. Por ejemplo, si $m = 10$ marcadores, entonces $3^{10} = 59.049$ combinaciones diferentes posibles (individuos que podríamos distinguir), pero si pasamos a $m = 20$, entonces $3^{20} = 3.486.784.401$, es decir casi 3500 millones de combinaciones diferentes posibles. Los chips que se utilizan normalmente (densidad media) en mejoramiento genético animal

contienen del orden de los 50.000 marcadores y el número $3^{50.000}$ escapa a la representación por la mayor parte de las computadoras disponibles (que indican que el mismo es igual a infinito para ellas).

En el cuadro siguiente vemos un ejemplo de uno de los formatos típicos para el genotipado, el formato conocido como **tped**, para 3 individuos. Las primeras 4 columnas corresponden a información sobre cada locus genotipado: 1) el cromosoma, 2) la identificación del SNP, 3) la posición en el genoma en centi-Morgans (usualmente sustituida por un 0 cuando no se conoce) y 4) la posición en el cromosoma, en bases. A partir de la quinta columna en adelante, cada par de columnas se corresponde a un individuo, siendo cada columna la correspondiente a uno de los dos alelos y la otra columna al otro alelo. Por ejemplo, para el primer SNP, su identificación es «snp1», el mismo se encuentra en el cromosoma 1, en la posición 50.123 (respecto a algún genoma de referencia). El primer individuo (columnas 5 y 6) posee el genotipo *AA*, el segundo individuo (columnas 7 y 8) el genotipo *AG* y el tercer individuo (columnas 9 y 10) el genotipo *GG*. Obviamente, los dos alelos posibles son **A** y **G** para este SNP. Para el segundo SNP (cromosoma 1, posición 61.230), los genotipos son *CT*, *CC* y *CT*. De acuerdo con el último SNP que aparece en la tabla, el chip tendría del orden de los 52 mil SNP.

Cromosoma	ID	Pos (cM)	Bases						
1	snp1	0	50123	A	A	A	G	G	G
1	snp2	0	61230	C	T	C	C	C	T
1	snp3	0	97333	G	G	T	T	G	T
...	...	0	...	...	...	...	...	...	...
22	snp52125	0	235243	A	A	A	A	A	G

La pregunta que podemos hacernos ahora es cómo transformar la información de secuencias en cada individuo, los genotipos, en alguna forma numérica que nos sirva para calcular similaridades entre individuos. Como veremos enseguida, existen muchas alternativas posibles para esta conversión, aunque la idea de base es siempre la misma. Una distinción muy importante a conservar en mente es que mientras en el caso de parentesco estadístico estamos hablando siempre de alelos **IBD**, en el caso de parentesco genómico nos vamos a enfocar en alelos **IBS** (idénticos en estado), ya que no conocemos las relaciones entre los individuos y apenas el genotipo observado para cada locus.

Contenido génico

Al observar el cuadro de arriba, posiblemente haya pasado por nuestra mente si no podíamos reducir la información para cada individuo de dos columnas a una sola columna. De hecho, al existir para cada SNP solamente 3 combinaciones posibles, si despreciamos la información de **fase**[2] (que, por otra parte, en el genotipado no disponemos de ella), claramente alcanzaría con colocar un número distinto para cada combinación. Si pensamos un poco más la idea, en lugar de colocar un número arbitrario a cada combinación (por ejemplo, 1 para un homocigoto, 7 para el otro y 9 para el heterocigoto), podríamos pensar en el marco de nuestro modelo genético básico, y el parentesco aditivo. Si el efecto de los

[2] Llamamos **información de fase** a la que nos indica qué alelo de un locus va junto en el cromosoma con qué alelo de otro locus, o de otra forma, la que nos permite reconstruir los haplotipos.

alelos es aditivo, entonces cuando tengo dos alelos del mismo tipo debería tener el doble de efecto. Entonces, si pensamos en el número de copias de uno de los dos alelos elegido en forma arbitraria, por ejemplo, el segundo en orden alfabético, tenemos que para un locus con alelos posibles **A** y **G**, las combinaciones y el número de copias del alelo **G** (el segundo por orden alfabético entre **A** y **G**) estarían dadas por

Cromosoma	Cantidad de copias de G
AA	0
AG	1
GG	2

Con esta información, podemos convertir los genotipos de los 3 individuos para el primer SNP de la tabla con los genotipados en $AA = 0$ (primer individuo), $AG = 1$ (segundo individuo) y $GG = 2$ para el tercer individuo.

Para el segundo SNP tenemos los alelos **C** y **T**, por lo que alfabéticamente el segundo es **T** y el cuadrito resumen de las combinaciones y su valor de contenido génico sería

Genotipo	Cantidad de copias de T
CC	0
CT	1
TT	2

y, por lo tanto, los valores para los 3 individuos serían $CT = 1$ (individuo 1), $CC = 0$ (individuo 2) y $CT = 1$ (individuo 3). Obviamente, como la elección del alelo de referencia fue completamente arbitraria, para marcadores bialélicos el conteo para el otro alelo es igual a dos menos el conteo del alelo de referencia.

Si aplicamos el método anterior a cada SNP de la tabla de genotipado, tenemos ahora que podemos escribirla como

Cromosoma	ID	Pos.cM.	Bases	Individuo 1	Individuo 2	Individuo 3
1	snp1	0	50123	0	1	2
1	snp2	0	61230	1	0	1
1	snp3	0	97333	0	2	1
...	...	0	...	...	...	...
22	snp52125	0	235243	0	0	1

Claramente, además de reducir el número de columnas necesarias para representar cada genotipo, tenemos ahora una representación numérica que se asocia claramente a nuestro modelo genético básico, en particular a los efectos aditivos del mismo. Es más, ahora queda muy claro que podemos descomponer la tabla de genotipado en dos partes, aquella con información sobre los marcadores (columnas 1 a 4) y aquella con información de los genotipos de los individuos (columna 5 en adelante). Con esta segunda parte podemos formar una matriz numérica de **contenido génico**, a la que por comodidad le ponemos nombres de

filas (los identificadores de cada marcador) y de columnas (identificación de individuo), lo que nos deja con algo así

	Individuo 1	Individuo 2	Individuo 3
snp1	0	1	2
snp2	1	0	1
snp3	0	2	1
...	...	...	...
snp52125	0	0	1

Al tratarse de una tabla (matriz) numérica, podemos aplicar sobre ella distintas operaciones, como veremos más adelante. Como cada celda de la tabla es una variable aleatoria de conteo, si llamamos z_i al conteo del alelo de referencia para el individuo i, entonces podemos calcular para el mismo estadísticos resumen, como la media y la varianza. Si asumimos que cada marcador se encuentra en equilibrio de Hardy-Weinberg (algo que es bastante común para la mayor parte de los marcadores) y llamamos A_1 al alelo de referencia y A_2 al alternativo, entonces podemos resumir la información para el cálculo de la media y la varianza en el siguiente cuadro

Genotipo	Frecuencia	z	z^2
A_1A_1	p^2	2	$2^2 = 4$
A_1A_2	$2pq$	1	$1^2 = 1$
A_2A_2	q^2	0	$0^2 = 0$
$\mathbb{E}()$		$2p$	$4p^2 + 2pq$

La esperanza de cada una de las variables aleatorias ($\mathbb{E}(z)$ y $\mathbb{E}(z^2)$) se calculan como de costumbre, sumando los productos de los valores por sus frecuencias correspondientes. Es decir

$$\begin{aligned} \mathbb{E}(z) &= 2 \times p^2 + 1 \times 2pq + 0 \times q^2 = 2p^2 + 2pq = 2p(p+q) = 2p \\ \mathbb{E}(z^2) &= 4 \times p^2 + 1 \times 2pq + 0 \times q^2 = 4p^2 + 2pq \end{aligned} \tag{9.65}$$

De lo anterior surge claramente que la media del contenido génico es igual a $2p$, lo que era de esperar ya que p es la frecuencia del alelo de referencia y en cada genotipo hay dos alelos, por lo que el conteo esperado es dos veces la frecuencia del alelo de referencia. Para calcular la varianza del contenido podemos recurrir a la relación $\mathbb{V}(z) = \mathbb{E}(z^2) - \mathbb{E}(z)^2$, por lo que

$$\mathbb{V}(z) = \mathbb{E}(z^2) - \mathbb{E}(z)^2 = (4p^2 + 2pq) - (2p)^2 = 4p^2 + 2pq - 4p^2 = 2pq \tag{9.66}$$

Es decir, la varianza en contenido génico (el número de alelos de referencia) es igual a $2pq$.

Si recordamos la definición de parentesco aditivo como correlación entre valores de cría de dos individuos, entonces para calcular el coeficiente de correlación vamos a precisar conocer la covarianza primero.

Como vimos previamente, para dos individuos i y j, al extraer un alelo en cada uno de ellos al azar, la probabilidad de que los mismos sean **IBD** es $f_{ij} = \Theta_{ij} = A_{ij}/2$, que es el coeficiente de coancestría de Malécot. Para calcular la covarianza de conteo génico entre dos individuos podemos recurrir a la fórmula de covarianza

$$Cov(z_i,z_j) = \mathbb{E}(z_i z_j) - \mathbb{E}(z_i)\mathbb{E}(z_j) \tag{9.67}$$

Por otro lado, las esperanzas de z_i y z_j son iguales a $\mathbb{E}(z) = 2pq$.

Para calcular la esperanza del producto cruzado $\mathbb{E}(z_i z_j)$ podemos razonar de la siguiente forma: tenemos cuatro formas de elegir un alelo en cada uno de los dos individuos; la única forma en que el producto $z_i z_j$ sea igual a 1 es que z_i sea A_1 y z_j también lo sea. O sea, el primer alelo debe ser A_1, con probabilidad de extracción p y el segundo también, pero como ya vimos antes, esto puede ocurrir por dos razones: que los dos alelos extraídos sean **IBD**, con probabilidad $\Theta_{ij} = A_{ij}/2$, o que los dos alelos no sean **IBD** pero que por azar se extraiga otro alelo A_1 (**IBS**), esta segunda posibilidad entonces con probabilidad $(1 - \Theta_{ij})p = (1 - A_{ij}/2)p$. Multiplicando ambas por la probabilidad p de que el primer alelo sea A_1 y multiplicando por 4 (por las cuatro posibles combinaciones), tenemos

$$\mathbb{E}(z_i z_j) = 4p[\frac{A_{ij}}{2} + \left(1 - \frac{A_{ij}}{2}\right) p] = 2pA_{ij} + 4p^2 - 2p^2 A_{ij} \therefore \tag{9.68}$$

$$\mathbb{E}(z_i z_j) = 2A_{ij}(p - p^2) + 4p^2 = 2p(1-p)A_{ij} + 4p^2 \therefore \tag{9.69}$$

$$\mathbb{E}(z_i z_j) = 2pqA_{ij} + 4p^2 \tag{9.70}$$

Sustituyendo el resultado de la ecuación (9.70) en la ecuación (9.67) tenemos

$$\begin{aligned} Cov(z_i,z_j) = \mathbb{E}(z_i z_j) - \mathbb{E}(z_i)\mathbb{E}(z_j) = 2pqA_{ij} + 4p^2 - [(2p)(2p)] \therefore \\ Cov(z_i,z_j) = 2pqA_{ij} + 4p^2 - 4p^2 = 2pqA_{ij} \end{aligned} \tag{9.71}$$

Este es el resultado de la covarianza en contenido génico entre dos individuos para un locus determinado. La correlación entre los contenidos génicos de los individuos i y j es igual a la covarianza entre z_i y z_j dividida entre el producto de las raíces cuadradas de las varianzas, es decir $\sqrt{\mathbb{V}(z_i)\mathbb{V}(z_j)} = \sqrt{(2pq)(2pq)} = \sqrt{(2pq)^2} = 2pq$, por lo que, poniendo todo junto, tenemos

$$\begin{aligned} r_{ij} = \frac{Cov(z_i,z_j)}{\sqrt{\mathbb{V}(z_i)\mathbb{V}(z_j)}} = \frac{2pqA_{ij}}{2pq} \therefore \\ r_{ij} = A_{ij} \end{aligned} \tag{9.72}$$

o lo que es lo mismo, el valor esperado de la correlación de contenido génico en dos individuos es igual al parentesco aditivo entre los mismos.

Extender el resultado previo de la ecuación (9.71) a la covarianza en contenido génico entre más individuos es inmediato si entendemos que A_{ij} es un elemento de la matriz de

parentesco aditivo entre individuos ($\mathbf{A}$). Si ahora utilizamos la notación vectorial/matricial, entonces los resultados previos pueden resumirse como

$$\begin{aligned} \mathbb{E}(\mathbf{z}) &= \mathbf{2}p \\ \mathbb{V}(\mathbf{z}) &= 2pq\mathbf{A} \end{aligned} \tag{9.73}$$

La matriz de similaridad genómica

En un artículo notable, VanRaden (2008) plantea tres alternativas para representar la similaridad (parentesco) genómica entre individuos, de las que solo consideraremos la primera. Es claro que la matriz de conteo que vimos previamente y que en este trabajo se nombra $\mathbf{M}$ es el elemento clave para determinar la similaridad entre individuos. Sin embargo, de la métrica elegida dependerán las propiedades de la matriz genómica. En lo que sigue asumiremos que la matriz $\mathbf{M}$ posee n filas representando los individuos y m columnas representando los marcadores, por lo que sería la matriz transpuesta respecto a como la habíamos visto previamente. Por ejemplo, si en lugar de utilizar los valores $\{0,1,2\}$ para los tres genotipos utilizamos $\{-1,0,1\}$ los elementos de la diagonal de $[\mathbf{MM}']_{\mathbf{nxn}}$ representan el número de loci homocigotos para cada individuo, mientras que los elementos fuera de la diagonal representan el número de alelos compartidos menos m.

Consideremos ahora la matriz $\mathbf{M}$ tomando los valores de los genotipos como $\{0,1,2\}$. La matriz $[\mathbf{MM}']_{\mathbf{nxn}}$ es una matriz cuadrada de $n \times n$ elementos. Para cada par de individuos i_1 y i_2 esta matriz contiene el resultado de $\sum_{j=1}^{m} z_{i_1,j} z_{i_2,j}$. Claramente, con esta codificación, si todos los loci en los dos individuos son homocigotos para el alelo alternativo, entonces $\sum_{j=1}^{m} z_{i_1,j} z_{i_2,j} = \sum_{j=1}^{m} 0 \times 0 = 0$. En cambio, si ambos son completamente homocigotos para el alelo de referencia, entonces $\sum_{j=1}^{m} z_{i_1,j} z_{i_2,j} = \sum_{j=1}^{m} 2 \times 2 = 4m$. Sin embargo, ambas situaciones deberían indicar exactamente lo mismo ya que la elección de qué alelo es el de referencia y cuál el alternativo es algo totalmente arbitrario, sin significado biológico hasta ahora. Por otro lado, es claro que con este producto matricial la similaridad entre individuos es función directa del número de marcadores, lo que no parece una buena idea (por ejemplo, en el caso anterior $4m$), por lo que deberíamos considerar su estandarización.

La alternativa propuesta por VanRaden (2008) consiste en transformar los conteos génicos restándole el doble de la frecuencia media del alelo de referencia en cada locus y luego dividiendo entre la suma de varianzas de conteo de todos los loci. Para eso, si p_j es la **frecuencia de base** del alelo de referencia en el locus j, vamos a definir una matriz $\mathbf{P}$ cuyas columnas están dadas por

$$\mathbf{P}_{\cdot,j} = 2(p_j) \tag{9.74}$$

Es decir, cada columna j de $\mathbf{P}$ contiene exactamente n elementos idénticos a $2p_j$. A partir de esto, basta sustraer esta matriz de la matriz de conteo para obtener una matriz $\mathbf{Z}$

$$\mathbf{Z} = \mathbf{M} - \mathbf{P} \tag{9.75}$$

La matriz $\mathbf{Z}$ tendrá los conteos, pero ahora «centrados» de alguna manera, tal que la $\mathbb{E}(Z_{\cdot,j}) = 0$, es decir, la esperanza de cada marcador en esta matriz es cero. Esta resta, además de centrar pone el énfasis en aquellos alelos raros, más que en los comunes. Supongamos, por ejemplo, que para un marcador determinado j la frecuencia

del alelo de referencia es 0,99. Entonces, mientras que para los homocigotos A_1A_1 el valor de $Z_{.,j} = 2 - (2 \times 0{,}99) = 0{,}02$, el otro homocigoto ($A_2A_2$) tendrá un valor de $Z_{.,j} = 0 - (2 \times 0{,}99) = 1{,}98$, cuyo valor absoluto es 99 veces mayor que el valor absoluto del A_1A_1. Ahora podemos hacer con comodidad el producto de $\mathbf{ZZ}'$ como un buen indicador de similaridad entre individuos, aunque aún dependiente del número de loci.

Finalmente, el último paso consiste en dividir el producto $\mathbf{ZZ}'$ entre la suma de las varianzas en conteo para todos los loci (ecuación (9.66)),

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{m}(p_i)(1-p_i)} \tag{9.76}$$

De acuerdo con VanRaden (2008), con la estandarización propuesta el coeficiente de consanguinidad observado de un individuo i es igual a $F_i = G_{ii} - 1$ y las relaciones de parentesco genómico entre dos individuos i y j, análogas al coeficiente de parentesco de Sewall Wright (1922) que ya vimos en la ecuación (9.30), se obtienen como $\frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}}$.

Ejemplo 9.6

Cuatro individuos fueron secuenciados para los 8 marcadores siguientes:

Marcador	Alelo 1	Alelo 2
snp1	C	G
snp2	A	C
snp3	C	T
snp4	G	T
snp5	A	G
snp6	A	T
snp7	C	T
snp8	A	C

La tabla de genotipado, con la información relevante para el cálculo de la similaridad genómica para los 4 individuos (en formato **tped**, excluyendo los datos de posicionamiento de cada marcador) fue la siguiente:

Marcador								
snp1	G	G	G	G	G	G	G	G
snp2	C	C	C	C	C	C	C	C
snp3	C	T	C	T	T	T	T	T
snp4	G	T	G	G	G	T	G	T
snp5	A	A	A	G	A	G	A	G
snp6	T	T	A	A	A	A	A	T
snp7	T	T	T	T	T	T	C	T
snp8	A	C	A	C	C	C	A	C

Por otra parte, las **frecuencias de base** de los alelos de referencia (el segundo por orden alfabético) para cada SNP es la siguiente: snp1 = 0,45; snp2 = 0,47; snp3 = 0,47; snp4 = 0,5; snp5 = 0,5; snp6 = 0,75; snp7 = 0,75; snp8 = 0,65

A partir de la información provista, calcular la matriz de similaridad genómica de acuerdo con VanRaden (2008). Calcular para cada uno de los individuos el coeficiente de consanguinidad genómico (observado). ¿Qué individuos se encuentran más emparentados, de acuerdo con VanRaden (2008) (siguiendo a Sewall Wright 1922) usando la matriz **G**?

El primer paso consiste en pasar la tabla de genotipado a una matriz de conteo alélico. Para eso debemos recordar que cada dos columnas indican el genotipo de un individuo. Utilizando la regla arbitraria de contar el número de copias del segundo alelo en orden alfabético (para los que tenemos a su vez las frecuencias de los alelos), contando en cada par de columnas y transponiendo obtenemos la siguiente matriz de conteo **M**:

Individuo	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
Indiv 1	2	2	1	1	0	2	2	1
Indiv 2	2	2	1	0	1	0	2	1
Indiv 3	2	2	2	1	1	0	2	2
Indiv 4	2	2	2	1	1	1	1	1

Para armar la matriz **P** alcanza con repetir tantas veces como individuos tengamos los valores de las **frecuencias de base** de los alelos **multiplicadas por** 2, es decir

Individuo	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
Indiv 1	0,90	0,94	0,94	1,00	1,00	1,50	1,50	1,30
Indiv 2	0,90	0,94	0,94	1,00	1,00	1,50	1,50	1,30
Indiv 3	0,90	0,94	0,94	1,00	1,00	1,50	1,50	1,30
Indiv 4	0,90	0,94	0,94	1,00	1,00	1,50	1,50	1,30

Ahora, utilizando la ecuación (9.75), calculamos $\mathbf{Z} = \mathbf{M} - \mathbf{P}$

Individuo	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
Individuo 1	1,10	1,06	0,06	0,00	−1,00	0,50	0,50	−0,30
Individuo 2	1,10	1,06	0,06	−1,00	0,00	−1,50	0,50	−0,30
Individuo 3	1,10	1,06	1,06	0,00	0,00	−1,50	0,50	0,70
Individuo 4	1,10	1,06	1,06	0,00	0,00	−0,50	−0,50	−0,30

El cálculo de $\mathbf{ZZ}'$ no es difícil (ver «Apéndice A: Conceptos matemáticos básicos» para refrescar como se realiza la multiplicación de matrices) pero algo tedioso de hacer a mano, por lo que nosotros utilizamos el software estadístico **R** para realizarla, pero también puedes hacer esta operación utilizando páginas web que lo realizan.

El resultado de la multiplicación $\mathbf{ZZ}'$ es

	Individuo 1	Individuo 2	Individuo 3	Individuo 4
Individuo 1	3,9272	1,9272	1,6872	1,9872
Individuo 2	1,9272	5,9272	4,6872	2,9872
Individuo 3	1,6872	4,6872	6,4472	3,7472
Individuo 4	1,9872	2,9872	3,7472	4,0472

Notemos que la matriz $\mathbf{ZZ}'$ es una matriz cuadrada y simétrica, de dimensiones $n \times n$ (con n el número de individuos). Finalmente, la matriz de similaridad genómica se calcula de acuerdo con la ecuación (9.76), es decir

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{2\sum_{i=1}^{m}(p_i)(1-p_i)} \tag{9.77}$$

Por lo tanto, necesitamos calcular el denominador de la misma, que teniendo en cuenta los valores de las frecuencias de base de los SNPs es $2\sum_{i=1}^{m}(p_i)(1-p_i) = 3{,}6964$. Dividiendo $\mathbf{ZZ}'$ entre este último valor tenemos la matriz $\mathbf{G}$ deseada

	Individuo 1	Individuo 2	Individuo 3	Individuo 4
Individuo 1	1,0624391	0,5213721	0,4564441	0,5376042
Individuo 2	0,5213721	1,6035061	1,2680446	0,8081376
Individuo 3	0,4564441	1,2680446	1,7441835	1,0137431
Individuo 4	0,5376042	0,8081376	1,0137431	1,0949031

Para el cálculo de los coeficientes de consanguinidad de los individuos recordamos que $F_i = G_{ii} - 1$, por lo que extraemos los valores de la diagonal en $\mathbf{G}$ y les restamos 1 a cada uno (o lo que es lo mismo $\mathbf{F} = \mathbf{G} - \mathbf{1}$). Esto nos deja con

	G_{ii}	F_i
Individuo 1	1,0624391	0,0624391
Individuo 2	1,6035061	0,6035061
Individuo 3	1,7441835	0,7441835
Individuo 4	1,0949031	0,0949031

Para determinar qué individuos se encuentran más emparentados de acuerdo con la matriz de parentesco genómico y VanRaden (2008) calculamos $\frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}}$ para cada celda fuera de la diagonal y observamos qué valor es el mayor.

La matriz resultante de este cálculo es

	Individuo 1	Individuo 2	Individuo 3	Individuo 4
Individuo 1	–	0,3994482	0,3353046	0,4984513
Individuo 2	0,3994482	–	0,7582338	0,6099050
Individuo 3	0,3353046	0,7582338	–	0,7335741
Individuo 4	0,4984513	0,6099050	0,7335741	–

por lo que el parentesco genómico mayor sería entre el **individuo 3** y el **individuo 2.**

PARA RECORDAR

- Una forma de ver la relación entre el **parentesco estadístico** y el **genómico** es la de considerar al primero como el parentesco esperado y al segundo como el parentesco realizado (u observado).
- Mientras que en el **parentesco estadístico** estamos siempre hablando de alelos **IBD**, en el caso de parentesco genómico nos vamos a enfocar en alelos **IBS** (idénticos en estado), ya que no conocemos las relaciones entre los individuos y apenas el genotipo observado para cada locus.
- El valor esperado de la correlación de contenido génico en dos individuos es igual al parentesco aditivo entre los mismos, tal que

$$r_{ij} = \frac{Cov(z_i, z_j)}{\sqrt{\mathbb{V}(z_i)\mathbb{V}(z_j)}} = \frac{2pqA_{ij}}{2pq} r_{ij} = A_{ij} \tag{9.78}$$

9.7 Estudios de ascendencia (ancestría)[3]

TOPICO AVANZADO

Las tecnologías de genotipado masivo, así como las de secuenciación masiva han permitido extender los estudios de relaciones entre individuos y poblaciones al nivel comercial, por lo que en la actualidad resulta fácil y económico conocer nuestros «orígenes». Sin embargo, la interpretación de los resultados de dichos estudios no es tan trivial como parece y debemos ser bastante cautelosos a la hora de interpretarlos.

Ascendencia (correspondiente a la palabra inglesa *ancestry*), puede definirse como el conjunto de los ancestros para un individuo, un grupo o un linaje dados. En el presente contexto nos referiremos a ella como la proporción de alelos (variantes genéticas) de un individuo que corresponden a diferentes grupos étnicos ancestrales.

[3] La siguiente sección fue inicialmente escrita para explicar los estudios de ascendencia en humanos y utilizada por la empresa **genLives** (https://wwww.genlives.com) para facilitar la comprensión del alcance de los mismos. Aquí es usada y modificada con conocimiento y autorización de dicha empresa.

Si bien la información contenida en el ADN de un individuo es relativamente inmutable a lo largo de la vida (con pequeñas variaciones debidas a mutaciones somáticas o a errores en el proceso de lectura de la secuencia), el proceso de asignación de las fracciones correspondiente a distintos grupos genéticos (o etnias) no lo es. Para entender la razón de esto, es necesario entender cómo funciona el proceso de transmisión de información genética a lo largo de las generaciones. Para ello usaremos una simple analogía que simplifica muchísimo el complejo proceso por detrás, eliminando varios detalles que por ahora no nos serán necesarios.

Supongamos que podemos representar la información del ADN de cada individuo en una población con un par de cajas que cada una contiene 50 bolitas de colores. Estás cajas vienen a representar un juego de cromosomas, heredado uno del padre (una caja) y el otro de la madre (la otra caja). Cada bolita correspondería entonces a la información en un locus (alelo) en cada cromosoma paterno o materno. Ahora, supongamos que al comienzo del experimento todas las bolitas de todas las cajas son rojas. Nuestra población es homogénea genéticamente. En nuestro modelo muy simplificado, para crear un nuevo individuo (un hijo), elegimos un padre y una madre, cada uno de los cuales aporta la copia de una caja (o sea, elige las mismas bolitas que tiene en una de las cajas de un reservorio con infinitas bolitas de todos los colores). Cada progenitor (padre o madre) aporta la mitad de las bolitas que tendrá el nuevo individuo, es decir, le aporta la mitad de su ADN, que a su vez será la mitad del ADN del nuevo individuo. Como todos los individuos solo tienen bolitas rojas, sus hijos solo tendrán información roja.

Ahora, supongamos que por error al copiar la información para hacer un nuevo individuo, uno de los padres introduce una pelota verde en su caja copia. En cada proceso de copia hay una probabilidad mayor que cero de cometer un error. Si la tasa de error es de 2 % (1 en 50), en promedio en cada copia el padre introducirá una pelota de color diferente al original en la caja. A esto, en genética le llamamos una mutación. Esto hace que el individuo hijo reciba una caja de 50 bolitas rojas y otra con 49 rojas y una verde. En promedio, considerando que este hijo tiene muchos descendientes, la mitad de estos recibirá una caja con una pelota verde y 49 rojas, mientras que la otra mitad recibirá una caja con todas rojas.

En realidad, el proceso de segregación genética (elegir la mitad del material genético del individuo para transmitirlo a la siguiente generación) es un poco más complejo que elegir una de las dos cajas para copiar su contenido. Existe un proceso llamado recombinación, que lo podemos ejemplificar de la siguiente manera: elegimos una bolita de cada caja y las retiramos. Luego, al azar elegimos una de las dos para copiarla en la caja del hijo. Repetimos el proceso hasta que no queden más bolitas en las cajas de los padres (50 veces en nuestro ejemplo). Obviamente debemos devolverle al padre sus dos cajas con las bolitas, tal cual las tenía antes de comenzar el proceso. De esta forma, la caja que el padre armó para su hijo es una combinación del contenido de sus dos cajas.

Supongamos ahora que repetimos el procedimiento de construcción de hijos y luego de varias generaciones muestreamos dos individuos. Uno de ellos, que llamaremos **I1** tiene una caja con 35 bolitas rojas, 5 azules, 5 verdes y 5 amarillas, mientras que la otra caja tiene 30 rojas, 6 azules, 7 verdes y 7 amarillas. El otro individuo, **I2**, tiene 47 bolitas rojas en una caja, dos azules y una amarilla, mientras que en la otra caja tiene 49 rojas y una amarilla. Podríamos suponer incluso que los individuos que tienen una pelota verde en esta población confunden más a menudo los colores que copian, lo que explica en cierta forma la diferencia en el número de bolitas «no rojas» en las 4 cajas.

Hagamos algunos cálculos en este momento. Si no supiéramos de dónde vinieron estos individuos, pero suponemos que en algún momento todas las poblaciones que existían eran de un solo color, cualquiera fuera este, diríamos que los dos individuos tienen diferentes proporciones de genes (bolitas) de las poblaciones originales o ancestrales (a veces lo escuchamos como porcentaje de sangre, por más que no tenga nada que ver con la realidad biológica). Para **I1** el cálculo nos daría

$35+30=65$ rojas, en 100 bolitas totales, o 65 % de genes de una población roja $5+6=11$ azules, en 100 bolitas totales, o 11 % de genes de una población azul $5+7=12$ verdes, en 100 bolitas totales, o 12 % de genes de una población verdes $5+7=12$ amarillas, en 100 bolitas totales, o 12 % de genes de una población amarillas.

Además, $65\,\% + 11\,\% + 12\,\% + 12\,\% = 100\,\% = 1$ (el total de los genes del individuo fueron asignados a alguna población ancestral).

Por otra parte, para **I2** el cálculo nos daría

$47+49=96$ rojas, en 100 bolitas totales, o 96 % de genes de una población roja $2+0=2$ azules, en 100 bolitas totales, o 2 % de genes de una población azul $1+1=2$ amarillas, en 100 bolitas totales, o 2 % de genes de una población amarillas.

$96\,\% + 2\,\% + 2\,\% = 100\,\% = 1$ (el total de los genes del individuo fueron asignados a alguna población ancestral).

Pese a que estos dos individuos los muestreamos de la misma población, si no conocemos cuántas poblaciones ancestrales realmente existían y confiamos que cada color se correspondía con una población homogénea para ese color (despreciamos la mutación como el mecanismo que introduce la variabilidad), entonces llegaremos a la conclusión de que el individuo **I1** tiene un 65 % de sus ancestros de la población roja, mientras que el **I2** tiene un 96 % de ancestros originarios de esa población. A su vez, arribamos a la conclusión de que **I1** tiene 11 % de ancestros azules y 12 % de cada uno de amarillos y verdes, mientras que **I2** solo tiene 2 % de ancestros azules y amarillos. Lo que usamos en este proceso fue asumir que las diferencias en colores se corresponden enteramente a poblaciones originales diferentes, lo que es equivalente a asumir que el mestizaje y no la mutación explican enteramente el proceso.

Pero supongamos que ahora tenemos información «histórica» que refiere que originalmente, solo existían dos poblaciones ancestrales, una **roja** y otra **azul**. Ahora nuestro modelo puede ser un poco más complejo. Vamos a asignar la variabilidad roja y azul a esas poblaciones ancestrales y vamos a suponer que verde y amarillo son en realidad mutaciones. Eso nos llevará a que para **I1** tenemos

- $35 + 30 = 65$ rojas, en 76 bolitas no verdes ni amarillas, o 85,5 % de genes de una población roja
- $5 + 6 = 11$ azules, en 76 bolitas no verdes ni amarillas, o 14,5 % de genes de una población azul

Como verificación, $85{,}5\,\% + 14{,}5\,\% = 100\,\% = 1$. Esto es, solo pudimos asignar 76 bolitas, de las 100 que tenemos, a alguna de las poblaciones ancestrales. Además, tenemos $(5+7=12)$ **verdes** + $(5+7=12)$ **amarillas**, en 100 bolitas totales (es decir, 24 % de «genes» mutados que no pudimos asignar).

Para el otro individuo, el cálculo nos daría

- 96 rojas, en 98 bolitas no verdes ni amarillas, o $\approx 98\,\%$ de genes de población roja
- 2 azules, en 98 bolitas no verdes ni amarillas, o $\approx 2\,\%$ de genes de población azul.
- Además $(1 + 1 = 2)$ amarillas en 100 bolitas totales, o $2\,\%$ de genes mutados.

De esto, es fácil inferir que la proporción de genes que corresponden a cada población ancestral depende del número de poblaciones ancestrales que coloquemos en el modelo. Usualmente, cuantas menos poblaciones ancestrales asumamos, mayor porcentaje de lo asignado le corresponderá a cada una. Además, es muy importante entender los procesos biológicos que existen por detrás de la variabilidad, así como sus parámetros, ya que de lo contrario nuestras conclusiones podrían ser completamente equivocadas (como en nuestro ejemplo, que asignamos la variabilidad a la existencia de distintas poblaciones ancestrales, cuando solo existía una).

Como dijimos antes, nuestro primer modelo era extremadamente simple y es tiempo de agregarle un poco más de realismo. Normalmente, aún los grupos étnicos más cerrados suelen tener algún contacto con sus vecinos, lo que produce que algunos hijos no reciban dos cajas de bolitas del mismo color que la población en las que les tocará vivir. Es más, las poblaciones vecinas suelen compartir una parte muy importante de los colores de sus cajas, ya que si nos remontamos un poco en el tiempo probablemente las dos surgieron de una población más ancestral (no hay que olvidarse que originalmente solo existía una población de *Homo sapiens*, que habitaba en África).

Supongamos entonces que tenemos dos poblaciones ancestrales vecinas **A** y **B**, la primera cuyas cajas tienen en promedio 70 bolitas rojas y 30 azules, mientras que en la segunda poseen en promedio 50 bolitas rojas, 40 azules y 10 amarillas. Dentro de cada población sin duda existirá variación a nivel individual. Es decir, en la población **A** nos podemos encontrar con individuos **60r+40az**, **80r+20az** y hasta algún extraño (e insignificante en número) **70r+29az+1am**, etcétera. Lo que cuenta es que el promedio de la población es, como dijimos **70r+30az**. Supongamos ahora, que tenemos un individuo que fue encontrado en algún lugar en el espacio geográfico de las dos poblaciones. Cuando analizamos su ADN, nos encontramos que posee la fórmula **60r+35az+5am**. ¿A cuál de estas dos poblaciones pertenece, si es que pertenece a alguna? Si fuese el producto de una cruza entre individuos de las poblaciones **A** y **B**, ¿cuántas generaciones atrás ocurrió ese mestizaje?

Lo primero que se nos viene a la mente, teniendo en cuenta que solo en una de las poblaciones (**B**) la frecuencia de las bolitas (alelos) amarillas es significativa y que este individuo tiene una proporción no despreciable de las mismas, es que este individuo probablemente venga de dicha población **B**. Como comentamos antes, si bien la frecuencia promedio de una población es única, existe variabilidad individual en torno a la misma y las frecuencias que observamos en este individuo desconocido no parece alejarse demasiado de las frecuencias promedios de **B**. Más exactamente, dada la forma de muestreo de las bolitas, podemos asumir razonablemente que la distribución del número de cada color sigue una ley **multinomial** (una versión de la conocida binomial, para más de dos categorías). Pero dejemos, por ahora estos cálculos para otro momento.

En realidad, si bien el razonamiento anterior no parece desacertado, no es la única alternativa razonable. De hecho, posiblemente tampoco sea la más parsimoniosa, si consideramos que el individuo puede ser el producto (descendiente) de un cruzamiento entre individuos de las poblaciones parentales **A** y **B**. Supongamos que un **labrador** de la población **A** (población que tiene en promedio la fórmula **70r+30az**) se encuentra con una **pastora** de la

población **B** (población que tiene en promedio la fórmula **50r+40az+10am**), que está llevando a pastar a sus **ovejas verdes** en los límites de las tierras de la población **B** y deciden (romance de por medio) establecerse en esas nuevas tierras limítrofes. Simplemente por casualidad, ambos son individuos muy representativos de sus poblaciones, es más, las frecuencias de bolitas en sus cajas se corresponden exactamente con los promedios de cada población. La esperanza matemática es que pasen a sus hijos la mitad de sus bolitas, es decir, el **Labrador** (**A**) transmitirá **35r+15az**, mientras que la **Pastora** (**B**) pasará **25r+20az+5am** (en promedio).

Un hijo promedio de ambos tendrá entonces la suma, **60r+35az+5am**, lo que es idéntico a la composición de las cajas de nuestro individuo desconocido. Más aún, las dos alternativas que manejamos antes para explicar el posible origen genético de nuestro individuo (que provenga de una de las dos poblaciones o que sea el producto del apareamiento entre un individuo de una y el otro de la otra población) no son más que unos casos particulares de una solución general donde las proporciones de cada una de las poblaciones varían entre 0 y 1, con la restricción de que la suma de las proporciones debe dar 1 (100 % del origen identificado). Por ejemplo, un individuo producto del apareamiento de un individuo de la población **B** y un individuo hijo de un individuo de **A** y otro de **B**, tendrá 0,25 de **A** y 0,75 de **B** (hablaremos de esto más adelante en el capítulo «Endocría, exocría, consanguinidad y depresión endogámica»). A medida que pasa el tiempo, en generaciones de individuos, mayor el espectro real de proporciones de mezcla posibles y mayor la dificultad de reconocer con claridad el verdadero origen genético de los individuos.

Podemos preguntarnos a esta altura, ¿de dónde vino tanta complejidad? De hecho, si las poblaciones fueran todas homogéneas y de colores diferentes entre ellas, distinguir las proporciones de las diferentes poblaciones en un individuo cualquiera sería una tarea trivial: contamos cuantas bolitas de cada color hay en las cajas de este individuo y las dividimos entre el total. Eso no dará la frecuencia de cada población (si multiplicamos los resultados por 100 los tendremos en porcentaje). La complejidad aparece cuando las poblaciones que consideramos como referencia tienen variabilidad interna, que hace que se solapen los colores que tienen. Cuanto mayor el solapamiento, es decir mayor similitud en las frecuencias de los colores de las bolitas, mayor la dificultad para discernir en forma clara las proporciones de cada una en un nuevo individuo.

Consideremos, como ejemplo dos poblaciones **X** e **Y**, en dos escenarios: en el primer escenario, **X** tiene una frecuencia de rojas de 0,99 y 0,01 de azules, mientras que **Y** tiene 0,01 de rojas y 0,99 de azules (asumamos que todos los individuos son idénticos entre ellos). Ahora, en este escenario, si encontramos un individuo cualquiera en el mundo y tenemos que asignar qué proporción tiene de cada una de las poblaciones X e Y, la situación será muy fácil. Simplificando mínimamente, podemos asignar todas las bolitas rojas en el nuevo individuo como provenientes de **X** y todas las azules como provenientes de **Y** y tendremos las proporciones deseadas.

Ahora, en el segundo escenario, **X** tiene una frecuencia de rojas de 0,99 y 0,01 de azules, mientras que **Y** tiene 0,97 de rojas y 0,03 de azules (asumamos que todos los individuos son idénticos entre ellos). En este segundo escenario, aplicar la estrategia anterior no tiene sentido ya que la diferencia de frecuencias entre las poblaciones es mínima y solo podremos conjeturar, con enorme probabilidad de equivocarnos que si el individuo tiene más de 0,98 de rojas es más probable que venga de **X**, y si tiene menos de ese valor que provenga de **Y**.

El asunto aquí es entonces el solapamiento de frecuencias (de colores) entre las distintas poblaciones. Una alternativa muy interesante, es cambiar nuestra definición de poblaciones de referencia. Es decir, si en lugar de considerar como poblaciones de referencia **X** e **Y** (en cualquiera de los escenarios), consideramos que las mismas son a su vez producto de mestizaje entre las verdaderas poblaciones ancestrales, la roja y azul (que por definición, serán ahora completamente homogéneas). Esto simplifica enormemente el problema de determinar las frecuencias, ya que nuevamente solo se trata de contar en el nuevo individuo la proporción de las bolitas de cada uno de los colores. ¿Cuál es el problema entonces de esta aproximación en la realidad?

Consideremos en primer lugar lo que sucedería si en lugar de dos colores tenemos una paleta 80 millones de colores reales, es decir que tenemos algún individuo que tiene al menos una bolita con ese color (en realidad la paleta teórica es mucho mayor, del orden de los billones a cifras astronómicas, dependiendo de cómo consideremos los colores). Ahora el problema no aparece con la heterogeneidad de las poblaciones, sino en cómo identificamos los colores. Nuestra capacidad de distinguir colores está claramente por debajo de estas cifras, lo que nos imposibilitaría distinguir con seguridad las poblaciones posibles. Afortunadamente, también sabemos que las posibles poblaciones humanas que existen o existieron en nuestro planeta están claramente por debajo de este número. Digamos que tal vez podemos agrupar colores por similitud, de tal forma que nos quedemos con un número manejable de los mismos, pongamos, por ejemplo, unos 25 y nos quedamos con un representante de cada grupo como «prototipo» de dicho grupo. Ahora, dado un individuo cualquiera del que queremos saber las proporciones de las poblaciones «originales» el problema se trata simplemente de asignar las bolitas con colores reales del mismo a las categorías de los **colores prototípicos** y contar cuántas entran en cada uno.

El proceso real de asignación de proporciones de ascendencia, en alguna de sus versiones, es bastante parecido a lo que describimos hasta acá. Se trata en una primera etapa de identificar un conjunto de poblaciones, que llamaremos de referencia, a las que se le realiza el genotipado (determinación de las variantes en el ADN) de varios individuos representativos de cada una. El genotipado es un proceso caro, si bien su costo disminuye de forma continua, por lo que el número de poblaciones para las que se cuenta con esta información es (era) relativamente bajo. Si bien, idealmente, las poblaciones que integran el conjunto debieran ser seleccionadas para representar de la mejor manera posible la diversidad genética humana del planeta entero, esto dista muchísimo de ser la realidad. Por ejemplo, en el proyecto «1000 genomes» (1000 genomas, aunque superaron los 2 mil), posiblemente el más exhaustivo en cobertura del genoma y número de poblaciones (26), existe una clara subrepresentación de los nativos americanos. Además, este panel de poblaciones está integrado en muchos casos por poblaciones que pretenden representar un grupo étnico al que ya no pertenecen geográfica e históricamente, simplemente porque es mucho más fácil de extraer las muestras en ellos. Por ejemplo, los centro-europeos del panel (**CEU**), son residentes del estado de Utah en EE. UU., con ascendencia del norte y oeste de Europa. En otros casos, las poblaciones sobre las que se muestrea presentan un alto grado de mestizaje con otros grupos completamente diferentes en la evolución humana, lo que hace muy difícil de interpretar los resultados cuando aparece una proporción significativa de estas poblaciones de referencia. Un ejemplo es el grupo **MXL**, que podría intentar interpretarse como nativos de México, pero que solo representa individuos con algo de ascendencia mexicana, viviendo en el área de Los Ángeles (EE. UU.). En la tabla que sigue se observan las poblaciones de referencia que hay disponibles en proyecto «1000 genomes» (ver tabla más abajo). Para cada una de las poblaciones del proyecto «1000 genomes» se realizó el secuenciado

de varios individuos (del orden de los 100 por población) y se determinó por lo tanto que estado (base o nucleótido) existían en cada individuo. A partir de ahí es fácil obtener una estimación para cada población de cuál es la frecuencia de cada base (A, C, G, T) en cada posición del genoma. Como usualmente solo consideramos marcadores dimórficos (que en las poblaciones solo hay dos variantes), esta tabla tiene tantas filas como marcadores dimórficos y tantas columnas como poblaciones, con cada celda conteniendo el valor del alelo de referencia. La segunda parte del proceso de determinar los porcentajes de ascendencia de un individuo consiste en obtener su información genética. Para esto existen varias alternativas; las más simples y baratas consisten en genotipar (identificar variantes) un juego muy reducido y definido previamente de marcadores (posiciones en el genoma). Los marcadores más usuales para ascendencia son del tipo polimorfismos de un solo nucleótido (SNP en inglés). Las alternativas de mayor complejidad consisten en la secuenciación del exoma o aun del genoma completo de un individuo. Estos últimos enfoques, además de estimar las frecuencias de ascendencia, permiten obtener valiosa información médica, por lo que, al bajar el costo de secuenciación de forma dramática, cada vez son más usados. Si bien las aproximaciones de secuenciación masiva permiten distinguir diferentes tipos de variaciones, no solo SNP, dada la enorme cantidad que se obtienen con estas técnicas y la simplicidad de su comparación, suelen ser el conjunto de información más usado para ascendencia.

Luego de que tenemos un conjunto de poblaciones de referencia y que tenemos la información genómica del individuo (marcadores), el proceso consiste en intentar determinar qué proporciones de cada una de esas poblaciones de referencia mejor explican el genotipo del individuo. Una de las estrategias para esto, eficiente y suficientemente rápida, es considerar cada marcador como una pieza independiente de información (una simplificación sobre la que abundaremos más adelante) y utilizar el enfoque estadístico de máxima verosimilitud. Básicamente, el enfoque de **máxima verosimilitud** consiste en describir una función de verosimilitud, que se relaciona con la probabilidad de observar los datos dados diferentes valores de los parámetros del modelo, y hallar el máximo de esta función. En nuestro caso, los parámetros del modelo serían las distintas proporciones de cada una de las poblaciones de referencia y, por lo tanto, lo que buscamos es cuál combinación de los mismos permiten maximizar la probabilidad de observar el conjunto de marcadores que observamos en el individuo. Para hallar este máximo, existen algoritmos eficientes que nos garantizan la solución óptima si se cumplen una serie de requerimientos.

Tabla 9.1. Códigos poblacionales. Grupos: AFR, Africanos; AMR, Mestizos Americanos; EAS, Asiáticos del este; EUR, Europeos; SAS, Asiáticos del Sur

Código	Descripción	Grupo
CHB	Han Chinese en Bejing, China	EAS
JPT	Japoneses en Tokio, Japón	EAS
CHS	Han Chinese del sur	EAS
CDX	Chinese Dai en Xishuangbanna, China	EAS
KHV	Kinh en ciudad Ho Chi Minh, Vietnam	EAS
CEU	Residentes de Utah (CEPH) con ascendencia de Europa del norte y el oeste	EUR
TSI	Toscanos en Italia	EUR
FIN	Fineses en Finlandia	EUR
GBR	Británicos en Inglaterra y Escocia	EUR
IBS	Iberian Population in Spain	EUR
YRI	Yoruba en Ibadan, Nigeria	AFR
LWK	Luhya en Webuye, Kenia	AFR
GWD	Gambian en Divisiones Oeste en Gambia	AFR
MSL	Mende en Sierra Leona	AFR
ESN	Esan en Nigeria	AFR
ASW	Americanos de ascendencia Africana en el suroeste de EE. UU.	AFR
ACB	Caribeños Africanos en Barbados	AFR
MXL	Ascendencia Mexicana en Los Ángeles EE. UU.	AMR
PUR	Puertorriqueños de Puerto Rico	AMR
CLM	Colombianos de Medellín, Colombia	AMR
PEL	Peruanos de Lima, Perú	AMR
GIH	Gujarati Indian de Houston, Texas	SAS
PJL	Punjabi de Lahore, Pakistán	SAS
BE	Bengalíes de Bangladesh	SAS
STU	Tamiles (Sri Lanka) del Reino Unido	SAS
ITU	Indian Telugu del Reino Unido	SAS

¿Cuáles son los principales problemas con esta aproximación tan simple y elegante? Primero, las que tienen que ver con las violaciones a lo asumido por el algoritmo. En particular, la no independencia de las observaciones, algo que está asociado al proceso de muestreo genético, el desequilibrio de ligamiento. Desequilibrio de ligamiento es el término usado para describir el hecho de que el proceso de selección de bolitas no es independiente como lo describimos nosotros antes; es decir, al seleccionar una bolita de las dos cajas del padre para copiarla al hijo, la siguiente bolita a escoger tiene mayor probabilidad de ser de la misma caja del padre que lo que correspondería al azar (50 % de cada caja). Esto se debe, en nuestro genoma, a que los genes están físicamente organizados en forma lineal en los cromosomas y, por lo tanto, los que están cerca tienen mayor probabilidad de «ir juntos» en el muestreo. Si bien este es un problema real, su impacto a nivel práctico parece ser bastante menor y en todo caso relativamente fácil de solucionar, muestreando marcadores lo suficientemente espaciados como para que el desequilibrio de ligamiento entre los mismos sea mínimo o inexistente. Un problema más real (de casi todos los métodos), es la dependencia de los resultados del conjunto de poblaciones de referencia usados.

El primer problema y más obvio es que si en el conjunto de poblaciones de referencia no se encuentran las «verdaderas» de origen del individuo, el método le asignará de cualquier manera frecuencias a las poblaciones de las que dispone, de tal manera que su suma dará 1 (100 %). Si la(s) población(es) de donde desciende no se encuentran, pero están algunas muy cercanas (en términos de **similitud genética**), entonces el problema no será muy importante; cuanto mayor la distancia entre las poblaciones de origen real y las que se encuentran representadas, mayor también la posibilidad de diferencias en las frecuencias estimadas. Además, dependiendo del origen del individuo a analizar, los resultados pueden resultar demasiado *coarse grain* para nuestros conocimientos genealógicos; por ejemplo, en Uruguay una parte

importante de la población tiene origen francés, pero esta población no se encuentra entre las de referencia para el proyecto «1000 genomes», lo que hará que la proporción de ese origen aparezca inflando otras poblaciones de Europa que si están (GBR, CEU, IBS, TSI, FIN).

Otro problema, menos obvio pero no menos importante, es el que genera la introducción de poblaciones de referencia con un alto grado de mestizaje, en particular las que tienen aportes de otras poblaciones de referencia. Este problema puede sesgar en forma importante las estimaciones de los coeficientes de ascendencia de un individuo y es difícil de distinguir el origen del problema. Por esta razón, una estrategia un poco cruda pero efectiva consiste en eliminar del conjunto de poblaciones de referencia a estas poblaciones con una proporción importante de mestizaje. Dependiendo del criterio utilizado, unas 19 poblaciones del proyecto «1000 genomes» muestran poca proporción de mestizaje y estarían aptas para ser usadas en el método de máxima verosimilitud. En el caso de las poblaciones pertenecientes al grupo **AMR**, solamente **PEL** (peruanos de Lima) presenta un nivel de mestizaje suficientemente bajo para ser incluida como población de referencia. Esto produce que, en general, toda la variabilidad de un individuo correspondiente a nativos americanos vaya a ser asignada a PEL, aunque en la mayor parte de los casos se corresponda a otros grupos étnicos bien distantes (en nuestro país la proporción esperable de ascendencia de nativos andinos es muy baja, aún entre los descendientes de nativos americanos, que se corresponden más a la macroetnia charrúa o a los guaraníes). Otro grupo con grandes problemas es el **SAS**, pero en nuestro país suele tener poca relevancia ya que pocos individuos presentan ascendencia cercana de estos grupos.

9.8 Conclusión

La semejanza entre parientes es uno de los hechos que más relacionamos con la disciplina de la genética. En este capítulo nos acercamos al fenómeno utilizando la aproximación de la genética cuantitativa, la cual desarrollamos en el capítulo anterior. Conceptos como la consanguinidad y el análisis de diagramas de parentesco nos permiten realizar un análisis sistemático del grado de similitud genética que poseen individuos emparentados. Pero ¿cómo se vincula la similitud genética con la fenotípica? ¿Cumple la herencia de los caracteres fenotípicos las mismas regularidades que se pueden plantear en el caso genético? En el siguiente capítulo entraremos en estas cuestiones, presentando el concepto de *heredabilidad*. Este es clave, en tanto el objetivo del mejoramiento genético es en última instancia realizar selección sobre caracteres fenotípicos de interés; la viabilidad de dicha selección depende de cuánto se hereda un carácter fenotípico de una generación a la siguiente, ya que la selección solo puede operar en caracteres heredables (como vimos en la primera parte de este libro).

9.9 Actividades

9.9.1 Control de lectura

1. Dos individuos emparentados comparten material genético. ¿Cómo se relaciona esto con el concepto de parentesco entre ambos? ¿Y con el concepto de consanguinidad?
2. ¿Pueden dos medios hermanos tener todos sus alelos idénticos por ascendencia? ¿E idénticos en estado?
3. Bajo el modelo genético básico, la covarianza fenotípica entre parientes se puede desglosar en diferentes componentes de varianza. ¿Cuáles son los mismos en los siguientes casos? *i)* Caso padre-hijo, *ii)* abuelo-nieto, *iii)* hermanos enteros y *iv)* medios hermanos.
4. Define el concepto de parentesco aditivo y discuta su importancia en el contexto del mejoramiento genético.
5. ¿Por qué pares de medios hermanos no pueden compartir varianza de dominancia? Desarrolla teniendo en cuenta que la varianza de dominancia se relaciona con la probabilidad de heredar un mismo genotipo.

9.9.2 ¿Verdadero o falso?

1. Los individuos pertenecen a poblaciones finitas, por lo cual siempre comparten cierta cantidad de material genético. Por lo tanto, el coeficiente de parentesco entre dos individuos de una misma población debe ser distinto de cero.
2. El parentesco de dominancia entre dos hermanos es distinto de cero.
3. La consanguinidad entre dos individuos emparentados siempre es mayor a cero.
4. Consideremos la distribución de la variabilidad dentro de grupos y entre grupos de individuos. Cuanto mayor es la covarianza entre individuos pertenecientes a un mismo grupo, mayor es la proporción de la variabilidad total que se explica por la varianza entre grupos (y viceversa).
5. Es posible estimar la varianza aditiva a través de la covarianza y otros componentes de varianzas.

Soluciones

1. **Verdadero.** Recordemos que el coeficiente de parentesco se vincula conceptualmente con la probabilidad de que dos alelos de dos individuos distintos sean idénticos por ascendencia. En una población finita, debido al efecto de la deriva, en última instancia todos los individuos comparten alelos (recordemos la teoría del coalescente y su relación con la endocría en una población, capítulo 6). Durante el cálculo del coeficiente de consanguinidad de un individuo se asume que los últimos ancestros del mismo poseen un coeficiente de parentesco de cero solo por motivos prácticos.
2. **Falso.** La afirmación es falsa ya que esto no aplica para todos los casos: en el caso de hermanos enteros sí ocurre esto, pero para medios hermanos el coeficiente de parentesco es siempre cero (ya que no es posible que los individuos compartan genotipos).

3. **Falso.** El coeficiente de consanguinidad solo se puede calcular para un individuo. El concepto no aplica a relaciones entre individuos, sean estos parientes o no.

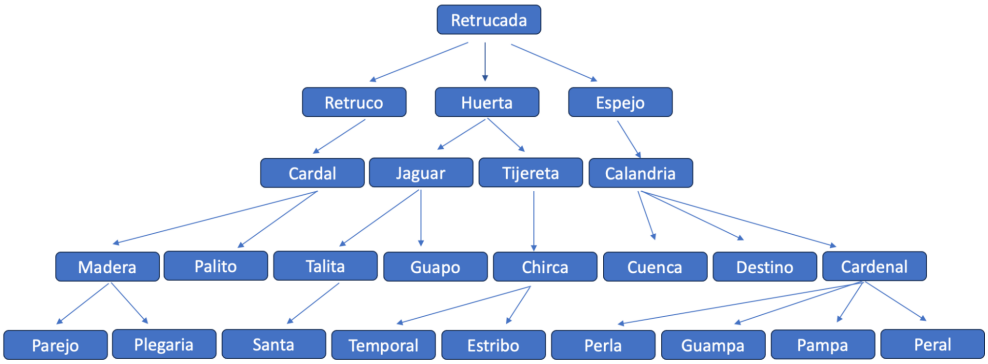
4. **Verdadero.** Esto se puede ver en el ejemplo de la figura 9.9. Dicho de otra forma, mayor covarianza entre individuos implica mayor similitud entre los mismos. Por lo tanto, una mayor proporción de la variabilidad total estará explicada por la diferencia entre los grupos (ya que dentro de cada grupo la variabilidad es baja). En el caso contrario una menor covarianza implica mayor diferencia entre los individuos de un mismo grupo, por lo que una mayor proporción de la variabilidad total estará explicada por estas diferencias dentro de los grupos (y no entre los mismos).

5. **Verdadero.** Si bien la covarianza fenotípica es la única cantidad directamente medible a partir de los datos, es posible estimar otros componentes (como la variabilidad aditiva) partiendo del modelo genético básico y realizando algunas asunciones razonables (véase sección «Estimación de las varianzas aditiva y de dominancia»).

9.9.3 Ejercicios

Ejercicio 9.1

A continuación se presenta un diagrama de parentesco de caballos criollos:



Basados en esta información:

a. ¿Cuál es el parentesco entre Tijereta y Plegaria?

b. ¿Cuánto valdría la consanguinidad de un hijo de Guampa y Cuenca?

Solución

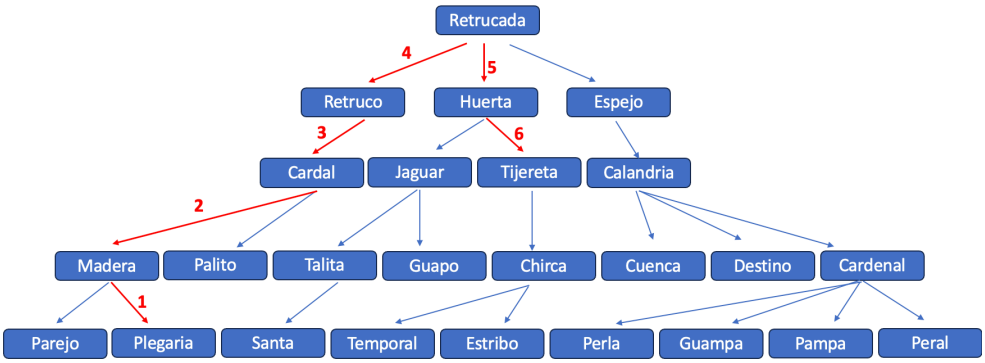
a. Podemos constatar que Tijereta y Plegaria comparten *un único* ancestro común (que es Retrucada), por lo que podremos expresar la ecuación 9.10 como

$$a_{TP} = \frac{1}{2}^{n} (1 + F_{Retrucada}) \tag{9.79}$$

En primera instancia, podremos asumir que $F_{Retrucada} = 0$, debido a que no tenemos información de que el animal posea consanguinidad (no tenemos la información y no podemos determinarla a través del pedigrí). Por lo tanto

$$a_{TP} = \frac{1}{2}^{n} \tag{9.80}$$

El único dato que necesitaremos entonces es el valor de n, el cual obtendremos contando las flechas que separan a ambos individuos.



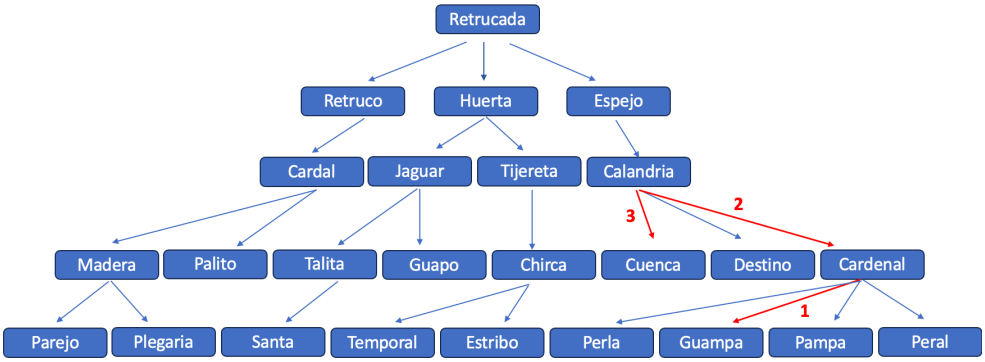
Procediendo con el método determinamos que $n = 6$, por lo tanto, el parentesco entre Tijereta y Plegaria es de

$$a_{TP} = \frac{1}{2}^{6} = \frac{1}{64} = 0{,}015625 \tag{9.81}$$

b. En primera instancia, podemos recordar que el coeficiente de consanguinidad (F) de un individuo equivale a la mitad del parentesco de sus padres. Por este motivo, para determinar la consanguinidad de un hijo entre Guampa y Cuenca deberemos determinar el parentesco entre ellos.

Podremos comprobar, al igual que en el caso anterior, que ambos animales comparten un único ancestro en común (Calandria). Repetiremos entonces lo aplicado en la parte a, tal que $a_{GC} = \frac{1}{2}^{n}(1 + F_{Calandria})$

Nuevamente asumimos que $F_{Calandria} = 0$ por las mismas consideraciones detalladas en el caso anterior, por lo que solo resta determinar n:



Comprobamos entonces que $n = 3$ por lo que

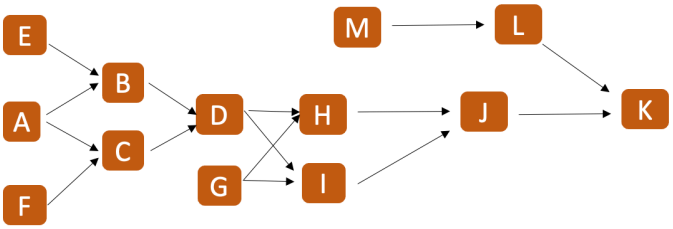
$$a_{GC} = \frac{1}{2}^3 = \frac{1}{8} = 0{,}125 \tag{9.82}$$

Entonces, la consanguinidad de un hijo de Guampa y Cuenca valdría

$$F_{Hijo} = \frac{1}{2} a_{GC} = 0{,}0625 \tag{9.83}$$

Ejercicio 9.2

A continuación se presenta un diagrama de pedigrí de una familia de ratones:



Basados en esta información:

a. ¿Cuánto vale el parentesco entre H e I?

b. ¿Cuánto vale la consanguinidad de J?

c. ¿Cuánto vale la consanguinidad de K?

Solución

a. Como podemos observar en el esquema, H e I poseen dos ancestros en común que son D y G, por lo que

$$a_{HI} = \frac{1}{2}^n (1 + F_D) + \frac{1}{2}^n (1 + F_G) \tag{9.84}$$

En este sentido, $F_G = 0$ mientras que, observando el esquema, podemos ver que los padres de D están emparentados por lo que F_D no será 0.

Calculemos entonces F_D

$$F_D = \frac{1}{2} a_{BC} \tag{9.85}$$

$$a_{BC} = \frac{1}{2}^n (1 + F_A) = \frac{1}{2}^2 (1 + 0) = \tag{9.86}$$

$$F_D = \frac{1}{2}\frac{1}{4} = \frac{1}{8} \tag{9.87}$$

Por lo que ahora contamos con esta información para poder calcular el parentesco entre H e I:

$a_{HI} = \frac{1}{2}^2(1 + \frac{1}{8}) + \frac{1}{2}^2 = \frac{1}{4}\frac{9}{8} + \frac{1}{4} = \frac{9}{32} + \frac{8}{32} = \frac{17}{32} = 0{,}53125$

b. Como hemos visto, la consanguinidad de J será equivalente a la mitad del parentesco de sus padres y que sabemos del punto anterior que vale $\frac{17}{32}$. Entonces, podemos calcular la consanguinidad de J como

$$F_J = \frac{1}{2}\frac{17}{32} = \frac{17}{64} = 0{,}265625 \tag{9.88}$$

c. Como nos ha sucedido en otras ocasiones, podemos comprobar que $F_K = 0$ ya que no existe relación de parentesco entre J y L (padres de K).

Ejercicio 9.3

En una población finita un individuo A se reproduce con un individuo B no emparentado, teniendo como progenie al individuo C. Un estudio revela que el coeficiente de consanguinidad en el individuo C es de 0,13. ¿Cómo es esto posible? Nota: calcula teniendo en cuenta que nos encontramos estudiando una población finita (donde la deriva opera generando endocría).

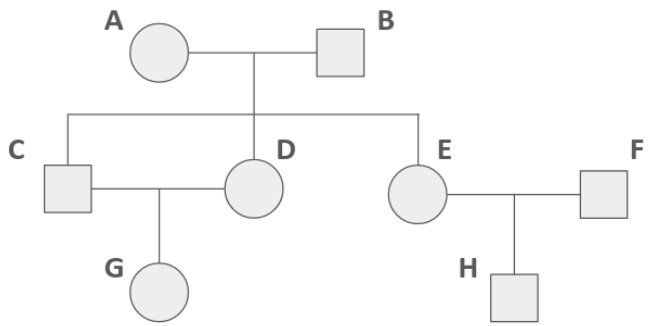
Solución

Notemos que A y B son individuos no emparentados. La aproximación usual al momento de evaluar el coeficiente de consanguinidad en C es asumir que el parentesco aditivo $a_{AB} = 0$, lo cual implica también que el coeficiente de consanguinidad $F_C = 0$, ya que

$F_C = \frac{1}{2}a_{AB}$. No obstante, esto es una simplificación de la realidad: A y B pertenecen a una población real de tamaño finito, y como vimos en la segunda parte de este libro esto implica que compartirán alelos idénticos por ascendencia simplemente por efecto de la deriva genética (capítulos «Deriva genética» y «Apareamientos no-aleatorios», recordar además el concepto del modelo del coalescente). ¿Cuál es la probabilidad de que dos alelos tomados al azar de A y B sean idénticos por ascendencia? Justamente el coeficiente de fijación de la población, que será a su vez igual al coeficiente de parentesco aditivo, si repasamos su definición (es decir, $a_{AB} = F = 0{,}13$). Por lo tanto, el coeficiente de consanguinidad de C no será de cero, si no de $F_C = \frac{1}{2}0{,}13 = 0{,}625$

Ejercicio 9.4

A continuación se muestra una genealogía que detalla las relaciones de parentesco entre 8 individuos.



a. Calcula el coeficiente de consanguinidad para el individuo G.

b. Un estudio poblacional muestra que en la población el coeficiente de fijación para la población es de aproximadamente $F = 0{,}14$. Teniendo este dato responda: ¿cuál es el coeficiente de consanguinidad para el individuo H? ¿Cuál es la probabilidad de que dos alelos sean IBD en el individuo G?

Solución

a. Recordemos que el coeficiente de consanguinidad de G equivale a $\frac{1}{2}a_{CD}$, donde a_{CD} es el parentesco aditivo de C y D. Para calcular a_{CD} se puede recurrir al método del diagrama de flechas, según el cual el parentesco aditivo entre dos individuos X e Y equivale a

$$a_{XY} = \sum_{A_i} (\frac{1}{2})^{n_{i1}+n_{i2}}(1 + F_{A_i}) \tag{9.89}$$

Notemos no obstante que los individuos C y D se encuentran emparentados, y son hermanos completos. El coeficiente de parentesco en este caso ya fue desarrollado en el capítulo (siendo $a_{CD} = \frac{1}{2}$), por lo que no nos extenderemos nuevamente el su cálculo. No obstante, remarcaremos (ya que es importante para la siguiente parte del ejercicio) que esta aproximación tiene en cuenta que dicho cálculo tiene en cuenta que A y B son individuos no emparentados (es decir, sin alelos IBD).

De lo anterior se tiene que el coeficiente de consanguinidad de G es de $(\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{4}$

b. ¿Qué sucede en el caso del individuo H? Notemos que sus progenitores, E y F, no están emparentados. Esto implica, en principio, que el coeficiente de consanguinidad de H debería ser de 0 (razonamiento análogo al utilizado para calcular los coeficientes de consanguinidad de los individuos C o D en el ejercicio anterior). No obstante, se reporta que el F en la población es de 0,14. Si recordamos de capítulos anteriores, la probabilidad de que dos alelos tomados al azar en la población sean IBD; ya que los individuos E y F son técnicamente dos individuos tomados al azar de la población, F debería ser también la probabilidad de que dos alelos de estos individuos sean IBD. Es decir, F es en efecto equivalente a lo que antes definimos como a_{EF}. Teniendo esto en cuenta tenemos que la consanguinidad de H es de $(\frac{1}{2}) \times a_{EF} = (\frac{1}{2}) \times 0{,}14 = 0{,}07$.

Notemos que el dato de que $F \neq 0$ afecta también a la estimación que hicimos en la primera parte del ejercicio. Por razonamiento análogo al empleado entre los individuos E y F, el coeficiente de parentesco $a_{AB} = 0{,}14$. Esto implica a su vez que los individuos C y D tienen coeficiente de consanguinidad de 0,07 cada uno. Esto debe contemplarse al momento de calcular el parentesco aditivo entre C y D para calcular la consanguinidad en G. La consanguinidad entre C y D ya no es de $\frac{1}{2}$, sino que

$$\begin{aligned} a_{CD} &= \frac{1}{2}^2 (1 + F_A) + \frac{1}{2}^2 (1 + F_B) \\ a_{CD} &= \frac{1}{2}^2 (1 + 0{,}07) + \frac{1}{2}^2 (1 + 0{,}07) \\ a_{CD} &= 0{,}535 \end{aligned} \tag{9.90}$$

De esto se desprende a su vez que $F_G \neq \frac{1}{4}$, sino que ahora aproximamos que $F_G = (\frac{1}{2}) \times a_{CD} = (\frac{1}{2}) \times 0{,}535 = 0{,}2576$

Ejercicio 9.5

Se conoce que la varianza aditiva del peso de carcasa en corderos pesados es $2{,}4\,\text{kg}^2$ a la vez que la varianza de dominancia es $1{,}2\,\text{kg}^2$ y la de ambiente común es $2{,}7\,\text{kg}^2$.

¿Cuál es la covarianza entre medios hermanos para esta característica? ¿Cuál es la covarianza entre hermanos enteros para esta característica?

Solución

En el caso de medios hermanos $\text{Cov}_{ME} = \frac{1}{4}V_A$, mientras que en hermanos enteros la $\text{Cov}_{HE} = \frac{1}{2}V_A + \frac{1}{4}V_D + V_{E\text{común}}$. Por lo tanto, en este caso $\text{Cov}_{ME} = \frac{1}{4} \times 2{,}4 = 0{,}6\,\text{kg}^2$ y $\text{Cov}_{HE} = \frac{1}{2} \times 2{,}4 + \frac{1}{4} \times 1{,}2 + 2{,}7 = 4{,}2\,\text{kg}^2$

Capítulo 10

Parámetros genéticos: heredabilidad y repetibilidad

È mia ferma convizione, sostenuta da anni di osservazioni e sperimentazioni, che gli uomini non sono uguali, che alcuni sono stupidi ed altri non lo sono, e che la differenza è determinata non da forze o fattore culturali ma dalle mene bio-genetiche di una imperscrutabile Madre Natura. Uno è stupido nelle stesso modo in cui altro ha i capelli rossi; uno appartiene al gruppo degli stupidi come un altro appartiene ad un gruppo sanguigno. Insomma uno nasce stupido per volere imperscrutabile e insindacabile della Divina Provvidenza.[1]

Carlo Cipolla

En el mejoramiento genético, tanto animal como vegetal, el objetivo central suele ser cambiar la frecuencia de los alelos relevantes de tal forma de modificar los valores aditivos en la dirección deseada (selección) y, por lo tanto, el fenotipo, o buscar estrategias para incrementar las combinaciones de alelos favorables (cruzamientos). En la determinación de la mejor estrategia, tanto la elección de una u otra, como de la combinación de las mismas, resulta fundamental entender la arquitectura genética de la característica a mejorar. Entendemos por **arquitectura genética** de una característica tanto los componentes que la afectan (número de genes, número de alelos en esos genes, distribución espacial de los mismos) como su interrelación y su regulación.

Desafortunadamente, en general esto no se trata más que de una expresión de deseo ya que los estudios requeridos para comprender a cabalidad dicha arquitectura suelen ser imposibles excepto para características determinadas por muy pocos genes. Sin embargo, pese a estas limitaciones, resulta fundamental comprender cómo se comporta la característica desde el punto de vista genético. ¿Tiene alguna base genética la característica? ¿Se parecen los progenitores a sus hijos más de lo que se parecen a la progenie de otros progenitores? ¿Existe variación genética importante en la población? ¿Qué rol juega el ambiente en la característica? Como veremos en capítulos siguientes, la **heredabilidad** de una característica es un parámetro central a la hora de seleccionar los individuos que serán los progenitores de la siguiente generación, así como para definir la mejor estrategia de mejoramiento.

[1]Segunda Ley Fundamental de la estupidez humana, *Allegro ma non troppo*, Carlo Cipolla (1988), un libro que en su segunda parte establece la teoría de la estupidez humana. En la fundamentación, Cipolla establece (traducción nuestra): «Tengo la firme convicción, avalada por años de observación y experimentación, de que los seres humanos no son iguales, de que unos son estúpidos y otros no, y de que la diferencia no viene determinada por fuerzas o factores culturales, sino por los dictados biogenéticos de una Madre Naturaleza inescrutable. Uno es estúpido del mismo modo que otro es pelirrojo; uno pertenece al grupo de los estúpidos del mismo modo que otro pertenece a un grupo sanguíneo. En definitiva, se nace estúpido por la inescrutable e incuestionable voluntad de la Divina Providencia». En particular, vinculado con el presente capítulo, la segunda ley viene a cuento cuando vamos a hablar de características heredables o repetibles a lo largo de una vida.

Por otra parte, ciertas características parecen repetirse en los individuos, alguna en forma espacial (es decir en distintas partes del individuo) y otras en forma temporal. Por ejemplo, los árboles frutales tenderán a reproducir anualmente su producción de frutos (medida como diferencia respecto a la media de su población), tanto en número como en peso. Los ovinos producirán todos los años vellones y si comparamos las diferencias del peso del mismo respecto a la media de los animales contemporáneos, estos desvíos de la media (hacia arriba y hacia abajo) tenderán a repetirse en cada zafra. En vacas lecheras, la producción total de cada lactancia, medida como desvío de la media del grupo contemporáneo y corregida por diferencias sistemáticas entre lactancias, también tenderá a repetirse. Esto no quiere decir que las diferencias sean exactamente las mismas, ya que el ambiente no afectará a todos los animales de la misma forma en distintos años. Tampoco todas las características tienen la misma **repetibilidad**.

Para entender mejor cómo juegan los distintos componentes de nuestro modelo genético en la arquitectura de una característica, de lo heredable de ella y de su repetibilidad (en caso de que se exprese varias veces en la vida del individuo) es que vamos a volver al modelo genético básico y utilizar distintas herramientas estadísticas, tanto para entender estos parámetros genéticos, como para predecir el valor de cría de los individuos o la producción más probable en una próxima medición en el mismo individuo.

OBJETIVOS DEL CAPÍTULO

☐ Definir el concepto de heredabilidad y discutir su relevancia en la selección de individuos.

☐ Comprender las diferencias entre la heredabilidad en sentido amplio y sentido estricto.

☐ Analizar peculiaridades de la heredabilidad en diferentes contextos.

☐ Indagar en algunos aspectos evolutivos de la heredabilidad.

☐ Analizar cómo es posible estimar la heredabilidad en la era genómica.

☐ Comprender métodos estadísticos y computacionales utilizados para estimar la heredabilidad.

☐ Definir el concepto de repetibilidad y comprender su uso como herramienta.

10.1 Heredabilidad

Intuitivamente, cuando hablamos de heredabilidad de una característica tendemos a pensar en el parecido entre progenitores y progenie. Sin embargo, el parecido entre progenitores y progenie puede tener causas genéticas como no genéticas. Por ejemplo, es muy conocido el problema (en realidad casi todos son problemas en esa área) de determinar la heredabilidad de una característica como el coeficiente intelectual (en seres humanos): si bien parece existir cierta base genética en las diferencias observadas entre individuos al realizar las pruebas, es indudable el rol que juega el ambiente donde crecieron y se educaron los individuos; padres que han alcanzado un importante nivel educativo suelen invertir más en la educación y estimulación de sus hijos que aquellos padres que cuentan con menos posibilidades. Esto hace que los padres tiendan a parecerse a sus hijos en los logros académicos, más allá de la posible existencia de una base genética en la característica.

Otras características tienen una base totalmente genética, pero sin embargo las, diferencias que podemos observar entre padres e hijos no siempre se deben a la genética. Por ejemplo, el número de patas en los mamíferos terrestres es de cuatro (tetrápodos), un número determinado por el genoma y su regulación, así como la disposición de los miembros. La aparición de animales con tres patas o con una disposición u orientación alterada de las extremidades, debido a causas ambientales durante el desarrollo embrionario (como la exposición a drogas, por ejemplo, la Talidomida), no modifica el determinismo genético de dicha característica, pese a introducir variaciones. Por ello, es necesario definir con precisión qué entendemos por heredabilidad de una característica.

La heredabilidad como relación de varianzas

Como vimos previamente, podemos descomponer las causas biológicas de los fenotipos que observamos en distintas componentes, tanto genéticas como ambientales. En general, para la mayoría de las características cuantitativas, dado el fenotipo observado en los progenitores, esperamos que en alguna medida su progenie se les parezca, aunque difícilmente el parecido sea total. Ya en el año 1886, sir Francis Galton[2] había reportado este hecho, curioso para la época (Galton 1886). El estudio de Galton involucraba las alturas de 205 parejas de padre y madre, así como las de 930 hijos de estas, todos en edad adulta (y multiplicando las alturas de las mujeres por el factor de corrección 1,08, para que fuesen comparables). A partir de graficar la altura promedio de los padres (en las abscisas) y las de los hijos (en las ordenadas), Galton observaba que los puntos caían dentro de una elipse, cuyo eje mayor sería la recta de regresión (en términos modernos) y que la pendiente de esta recta era menor a 1. Es decir, respecto a la altura, en promedio los hijos se desviaban de la media poblacional menos que el promedio de sus padres, como se puede apreciar en la figura 10.1, en la relación $1 \rightarrow \frac{2}{3}$. A este fenómeno Galton lo llamó «regresión hacia la mediocridad» y el mismo fue una de las fuentes de discrepancia durante años entre biométricos y mendelianos (si en cada generación los valores más alejados tienden hacia el promedio, en pocas generaciones todos los individuos estarán muy cerca del promedio, desapareciendo la variabilidad). Como ya vimos parcialmente antes, y como dejaremos en claro ahora, la perspectiva biométrica y la

[2] Sir Francis Galton (16 febrero de 1822 – 17 enero de 1911) fue un científico y *polymath* inglés de la era victoriana. Entre otras área del conocimiento abundó en estadística, sociología, sicología, antropología, metereología y geografía. Además, propuso las técnicas de análisis de correlación y regresión que serían fundamentales para el desarrollo temprano de la genética.

mendeliana son claramente compatibles y la explicación de este fenómeno es relativamente sencilla: no todas las diferencias que observamos en los padres tienen base genética y, por lo tanto, no es esperable que dichas diferencias aparezcan totalmente incambiadas en los hijos.

Nuestro modelo genético establecía que el fenotipo era la suma de los efectos genéticos y ambientales, así como sus interacciones. De forma simplificada, nuestro modelo era $P = A + D + I + E$. Si un individuo posee un valor fenotípico elevado respecto al promedio de la población, en altura, por ejemplo, desde el punto de vista matemático cualquier combinación de estos efectos podría producir este efecto. Sin embargo, en general, los individuos más altos de la población serán aquellos en los que tanto los componentes genéticos como los ambientales se alineen para arriba, es decir que tanto los genéticos son positivos, como los ambientales son positivos respecto de la media. Como ya vimos antes, los padres no transmiten a sus hijos sus propias combinaciones de alelos, que son las bases de la dominancia y de la epistasis. Tampoco transmiten en ambiente, ya que es algo externo a ellos. Por lo tanto, el único componente transmisible de padres a hijos sería el componente de los efectos genéticos aditivos.

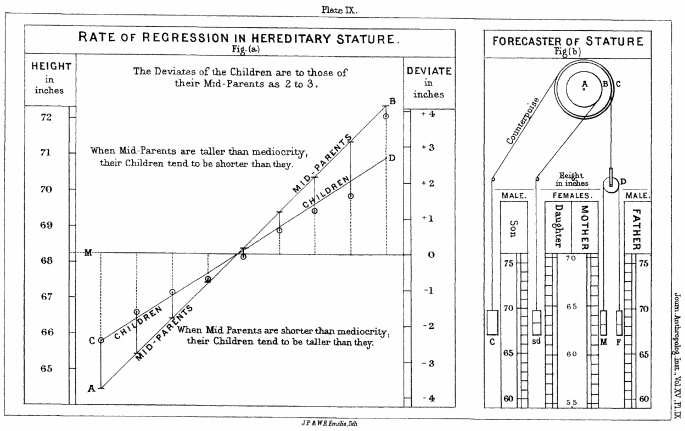


Figura 10.1. Captura de imagen de la figura original (Plate IX) en el artículo de Galton (1886). A la izquierda (Fig. (a)) se aprecia la forma en que dedujo la relación $1 \rightarrow \frac{2}{3}$ para las desviaciones de altura de padres e hijos, ya que los segmentos de recta correspondientes a los hijos tienen aproximadamente $\frac{2}{3}$ del largo correspondiente al de sus progenitores. Notemos que la forma del gráfico de regresión difiere respecto a cómo lo graficaríamos en la actualidad, ya que tanto abscisas como ordenadas se refieren a la medida en los padres (por esa razón el ajuste de la recta de padres es perfecto). La escala de la derecha es en términos absolutos, mientras que la escala a la izquierda del gráfico marca los puntos como desvío de la media. A la derecha (Fig. (b)) aparece el diagrama de una máquina para pronosticar la altura de los hijos a partir de la altura de los padres, donde se aprecia la diferencia en las escalas de los dos sexos. (Trabajo de dominio público en casi todos los países al haber pasado más setenta años de la muerte del autor).

El resto de los efectos esperamos que se comporten de forma aleatoria con media cero y, por lo tanto, no aportarán más altura a los hijos que el promedio de los efectos aditivos de sus padres, que era menor a su vez que el promedio fenotípico. Esto explica perfectamente lo observado por Galton de que los hijos de padres altos suelen estar también sobre el promedio, pero menos que el promedio de sus padres. Idéntico razonamiento se aplica a los

padres bajos, pero en este caso sus hijos serán menos bajos que el promedio de sus padres, acercándose también al promedio general.

Ahora que nos resulta más clara la base conceptual de lo que esperar de la transmisión de una característica cuantitativa, veamos lo que ocurre con la variación. Vimos antes, en el capítulo «El modelo genético básico, la forma en que podíamos pasar de trabajar con los datos individuales a la descripción del modelo en forma de varianzas, es decir

$$V_P = V_G + V_E = V_A + V_D + V_I + V_E \tag{10.1}$$

Ahora, si queremos entender el papel que juega la única fuente de variación heredable, es decir la de los efectos genéticos aditivos, resulta natural expresar esta como la proporción de la variabilidad total correspondiente a los efectos aditivos. Por lo tanto, llamamos **heredabilidad** a la relación

$$h^2 = \frac{V_A}{V_P} = \frac{V_A}{V_A + V_D + V_I + V_E} \tag{10.2}$$

Notemos que el símbolo h^2 representa la heredabilidad, no el cuadrado de la misma. Esto se debe a que originalmente se definió el símbolo h como el coeficiente en el método de «path analysis» del artículo S. Wright (1921) y que, por lo tanto, su cuadrado (h^2) representa la proporción heredable de la variabilidad. Otra cosa fundamental a tener en cuenta, que a esta altura debería resultar obvia, cuando hablamos de **la heredabilidad de una característica** se trata de un **parámetro poblacional**. Esto se refiere, por un lado, a que los individuos ***no tienen*** una heredabilidad específica a cada uno de ellos (ya que se trata de una relación de varianzas, ambas determinadas en la misma población). Pero por otro lado, también se refiere a que la **heredabilidad de una característica** corresponde a la determinación de la misma en una población dada y en un momento dado.

Para entender esto último alcanza con analizar la definición en la ecuación (10.2). Por un lado, la varianza aditiva puede variar de población en población; bajo fuerte selección es esperable que la misma disminuya en la población a lo largo del tiempo ya que si de forma continua elegimos animales «mejores», irán disminuyendo su frecuencia en la población alelos y genotipos (fuentes de varianza aditiva). Por otro lado, mucho más importante aún es la variabilidad que podemos observar en el denominador. De acuerdo con nuestro modelo genético $P = G + E$ y despreciando las interacciones $V_P = V_G + V_E$. Más aún, habíamos acordado que en este modelo todas las variaciones no explicadas por lo genético las considerábamos como variaciones ambientales, entre ellas los errores de medición. Por ejemplo, si estamos considerando la heredabilidad de la característica producción de leche a 305 días en un determinado rodeo de gran tamaño y las mediciones de producción las realizamos con un equipamiento con gran error aleatorio, entonces la varianza «ambiental» será mayor y todo el denominador será mayor que si lo realizamos con un equipo más preciso. Como la varianza aditiva permanece incambiada, en el primer caso **nuestra estimación** de la heredabilidad de la característica será menor que en el segundo, exclusivamente debido a la forma de medir la característica.

Una representación gráfica de la heredabilidad como cociente de varianzas se observa en la figura 10.2.

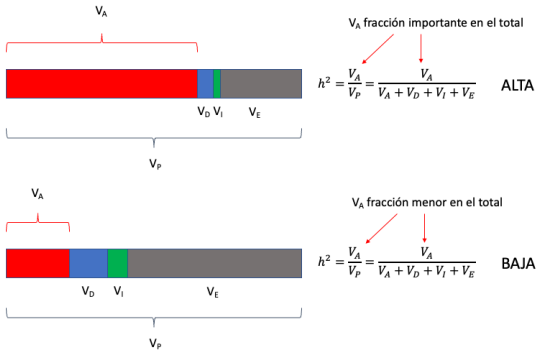


Figura 10.2. Representación gráfica de la heredabilidad como relación de varianzas. En la situación de arriba la varianza aditiva (en rojo) representa una fracción importante de la varianza fenotípica, es decir de la varianza total ($V_P = V_A + V_D + V_I + V_E$), por lo que la heredabilidad de la característica será alta. En la situación de abajo, la varianza aditiva (en rojo) es bastante pequeña en relación con total y, por lo tanto, la heredabilidad de la característica será baja.

En la misma, los distintos componentes de la varianza se representan como barras de colores, rojo para la varianza aditiva, azul para la de dominancia, verde para la de epistasis y gris para la varianza ambiental (incluyendo en esta todas las fuentes **no-genéticas**). De acuerdo con nuestro modelo, la varianza fenotípica es igual a la suma de todas estas fuentes de variación, es decir $\mathrm{V_P} = \mathrm{V_A} + \mathrm{V_D} + \mathrm{V_I} + \mathrm{V_E}$. En la situación ilustrada en la parte superior, la varianza aditiva es el componente más importante de toda la varianza y, por lo tanto, la relación $\frac{V_A}{V_A+V_D+V_I+V_E}$ será alta, o lo que es lo mismo, la heredabilidad de la característica será alta. En la situación de abajo, por otra parte, la varianza aditiva es relativamente pequeña en el total y, por lo tanto, $\frac{V_A}{V_A+V_D+V_I+V_E}$ será pequeña, lo que es decir que la heredabilidad será pequeña.

Claramente, como la heredabilidad la definimos acá como un cociente entre dos varianzas, la aditiva y la total (fenotípica), la heredabilidad estará entre 0 y 1, es decir $0 \leqslant h^2 \leqslant 1$. El límite inferior de h^2 se alcanza solamente en la situación en que la característica de interés no posee varianza aditiva ($V_A = 0$), mientras que el límite superior se alcanza en el caso hipotético donde la única fuente de varianza es la aditiva ($V_P = V_A$). En general, para la mayor parte de las características de interés agronómico existe algo de varianza aditiva y, por lo tanto, la heredabilidad es mayor a cero, mientras que el límite superior es casi imposible de alcanzar aun cuando toda la varianza teórica fuese aditiva ya que las mediciones conllevan error siempre. Notemos que como la heredabilidad es una relación de varianzas de la misma característica, siempre tendremos las mismas unidades en el numerador que en el denominador y, por lo tanto, estas desaparecerán de la heredabilidad, por lo que se trata de un **parámetro genético adimensional** (no tiene dimensiones).

Ejemplo 10.1

En un estudio previo ejemplo 9.4 para determinar los componentes genéticos de la varianza en diámetro de fibras en ovejas verdes llegamos a que nuestro mejor estimador de la varianza aditiva era $\hat{\sigma}^2_A = 1{,}80\ \mu\text{m}^2$, mientras que para la varianza de dominancia era $\hat{\sigma}^2_D = 1{,}40\ \mu\text{m}^2$. Asumiendo que para la característica la varianza de epistasis es prácticamente cero y que la varianza ambiental es de $\hat{\sigma}^2_E = 0{,}80\ \mu\text{m}^2$, obtener una estimación puntual de la heredabilidad.

De acuerdo con nuestra definición en la ecuación (10.2), la heredabilidad es igual a

$$h^2 = \frac{V_A}{V_P} = \frac{V_A}{V_A + V_D + V_I + V_E} \tag{10.3}$$

por lo que sustituyendo las varianzas por nuestros estimadores de las mismas tenemos

$$h^2 = \frac{\hat{\sigma}^2_A}{\hat{\sigma}^2_A + \hat{\sigma}^2_D + \hat{\sigma}^2_I + \hat{\sigma}^2_E} = \frac{1{,}80}{1{,}80 + 1{,}40 + 0 + 0{,}80} = 0{,}45 \tag{10.4}$$

Por lo tanto, nuestro estimador puntual de la heredabilidad de la característica diámetro de fibras en esta majada de ovejas verdes es $h^2 = 0{,}45$, un valor intermedio.

PARA RECORDAR

- El único componente transmisible del fenotipo de padres a hijos es el **aditivo**, el resto de los efectos esperamos que se comporten de forma aleatoria con media cero.
- La heredabilidad, expresada como relación entre varianzas puede definirse como $h^2 = \frac{V_A}{V_P} = \frac{V_A}{V_A+V_D+V_I+V_E}$
- La **heredabilidad** de una característica es un **parámetro poblacional**, por lo que los individuos **no tienen** una heredabilidad específica para cada uno.
- La **heredabilidad de una característica** corresponde a la determinación de la misma en una población dada y en un momento dado.
- Al estar definida como un cociente entre varianzas (la aditiva y la fenotípica), la heredabilidad estará entre 0 y 1, es decir $0 \leqslant h^2 \leqslant 1$. A su vez, al no tener unidades, se trata de un parámetro genético adimensional.

La heredabilidad como regresión

En la sección anterior vimos nuestra primera definición matemática de la heredabilidad de una característica y esta era la proporción de la varianza total atribuible a los efectos aditivos, es decir $h^2 = V_A/V_P$. Sin embargo, ya en el trabajo de Galton (1886) se encuentra implícito el concepto de predicción a partir de los datos fenotípicos (ver figura 10.1, «Fig. (b)» a la derecha). En el caso planteado por Galton (1886), la idea era predecir el fenotipo de los hijos a partir del fenotipo de los padres, pero nosotros podemos plantearnos algo un poco

más diferente y posiblemente más interesante: **predecir el valor de cría de los individuos a partir de su propio fenotipo**. Planteado de otra manera, sabemos que en general no podemos determinar directamente el valor de cría de un individuo ya que no es algo que observemos directamente. Sin embargo, el fenotipo aporta información clave para predecir el valor de cría. Entonces, ¿cómo podemos usar esta información?

Regresión del valor de cría en el fenotipo individual

Seguramente recordarás de los cursos previos de estadística que en el caso de que exista una asociación lineal entre dos variables podemos usar esta relación para inferir los parámetros de una ecuación de predicción. En particular, cuando centramos las dos variables (es decir, les restamos sus correspondientes medias), la recta de regresión debe pasar por la intersección de los ejes, el punto (0,0) y, por lo tanto, el valor del intercepto (el valor de las ordenadas cuando la abscisa es cero) es 0 también. Esto nos deja con un solo parámetro para estimar que es la pendiente de la recta de regresión. En nuestro caso estamos interesados en predecir el valor de cría de los individuos a partir de sus fenotipos, por lo que la pendiente de la recta de regresión (b_{AP}) me permitirá la mejor predicción posible del valor de cría a partir de un valor fenotípico dado.

La situación se ilustra en la figura 10.3. Supongamos que de alguna manera obtuvimos el verdadero valor de cría de un conjunto de animales, así como sus valores fenotípicos, para una característica determinada. Cada punto azul representa un individuo. En el eje de las abscisas tenemos los valores fenotípicos y en el eje de las ordenadas los correspondientes valores de cría (es decir, genéticos aditivos). A primera vista parece claro que existe una relación entre las dos variables, posiblemente lineal, ya que la densidad de puntos en los cuadrantes inferior derecho y superior izquierdo es mucho más baja que en los otros dos cuadrantes.

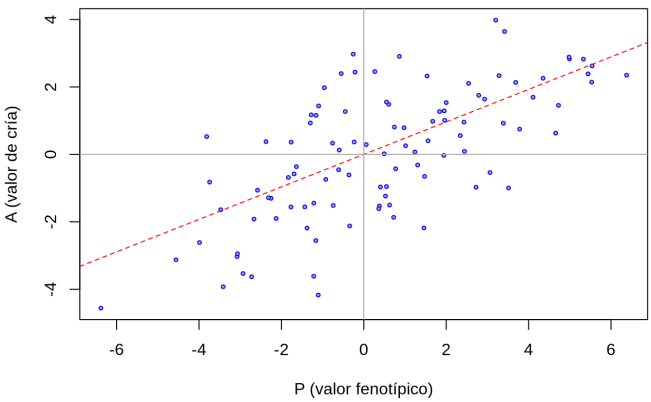


Figura 10.3. Regresión de los valores de cría en los valores fenotípicos (centrados). La pendiente de la regresión por el cero es $b_{AP} = \frac{V_A}{V_P} = h^2$

Veamos ahora qué ocurre si «ajustamos» una recta de regresión por mínimos cuadrados. De acuerdo con lo que has visto en cursos previos, el estimador por mínimos cuadrados de la pendiente entre X e Y es igual a $b_{XY} = \frac{Cov(X,Y)}{Var(X)}$, por lo que en nuestro caso tenemos

$$b_{AP} = \frac{Cov(A,P)}{V_P} = \frac{Cov(A,A+D+I+E)}{V_P} \therefore$$
$$b_{AP} = \frac{Cov(A,A) + Cov(A,D) + Cov(A,I) + Cov(A,E)}{V_P} \quad (10.5)$$

Sin embargo, como ya vimos antes, por construcción la covarianza entre el valor aditivo y de dominancia es cero, al igual que el que corresponde a aditivo y de epistasis (porque ambos se construyen como desvíos de lo explicado por el efecto anterior). Además, también hemos supuesto que en general lo ambiental no se correlaciona con lo genético, por lo que también la covarianza entre lo aditivo y lo ambiental será cero. Poniendo todo junto, como excepto la covarianza entre aditivo y aditivo el resto de los términos del numerador son cero y dado que la covarianza de una variable consigo misma es igual a la varianza, tenemos ahora que

$$b_{AP} = \frac{Cov(A,A)}{V_P} = \frac{V_A}{V_P} = h^2 \quad (10.6)$$

Es decir, **el valor de la pendiente de la recta de regresión entre el valor aditivo y el valor fenotípico es igual a la heredabilidad**. En la figura 10.4 podemos apreciar cómo funciona nuestra «máquina de predicción» (en lugar de los pesos y poleas de la máquina de Galton): para un nuevo individuo, entramos su valor fenotípico (como desvío de la media) en las abscisas y subimos (o bajamos) verticalmente hasta encontrar la recta de regresión, momento en que nos movemos horizontalmente hasta el corte con el eje de las ordenadas (el eje vertical); el punto de corte en las ordenadas será el valor de cría predicho para nuestro individuo.

En principio, hemos encontrado una herramienta fantástica. Luego de haber «calibrado» nuestra máquina con algunos pares de datos valor de cría y valor fenotípico, alcanza con tener el valor fenotípico de los individuos a los que les queremos predecir el valor de cría y nuestra «máquina» lo hará de inmediato por nosotros. En particular, el **valor de cría estimado** para el individuo i será

$$\hat{A}_i = (P_i - \bar{P})\, b_{AP} = (P_i - \bar{P})\, h^2 \quad (10.7)$$

Sin embargo, en general, nuestra máquina de predicción no será perfecta, como ocurre con la mayoría de las máquinas de predicción. El problema es fácil de apreciar en la figura 10.4. Mientras que para cualquier valor fenotípico existe una distribución de valores de cría que le corresponden (puntos grises con la misma abscisa), la máquina siempre me devolverá un solo valor: el de la ordenada del punto de corte de la abscisa con la recta de predicción. Dicho de otra manera, las distancias entre los puntos «grises» y la recta representan los errores de

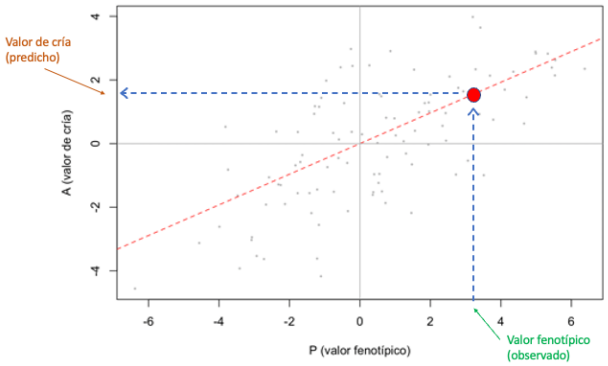


Figura 10.4. Predicción del valor de cría de un individuo a partir de su propio fenotipo. A partir de los pares de datos valor de cría y fenotipo (puntos grises) estimamos una recta de regresión cuya pendiente es la heredabilidad ($b_{AP} = V_A/V_P = h^2$). A partir de esta recta podemos predecir el valor de cría de nuevos individuos teniendo el valor fenotípico de los mismos. En la figura, entramos el valor fenotípico en el eje de las abscisas y luego «subimos» por la línea azul a trazos hasta el corte con la recta (roja a trazos) y en ese punto vamos hacia el eje de las ordenadas por la línea azul horizontal: el valor de corte en las ordenadas será el valor de cría de nuestro individuo.

predicción que cometemos usando la recta como **mejor predictor**. La pregunta que sigue es ¿cómo se relacionan los errores que cometemos con la heredabilidad?

Claramente, para cada punto (individuo) el error $\mathrm{e_i} = \mathrm{A_i} - \hat{\mathrm{A}}_\mathrm{i} = \mathrm{A_i} - (\mathrm{P_i} - \bar{\mathrm{P}})\ h^2$ será posiblemente diferente, por lo que estamos hablando de una distribución de errores. En el caso de dos variables aleatorias ligadas por una relación lineal y con distribución normal ambas (como es generalmente nuestro caso), una forma usual de medir el grado de asociación entre ellas, o la precisión de la predicción (lo que es equivalente), es el **coeficiente de correlación de Pearson** (por Karl Pearson).[3] Seguramente recordarás que para dos variables X e Y, dicho coeficiente se definía como $r_{XY} = \frac{Cov(X,Y)}{\sqrt{V_X V_Y}}$, por lo que en nuestro caso, el coeficiente de correlación entre el valor de cría y el valor fenotípico será

$$r_{AP} = \frac{Cov(A,P)}{\sqrt{V_A}\sqrt{V_P}} = \frac{V_A}{\sqrt{V_A}\sqrt{V_P}} = \frac{\sqrt{V_A}\sqrt{V_A}}{\sqrt{V_A}\sqrt{V_P}} = \sqrt{\frac{V_A}{V_P}} = \sqrt{h^2} = h \qquad (10.8)$$

Es decir, en este caso particular de predecir el valor de cría a partir del valor fenotípico del individuo, **la precisión es la raíz cuadrada de la heredabilidad** ($\sqrt{h^2} = h$). Cuanto más alta sea la heredabilidad, mayor será, por lo tanto, la precisión de la predicción. De hecho, un estadístico usado normalmente para comparar el ajuste de modelos es el coeficiente de determinación R^2, que en modelos de como el nuestro es el cuadrado del coeficiente de correlación de Pearson y por lo tanto (con datos infinitos) $R^2 = h^2$. La relación entre la heredabilidad y la precisión de la predicción del valor de cría a partir del fenotipo del individuo se observa claramente en la figura 10.5.

[3] Karl Pearson (27 marzo de 1857 – 27 abril de 1936) fue un matemático y estadístico inglés, fundador de la estadística matemática, propulsor del darwinismo social, la eugenesis y aun del «racismo científico». Curador y biógrafo oficial de Sir Francis Galton, sus aportes a la bioestadística fueron fundamentales para el desarrollo de las ciencias biológicas en general y de la genética en particular.

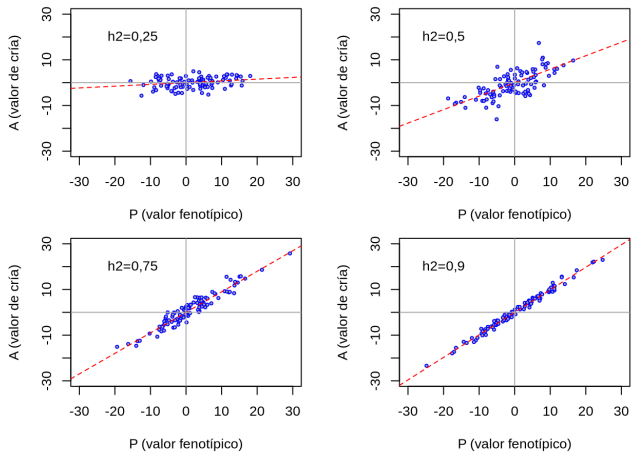


Figura 10.5. Relación entre la heredabilidad y la precisión de la predicción. La pendiente de la regresión es $b_{AP} = \frac{V_A}{V_P} = h^2$, mientras que la precisión es $r_{AP} = \frac{V_A}{\sqrt{V_P V_A}} = h$. A medida que aumenta la pendiente (es decir la heredabilidad), los puntos se distribuyen más cerca de la línea de regresión, o lo que es lo mismo, aumenta la precisión.

A medida que aumenta la heredabilidad de la característica, que es igual a la pendiente, los puntos se van «acercando» a la recta, lo que es equivalente a decir que disminuye el error ($e_i = A_i - \bar{A}_i$) y, por lo tanto, aumenta la precisión.

En general, la relación entre el coeficiente de regresión y el de correlación está dada por

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \Leftrightarrow r_{XY}\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_Y}{\sigma_X}\frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_Y}{\sigma_Y}\frac{\sigma_{XY}}{\sigma_X \sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2} = b_{XY} \therefore$$
$$b_{XY} = r_{XY}\frac{\sigma_Y}{\sigma_X} \tag{10.9}$$

por lo que en nuestro caso, como verificación de nuestros resultados previos, sustituyendo en la ecuación (10.10) A por Y y P por X, tenemos

$$h^2 = b_{AP} = r_{AP}\frac{\sigma_A}{\sigma_P} = h \times h = h^2 \tag{10.10}$$

Ejemplo 10.2

Dada la estimación de la heredabilidad obtenida en el ejemplo 10.1 para el diámetro de fibras (DF) en las ovejas verdes ($h^2_{DF} = ,45$) y sabiendo que la media de diámetro en la población de referencia es de 17,2μm, estimar el valor de cría para dicha característica de los carneros con los siguientes fenotipos:

Carnero	$\hat{A}$ Diámetro (μm)	$\hat{A}$ Peso del vellón sucio (kg)
Elmas Astado	16,8	7,635
Peloduro	19,3	9,457
Ram Bullero	17,2	6,356
Capicubo	16,1	7,574
Keto Maste	14,7	11,574

Conociendo que la heredabilidad de Peso del Vellón Sucio (PVS) en esta majada es de $h^2_{PVS} = 0{,}36$ y la media fenotípica es de 7,345 kg, estimar los valores de cría de los carneros. Calcular las precisiones de las estimaciones de valor de cría para ambas características.

De acuerdo con la ecuación (10.7)

$$\hat{A}_i = (P_i - \bar{P})\, h^2 \tag{10.11}$$

por lo que podemos transformar los valores en el cuadro anterior a

Carnero	$\hat{A}$ Diámetro (μm)	$\hat{A}$ Peso del vellón sucio (kg)
Elmas Astado	$(16{,}8 - 17{,}2)0{,}45 = -0{,}180$	$(7{,}635 - 7{,}345)0{,}36 = 0{,}10440$
Peloduro	$(19{,}3 - 17{,}2)0{,}45 = 0{,}945$	$(9{,}457 - 7{,}345)0{,}36 = 0{,}76032$
Ram Bullero	$(17{,}2 - 17{,}2)0{,}45 = 0{,}000$	$(6{,}356 - 7{,}345)0{,}36 = -0{,}35604$
Capicubo	$(16{,}1 - 17{,}2)0{,}45 = -0{,}495$	$(7{,}574 - 7{,}345)0{,}36 = 0{,}08244$
Keto Maste	$(14{,}7 - 17{,}2)0{,}45 = -1{,}125$	$(11{,}574 - 7{,}345)0{,}36 = 1{,}52244$

Como vimos previamente, de acuerdo con la ecuación (10.8), la precisión de las estimaciones a partir del fenotipo individual viene dada por

$$r_{AP} = h = \sqrt{h^2} \tag{10.12}$$

por lo que $r_{DF} = \sqrt{h^2_{DF}} = \sqrt{0{,}45} \approx 0{,}671$, mientras que para PVL será $r_{PVL} = \sqrt{h^2_{PVL}} = \sqrt{0{,}36} = 0{,}60$

Regresión del valor fenotípico de los hijos sobre el fenotipo de un progenitor

Antes veíamos la importancia de considerar la **heredabilidad como la pendiente de regresión entre el valor de cría de los individuos y su fenotipo** ya que de esta forma podíamos predecir el valor de cría de nuevos individuos. Sin embargo, el verdadero valor de cría de los individuos es usualmente desconocido, de ahí la trascendencia de poder predecirlo, por lo que para estimar esta pendiente es necesario hacerlo de otra forma. Es decir, para predecir los valores de cría precisamos de una pendiente cuya estimación precisa de valores de cría (además de los fenotípicos), por lo que de alguna manera no hay salida por ese lado.

Sin embargo, la **regresión de los valores fenotípicos entre progenie y progenitores** tiene también una relación directa con la heredabilidad, como veremos aquí. Para empezar por lo más sencillo vamos a estudiar la regresión entre los valores fenotípicos de una progenie en uno de sus progenitores. Esto es, vamos a formar parejas de datos de un progenitor (padre o madre) y uno de los descendientes. Para simplificar más las cosas, vamos a asumir que la varianza fenotípica de la característica es igual en ambos sexos, así como sus medias, aunque luego vamos a relajar alguno de estos supuestos.

Llamemos x_1 y x_2 a los progenitores y y a su progenie. Si V_P es la varianza fenotípica en los padres, la regresión del **fenotipo de un hijo** en el **fenotipo de un progenitor**, por ejemplo x_1, la podemos escribir como

$$b_{HP} = \frac{Cov_{HP}}{V_P} = \frac{Cov(y,x_1)}{V_P} \tag{10.13}$$

El denominador de la ecuación es fácilmente estimable en forma directa a partir de las observaciones fenotípicas de los padres, por ejemplo, con el estimador usual de la varianza ($\hat{\sigma}^2_P = \frac{1}{n-1}\sum_n (x_i - \bar{x})^2$) y es interpretable directamente en términos de nuestro modelo genético básico. Sin embargo, el numerador (que también podemos estimar directamente de los datos como $\hat{\sigma}^2_P = \frac{1}{n-1}\sum_n (x_i - \bar{x})(y_i - \bar{y})$) no nos dice directamente su significado en el modelo genético básico. Para obtener dicho significado debemos expresar los fenotipos de progenie y progenitor en términos del modelo genético básico, es decir

$$Cov_{HP} = Cov(y,x_1) = Cov(A_y + D_y + I_y + E_y, A_{x_1} + D_{x_1} + I_{x_1} + E_{x_1}) \tag{10.14}$$

Pero y es hijo de x_1 y x_2, por lo que su valor de cría esperado será la suma de la mitad de cría del valor de cada uno de los progenitores. Por lo tanto, podemos escribir el resultado de la ecuación (10.14) como

$$Cov(\left(\frac{A_{x_1} + A_{x_2}}{2}\right) + D_y + I_y + E_y, A_{x_1} + D_{x_1} + I_{x_1} + E_{x_1}) \tag{10.15}$$

Pero, como vimos previamente, los efectos de dominancia no covarían entre padres e hijos (al menos bajo apareamientos aleatorios), mientras que no es usual que los efectos ambientales en las dos generaciones covaríen, por lo que también los asumiremos iguales a cero.

La relación progenitor-progenie incluye efectos epistáticos ya que hay combinaciones de los alelos del padre que pueden «viajar juntas» ($\frac{1}{4}\mathrm{V_{AA}}$,$\frac{1}{8}\mathrm{V_{AAA}}$, etcétera), pero como hacemos usualmente, en general podemos despreciar estas interacciones. Esto nos deja con la siguiente relación

$$Cov(\left(\frac{A_{x_1}+A_{x_2}}{2}\right),A_{x_1}) = \frac{1}{2}[Cov(A_{x_1},A_{x_1}) + Cov(A_{x_2},A_{x_1})] = \frac{1}{2}Cov(A_{x_1},A_{x_1}) \tag{10.16}$$

La última igualdad en (10.16) es debida a que bajo apareamientos aleatorios la covarianza entre valores aditivos de los progenitores es cero, es decir, mejores padres no se aparean preferentemente con mejores o peores madres. A su vez, $\mathrm{Cov(A_{x_1},A_{x_1}) = V_A}$, ya que x_1 es un individuo tomado al azar de una población con varianza genética $\mathrm{V_A}$, por lo tanto,

$$Cov_{HP} = \frac{1}{2}Cov(A_{x_1},A_{x_1}) = \frac{1}{2}V_A \tag{10.17}$$

Tenemos ahora una expresión interpretable en términos del modelo genético básico de la covarianza entre un progenitor y su progenie, que es además el numerador del estimador del coeficiente de regresión del valor fenotípico de la progenie en el de uno de sus progenitores. Volviendo a la ecuación (10.13), si colocamos este resultado de la covarianza fenotípica entre padres e hijos tenemos que

$$b_{HP} = \frac{Cov_{HP}}{V_P} = \frac{\frac{1}{2}V_A}{V_P} = \frac{1}{2}\frac{V_A}{V_P} = \frac{1}{2}h^2 \tag{10.18}$$

Por lo tanto, **la pendiente de la regresión de los valores fenotípicos de una progenie en uno de sus progenitores igual a un medio de la heredabilidad**, por lo que al multiplicar dicho coeficiente de regresión por 2 tenemos un estimador insesgado de la heredabilidad, es decir

$$b_{HP} = \frac{1}{2}h^2 \Leftrightarrow h^2 = 2\, b_{HP} \tag{10.19}$$

La importancia de la ecuación (10.19) es obvia. A partir de valores fenotípicos de padres e hijos podemos estimar la heredabilidad de una característica, que como vimos previamente es a su vez la pendiente de la regresión de los valores de cría individuales en sus correspondientes valores fenotípicos. Esto nos permite salir de la encrucijada desde la que partimos y ahora sí (teniendo un estimado de la heredabilidad) podemos utilizar todo el poder predictor de los valores de cría que tienen los fenotipos.

Veamos ahora algunos aspectos prácticos de esta regresión fenotípica. Por un lado, hemos trabajado con un solo hijo por progenitor, pero es fácil de demostrar que hubiese sido lo

mismo trabajar con el promedio de varios hijos (en principio, un número determinado e igual para todos los padres). Es decir, mientras que las parejas se formen al azar, la covarianza entre valor de cría de un progenitor y el promedio de N de sus progenies se podrá descomponer como

$$Cov_{\bar{H}P} = \frac{1}{N}\left[\frac{1}{2}Cov(A_{x_1},A_{x_1}) + \frac{1}{2}Cov(A_{x_1},A_{x_1}) + \cdots + \frac{1}{2}Cov(A_{x_1},A_{x_1}) + \frac{1}{N}\left(\frac{1}{2}Cov(A_{x_?},A_{x_1}) + \frac{1}{2}Cov(A_{x_?},A_{x_1}) + \cdots + \frac{1}{2}Cov(A_{x_?},A_{x_1})\right)\right] \quad (10.20)$$

con $A_{x_?}$ representando el valor de cría de todos los individuos que no son x_1. Pero como ya vimos, al ser formadas al azar las parejas, la correlación entre valores de cría de x_1 con el resto de los $x_?$ será cero y, por lo tanto, todo el segundo término de la derecha en (10.21) es cero. Esto nos deja con

$$\begin{aligned} Cov_{\bar{H}P} &= \frac{1}{N}\left[\frac{1}{2}Cov(A_{x_1},A_{x_1}) + \frac{1}{2}Cov(A_{x_1},A_{x_1}) + \cdots + \frac{1}{2}Cov(A_{x_1},A_{x_1})\right] \\ &= \frac{1}{2}\frac{1}{N}\left[N\,Cov(A_{x_1},A_{x_1})\right] \\ &\therefore Cov_{\bar{H}P} = \frac{1}{2}V_A \end{aligned} \quad (10.21)$$

Por otro lado, asumimos que la varianza fenotípica de la característica era la misma en los dos sexos, algo que no siempre ocurre en las características de interés económico. Esto nos deja con dos posibilidades al menos: a) estimar la heredabilidad de la característica para cada sexo, y b) corregir los registros de alguna manera para compensar por estas diferencias. Más aún, algunas características solo se expresan en solamente uno de los dos sexos, por ejemplo, la producción de leche. En ese caso resulta obvia la regresión dentro de sexo, por ejemplo, vacas lecheras hijas de otras vacas lecheras. En el caso de la regresión de hijos e hijas en madres también debe considerarse la existencia de un componente de varianza materna (V_M), cuyo aporte será $\frac{1}{2}V_M$ y que puede no ser despreciable, especialmente en aquellas características como el crecimiento en bovinos donde el aporte de la madre es fundamental. En estos casos la no consideración de este componente sesgará hacia arriba el estimado de la heredabilidad ya que el numerador incluye elementos que no corresponden a la varianza aditiva.

Ejemplo 10.3

Diferentes estudios concluyen que la mejor raza de perros para el cuidado de las **ovejas verdes** es el **perro de montaña de los Pirineos** (figura 10.6), un perro de gran tamaño, rústico y suficientemente tranquilo como para no alterar el comportamiento de las ovejas. Sin embargo, dado el gran tamaño de las **ovejas verdes**, es necesario seleccionar perros de gran altura, por lo que fue encargado un estudio sobre la heredabilidad de esta característica entre los criaderos de la zona. Se seleccionaron 10 perros (machos) que habían sido criados en las mismas condiciones de alimentación y cuidados y para cada uno de ellos se eligió al azar un hijo (macho). Todos los animales se midieron a edad adulta (3 años).

Figura 10.6. Perro de montaña de los Pirineos, también conocido como Gran Pirineo, Chien des Pyrénées, Chien de Montagne des Pyrénées, Gigante de los Pirineos, Gran Pirineu, Gos de Muntanya dels Pirineus, Patou, entre otros. Un perro pastor de gran tamaño, con una altura a la cruz de entre 65 y 80 cm (foto de Wikipedia, HeartSpoon, CC BY-SA 3.0 via Wikimedia Commons).

En el cuadro siguiente se presentan las alturas de padres e hijos. A partir de estos datos estimar la heredabilidad de la característica altura a la cruz.

Padre	Altura cm.	Altura hijo cm.
Euler	76	70
Leonardo	72	77
Galileo	67	73
Poisson	75	74
Galois	80	75
Maxwell	63	62
Newton	62	70
Bernoulli	69	70
Gauss	75	72
Fermi	74	75

De acuerdo con lo que vimos previamente, la pendiente de la regresión lineal de progenie en progenitor es igual a la mitad de la heredabilidad (figura 10.7), por lo que un buen estimador de esta última es $\hat{\mathrm{h}}^2 = 2\ \mathrm{b_{HP}}$. El estimador **OLS** de la pendiente es

$$b_{HP} = \frac{Cov_{HP}}{V_P} = \frac{\sum_{n=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{n=1}^{N}(x_i - \bar{x})^2} \tag{10.22}$$

La media de los padres es $\bar{x} = 71{,}3$ cm, mientras que la media de hijos es $\bar{y} = 71{,}8$ cm.

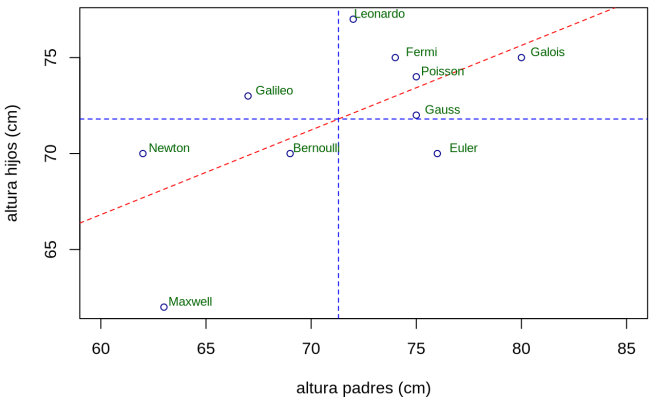


Figura 10.7. Regresión de los hijos en los padres para la altura a la cruz. Las medias fenotípicas de padres e hijos se corresponden con las líneas azules a trazos, mientras que la recta de regresión se marca en rojo.

Restando a los valores originales la correspondiente media, haciendo el cuadrado de los desvíos en padres y el producto cruzado de los desvíos, tenemos el siguiente cuadro

Padre	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
Euler	4,7	22,09	−1,39	−6,53
Leonardo	0,7	0,49	5,03	3,52
Galileo	−4,3	18,49	1,14	−4,89
Poisson	3,7	13,69	2,06	7,61
Galois	8,7	75,69	3,28	28,55
Maxwell	−8,3	68,89	−10,07	83,55
Newton	−9,3	86,49	−1,82	16,96
Bernoulli	−2,3	5,29	−1,31	3,02
Gauss	3,7	13,69	0,20	0,73
Fermi	2,7	7,29	2,89	7,79
SUMA				
	0,0	312,1	0,0	140,3

Por lo tanto, usando estos valores en la ecuación del estimador **OLS** de la pendiente obtenemos

$$b_{HP} = \frac{\sum_{n=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{n=1}^{N}(x_i - \bar{x})^2} = \frac{140{,}3\ \text{cm}^2}{312{,}1\ \text{cm}^2} = 0{,}44954 \tag{10.23}$$

Como $\hat{\text{h}}^2 = 2\ \text{b}_{\text{HP}}$, entonces $\hat{\text{h}}^2 = 2 \times 0{,}44954 \approx 0{,}899$. Como referencia, el proyecto **Framingham Heart Study**, sobre una cohorte de más de 5000 pacientes (humanos) con distintas mediciones en padres e hijos arrojó una estimación de la heredabilidad de la altura

(HT, por height) de $h^2_{HT} = 0{,}84 \pm 0{,}01$ (Byars et al. 2010), por lo que no parece tan descabellado el valor obtenido en nuestra «prospección» ficticia en perros.

Regresión del valor fenotípico de los hijos sobre el fenotipo de ambos progenitores

En características que se expresan en ambos sexos, es posible pensar la **regresión del fenotipo de los hijos (o su promedio) en el promedio del fenotipo de ambos progenitores.** Esta alternativa a la regresión en uno de los progenitores tiene la ventaja de que contempla posibles sesgos en la elección de las parejas (por ejemplo, cuando individuos altos buscan parejas altas, o cuando apareamos toros grandes con vacas grandes y vacas chicas con toros chicos). Claramente, además de que la característica debe expresarse en ambos sexos, las medidas deben poder promediarse. Esto plantea un problema cuando existen diferencias notorias en la distribución de la característica entre sexos. Por ejemplo, en muchas características productivas —como peso de vellón, ganancias diarias de peso, peso a determinada edad, etcétera— las diferencias entre sexos son notorias, tanto en medias como, en algunos casos, la varianza de la característica. Una posibilidad es considerar los fenotipos como desvíos del grupo contemporáneo del mismo sexo. Otra alternativa es aplicar algún tipo de factores de corrección que ponga las dos distribuciones en plano de igualdad. Más general, podrían incluirse todos los factores a corregir necesarios en un modelo lineal y de esta forma obtener una estimación corregida simultáneamente para todos ellos.

La pendiente de la regresión del fenotipo de hijos (o su promedio, todas las parejas con el mismo número de hijos) en el promedio de los dos padres estará dada por

$$b_{H\bar{P}} = \frac{Cov_{H\bar{P}}}{V_{\bar{P}}} \tag{10.24}$$

Precisamos ahora entender a qué equivale la covarianza fenotípica entre hijos (o su promedio dentro de padres) y el promedio del fenotipo de los padres. Si x_1 y x_2 son los padres, mientras que y es un hijo de la pareja (o el promedio de los mismos), entonces

$$Cov_{H\bar{P}} = Cov(\bar{y}, \frac{x_1 + x_2}{2}) = Cov(\bar{y}, \frac{x_1}{2}) + Cov(\bar{y}, \frac{x_2}{2}) = \tag{10.25}$$

$$= \frac{1}{2}Cov(\bar{y},x_1) + \frac{1}{2}Cov(\bar{y},x_2) = Cov_{HP} \ \therefore \tag{10.26}$$

$$Cov_{H\bar{P}} = \frac{1}{2}V_A \tag{10.27}$$

Por otro lado, asumiendo que la varianza fenotípica en ambos sexos es igual, la varianza del promedio fenotípico de los padres será igual a

$$V_{\bar{P}} = Var(\frac{x_1 + x_2}{2}) = Var(\frac{1}{2}(x_1 + x_2)) = \frac{1}{4}[Var(x_1) + Var(x_2)] = \tag{10.28}$$

$$= \frac{1}{4}[2\ Var(x)] = \frac{1}{4}[2\ V_P] = \frac{1}{2}V_P \tag{10.29}$$

Por lo tanto, sustituyendo los resultados de las ecuaciones (10.27) y (10.29) en (10.24), tenemos

$$\mathrm{b_{H\bar{P}}} = \frac{\frac{1}{2}\mathrm{V_A}}{\frac{1}{2}\mathrm{V_P}} = \frac{\mathrm{V_A}}{\mathrm{V_P}} = \mathrm{h}^2 \tag{10.30}$$

Por lo tanto, teniendo en cuenta este resultado, **la pendiente de la regresión del fenotipo de los hijos en el promedio del fenotipo de los padres es igual a la heredabilidad**, o en términos matemáticos $\mathrm{b_{H\bar{P}}} = \hat{\mathrm{h}}^2$

Precisión en las estimaciones de heredabilidad: varianza del estimador

Previamente demostramos que la regresión de los valores genéticos aditivos en los fenotípicos es igual a la heredabilidad. También vimos que regresiones de los fenotipos de la progenie en fenotipos de los progenitores son funciones lineales simples de la heredabilidad. En particular, la pendiente de las rectas de regresión son las que acarrean significado respecto a la heredabilidad y, por lo tanto, conocer la distribución del estimador resulta fundamental para entender cómo varía y establecer límites para la inferencia. Es decir, si conocemos las propiedades distribucionales del estimador y en particular el primer y segundo momento (media y varianza), podemos establecer un intervalo de confianza para nuestras estimaciones de la pendiente y, por lo tanto, de la heredabilidad. En general, no tiene mucho sentido conocer un valor puntual de estimación de heredabilidad si no conocemos alguna medida de incertidumbre de este estimado que nos permita afirmarnos en su relevancia para nuestras proyecciones ulteriores. Claramente, no es lo mismo decir que un estimado de heredabilidad es de $h^2 = 0{,}45 \pm 0{,}05$ (desvíos estándar) que $h^2 = 0{,}45 \pm 0{,}20$: mientras en el segundo caso (considerando una distribución aproximadamente normal), si el muestreo se repitiera infinitas veces, contendría el «verdadero valor» de la heredabilidad en, por ejemplo, 95 % de los intervalos calculados, entre $0{,}05 < h^2 < 0{,}85$, en el primero el intervalo se reduce a $0{,}35 < h^2 < 0{,}55$.

Una excelente introducción a este tema se encuentra en el libro Wooldrigde (2013), del que tomaremos prestada la notación inicial. Para encontrar estos momentos en el estimador de mínimos cuadrados ordinarios, partimos del modelo lineal corriente, es decir

$$y_i = \beta_0 + \beta_1\, x_i + e_i \tag{10.31}$$

con $e_i \sim N(0,\sigma_e^2)$. Antes de poder derivar la varianza del estimador de β_1 (la pendiente de la recta), es decir $\hat{\beta}_1$, vamos a demostrar un par de resultados accesorios. En primer lugar, vamos a demostrar que $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$, un resultado que puede

parecer a primera vista contraintuitivo. Sin embargo, si operamos con el lado izquierdo de la igualdad, tenemos

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\bar{x}y_i+\bar{x}\bar{y}-x_i\bar{y}= \tag{10.32}$$

$$\sum_{i=1}^{n}x_iy_i-\bar{x}y_i+\sum_{i=1}^{n}\bar{x}\bar{y}-x_i\bar{y}=\sum_{i=1}^{n}x_iy_i-\bar{x}y_i+n\bar{x}\bar{y}-\sum_{i=1}^{n}x_i\bar{y}= \tag{10.33}$$

$$\sum_{i=1}^{n}x_iy_i-\bar{x}y_i+n\bar{x}\bar{y}-\bar{y}\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i-\bar{x}y_i+n\bar{x}\bar{y}-\bar{y}(n\bar{x})= \tag{10.34}$$

$$\sum_{i=1}^{n}x_iy_i-\bar{x}y_i=\sum_{i=1}^{n}(x_i-\bar{x})y_i\ \therefore \tag{10.35}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}(x_i-\bar{x})y_i \tag{10.36}$$

que es lo que queríamos demostrar. Pensándolo un poco mejor, no debería resultarnos extraño ni contraintuitivo ya que $y_i-\bar{y}$ es lo mismo que desplazar los datos de la variable dependiente una cantidad fija, lo que alterará el intercepto, pero no la pendiente de la recta. A partir de esto, si aplicamos la igualdad de la ecuación (10.36) en el numerador de la definición clásica del estimador **OLS** (por *ordinary least squares*, o mínimos cuadrados ordinarios), tenemos que

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{10.37}$$

Si por conveniencia usamos la notación siguiente:

$$STC_x=\sum_{i=1}^{n}(x-\bar{x})^2 \tag{10.38}$$

entonces, sustituyendo y_i por $\beta_0+\beta_1\ x_i+e_i$ (el lado derecho de la ecuación (10.31)), el estimador **OLS** de la pendiente de la recta puede escribirse como

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{STC_X}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1\ x_i+e_i)}{STC_X} \tag{10.39}$$

Trabajemos ahora con el numerador de (10.39). Operando, tenemos

$$\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1\ x_i+e_i)=\beta_0\ \sum_{i=1}^{n}(x_i-\bar{x})+\beta_1\ \sum_{i=1}^{n}(x_i-\bar{x})x_i+\sum_{i=1}^{n}(x_i-\bar{x})e_i \tag{10.40}$$

El resultado de la ecuación (10.40) puede parecer complejo, pero como $\sum_{i=1}^{n}(x_i - \bar{x}) = (\sum_{i=1}^{n} x_i) - n\bar{x} = n\bar{x} - n\bar{x} = 0$, entonces el primer término del lado derecho de la ecuación (10.40) es cero. Utilizando el resultado de la ecuación (10.36) y la definición de (10.38), el segundo término es igual a

$$\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i = \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \beta_1 STC_X \qquad (10.41)$$

Si usamos la siguiente definición para simplificar la notación

$$d_i = x_i - \bar{x} \qquad (10.42)$$

El numerador de la ecuación (10.39) queda

$$\beta_1 STC_X + \sum_{i=1}^{n} d_i e_i \qquad (10.43)$$

por lo que sustituyendo esto por el numerador de la ecuación (10.39), tenemos ahora que

$$\hat{\beta}_1 = \frac{\beta_1 STC_X + \sum_{i=1}^{n} d_i e_i}{STC_X} = \beta_1 + (1/STC_X)\sum_{i=1}^{n} d_i e_i \qquad (10.44)$$

que es nuestro segundo resultado previo. El resultado de la ecuación (10.44) nos indica que nuestro estimador de la pendiente de regresión será igual al «verdadero valor de la pendiente» (β_1), más un término aleatorio que es una combinación lineal de los errores. De hecho, esto nos permite demostrar fácilmente que condicional a la muestra (x_1,x_2,x_3...,x_n), $\hat{\beta}_1$ es un estimador insesgado en **OLS** de β_1

$$\begin{aligned} E(\hat{\beta}_1) = E(\beta_1 + (1/STC_X)\sum_{i=1}^{n} d_i e_i) = \\ \beta_1 + E((1/STC_X)\sum_{i=1}^{n} d_i e_i) = \\ \beta_1 + (1/STC_X)\sum_{i=1}^{n} E(d_i e_i) = \\ \beta_1 + (1/STC_X)\sum_{i=1}^{n} d_i E(e_i) = \\ \beta_1 + (1/STC_X)\sum_{i=1}^{n} d_i \times 0 \ \therefore \\ E(\hat{\beta}_1) = \beta_1 \end{aligned} \qquad (10.45)$$

También se puede demostrar que $\hat{\beta}_0$ es un estimador insesgado de β_0, pero eso carece de importancia para nosotros ya que por un lado el intercepto no tiene significación genética

relevante en este contexto y por otro lado, en algunas regresiones al centrar la variable dependiente y la independiente, la recta debe pasar por el punto (0,0) (el origen del sistema cartesiano de ejes).

Veamos ahora si podemos derivar una expresión para la varianza del estimador de la pendiente, es decir $\hat{\beta}_1$. Si partimos de la ecuación (10.44)

$$\hat{\beta}_1 = \beta_1 + (1/STC_X)\sum_{i=1}^{n} d_i e_i \tag{10.46}$$

y teniendo en cuenta que β_1 es constante (el «verdadero valor» de la pendiente), mientras que condicional en los valores x_i, STC_X y d_i también lo son, entonces

$$Var(\hat{\beta}_1) = Var(\beta_1 + (1/STC_X)\sum_{i=1}^{n} d_i e_i) = (1/STC_X)^2 Var(\sum_{i=1}^{n} d_i e_i) \tag{10.47}$$

Además, considerando que los e_i son realizaciones independientes de una variable aleatoria (o sea, la misma distribución, con varianza σ_e^2), entonces la varianza de la suma es igual a la suma de las varianzas y por lo tanto

$$Var(\hat{\beta}_1) = (1/STC_X)^2 \sum_{i=1}^{n}(d_i^2 Var(e_i)) = (1/STC_X)^2 \sum_{i=1}^{n}(d_i^2 \sigma_e^2) = \tag{10.48}$$

$$\sigma_e^2 (1/STC_X)^2 \sum_{i=1}^{n} d_i^2 = \sigma_e^2 (1/STC_X)^2 STC_X \ \therefore \tag{10.49}$$

$$Var(\hat{\beta}_1) = \frac{\sigma_e^2}{STC_X} \tag{10.50}$$

Este último es un resultado fundamental para nosotros ya que venimos trabajando la heredabilidad como una regresión y, por lo tanto, necesitamos tener un estimador del error en nuestras estimaciones de heredabilidad.

Es momento de volver a la genética cuantitativa. Por comodidad, para simplificar la notación, a partir de acá definimos $b \equiv \hat{\beta}_1$, es decir que llamaremos b a nuestro estimador de la pendiente. Más aún, ahora X e Y representan dos variables aleatorias, por ejemplo, fenotipo de progenitores y fenotipos de progenie, y por lo tanto con varianzas σ_X^2 y σ_Y^2.

En primer lugar, dadas N observaciones, un estimador de la varianza en X a partir de las mismas es

$$\hat{\sigma}_X^2 = \frac{1}{N-2}\sum_{i=1}^{n}(x - \bar{x})^2 \Leftrightarrow STC_X = \sigma_X^2 (N-2) \tag{10.51}$$

Como vimos en la ecuación (10.50), utilizando la nueva notación, en **OLS** la varianza del estimador de la pendiente es

$$Var(b) = \sigma_b^2 = \frac{\sigma_e^2}{STC_X} \tag{10.52}$$

Por otro lado, en general es desconocida la varianza del error (el numerador de (10.52)), pero de acuerdo con nuestro modelo lineal

$$y_i = b\, x_i + e_i \Leftrightarrow e_i = y_i - b\, x_i \;\; \therefore \tag{10.53}$$

$$Var(e) = \hat{\sigma}_e^2 = Var(y_i - b\, x_i) = \sigma_Y^2 - Var(b\, x_i) \;\; \therefore \tag{10.54}$$

$$\hat{\sigma}_e^2 = \sigma_Y^2 - b^2\sigma_X^2 \tag{10.55}$$

Sustituyendo los resultados de las ecuaciones (10.51) y (10.55), condicional en el valor de b, en la ecuación (10.52), tenemos

$$\sigma_b^2 = \frac{\hat{\sigma}_e^2}{STC_X} = \frac{\sigma_Y^2 - b^2\sigma_X^2}{\sigma_X^2(N-2)} \;\; \therefore \tag{10.56}$$

$$\sigma_b^2 = \frac{1}{N-2}\left[\frac{\sigma_Y^2}{\sigma_X^2} - \frac{b^2\sigma_X^2}{\sigma_X^2}\right] \tag{10.57}$$

Finalmente, operando sobre (10.57), llegamos a una forma de la varianza del estimador de la pendiente de regresión que se presta para trabajarlo y sacar conclusiones.

$$\sigma_b^2 = \frac{1}{N-2}\left[\frac{\sigma_Y^2}{\sigma_X^2} - b^2\right] \tag{10.58}$$

Un tema donde la ecuación (10.59) puede ser de relevancia es a la hora de diseñar experimentos para la estimación de la heredabilidad, buscando optimizar la combinación de progenitores y progenie dadas determinadas capacidades experimentales (entendido esto como la varianza mínima del estimador). En ese sentido, Douglas S. Falconer (1983) plantea que asumiendo que b es por lo general pequeño, b^2 lo es mucho más y, por lo tanto, asumiendo además que N es grande, una aproximación razonable de la varianza del estimador sería

$$\sigma_b^2 \approx \frac{1}{N}\frac{\sigma_Y^2}{\sigma_X^2} \tag{10.59}$$

Para ser más específicos, supongamos que vamos a estimar la heredabilidad de una característica a partir de la regresión de $n \geqslant 1$ hijos para alguna combinación de $k = 1{,}2$ padres (por ejemplo, la media de n hijos para cada media de k padres) y que tenemos N pares de esas medidas. El total de individuos medidos será Nn progenies más kN progenitores ($k = 1$ si se mide uno solo, o $k = 2$ si se miden los dos), por lo que los totales de individuos medidos serán $T = N(n + k)$. Por ejemplo, cuando se trata de la regresión de un hijo en

uno de sus padres (asumiendo que las varianzas fenotípicas son iguales en ambos sexos), la varianza de los progenitores (σ^2_X) es igual a la varianza fenotípica V_P.

Por otro lado, (σ^2_Y) es igual a la varianza de las medias de familias de n individuos. Como demostraremos más adelante «La reducción en la varianza fenotípica con varias medidas», la varianza de la media familiar es igual a

$$\sigma^2_Y = \frac{1+(n-1)t}{n} V_P \tag{10.60}$$

donde t es el **coeficiente de correlación intraclase** $t = r_{A_1A_2}h^2$, es decir el parentesco aditivo entre miembros de la familia ($r_{A_1A_2}$) multiplicado por la heredabilidad. Ahora, sustituyendo σ^2_Y en la ecuación (10.59), tenemos

$$\sigma^2_b \approx \frac{1}{N} \frac{\frac{1+(n-1)t}{n} V_P}{V_P} = \frac{1+(n-1)t}{nN} \tag{10.61}$$

Más generalmente, en el caso del promedio de los dos progenitores, $\sigma^2_X = \frac{V_P}{k}$, con $k = 2$, por lo que de acuerdo con Falconer y Mackay (1996), la varianza del estimador será

$$\sigma^2_b \approx \frac{1}{N} \frac{\frac{1+(n-1)t}{n} V_P}{\frac{V_P}{k}} = \frac{k[1+(n-1)t]}{Nn} \tag{10.62}$$

En general $t = r_{A_1A_2}h^2$ no se conoce de antemano ya que no conocemos la heredabilidad y t es función de la misma. Si pretendemos usar la varianza del estimador para planificar un experimento de forma óptima, variando el número y tamaños de las familias, habrá que suponer diferentes valores de la heredabilidad y ver qué ocurre bajo los diferentes supuestos.

PARA RECORDAR

- El valor de la pendiente de la recta de regresión del valor de cría (valor aditivo) en el fenotipo individual es igual a la heredabilidad (h^2) puesto que $b_{AP} = \frac{Cov(A,P)}{V_P} = \frac{Cov(A,A)+Cov(A,D)+Cov(A,I)+Cov(A,E)}{V_P} = \frac{Cov(A,A)}{V_P} = \frac{V_A}{V_P} = h^2$
- Cuando predecimos el valor de cría a partir del valor fenotípico del individuo, **la precisión es la raíz cuadrada de la heredabilidad** ($\sqrt{h^2} = h$).
- Podemos describir la regresión del fenotipo de un hijo en el fenotipo de **un progenitor** como $b_{HP} = \frac{Cov_{HP}}{V_P} = \frac{Cov(y,x_1)}{V_P}$ donde x_1 y x_2 a los progenitores y y a su progenie.
- Si además sabemos que la covarianza entre hijos y padres es $Cov_{HP} = \frac{1}{2}Cov(A_{x_1},A_{x_1}) = \frac{1}{2}V_A$ entonces $b_{HP} = \frac{Cov_{HP}}{V_P} = \frac{\frac{1}{2}V_A}{V_P} = \frac{1}{2}\frac{V_A}{V_P} = \frac{1}{2}h^2$ Por lo tanto, la **pendiente de la regresión de los valores fenotípicos de una progenie en uno de sus progenitores igual a un medio de la heredabilidad.**
- Para realizar la regresión del fenotipo de los hijos en el **promedio del fenotipo de ambos progenitores**, la característica de interés debe expresarse en ambos sexos además de que las medidas deben ser promediables. La pendiente de la regresión será $b_{H\bar{P}} = \frac{Cov_{H\bar{P}}}{V_{\bar{P}}} = \frac{\frac{1}{2}V_A}{\frac{1}{2}V_P} = \frac{V_A}{V_P} = h^2$ por lo tanto la **pendiente**

de la regresión del fenotipo de los hijos en el promedio del fenotipo de los padres es igual a la heredabilidad.

- La varianza del estimador de la pendiente de la regresión es $Var(b) = \sigma_b^2 = \frac{\sigma_e^2}{STC_X}$

10.2 Heredabilidad en sentido amplio y sentido estricto

En especies animales la clonación no es una realidad extendida, ni en la producción ni en el mejoramiento, pero en vegetales o en procariotas sí lo es. Los clones son individuos idénticos desde el punto de vista genético (al menos si ignoramos la **epigenética**) y, por lo tanto, comparten tanto los efectos aditivos de todos sus alelos como las interacciones entre los mismos, dentro de locus (efectos de dominancia) y entre loci (epistasis).

Supongamos que de una planta extraemos muchos clones y los mismos crecen en diferentes partes de un campo. Como todos tienen la misma base genética, las diferencias que podamos observar en las características se corresponden con diferencias en el ambiente. Por otra parte, si colocamos diferentes individuos de la especie (elegidos al azar, por ejemplo) en un ambiente extremadamente homogéneo, de laboratorio, prácticamente toda la varianza fenotípica será de origen genético ya que no existe varianza ambiental. Esto nos lleva a la posibilidad de definir un nuevo parámetro genético que tenga en cuenta la **determinación genética** de la característica. Llamamos **heredabilidad en sentido amplio** o **coeficiente de determinación genética** (y la notamos como H^2) a la proporción de la varianza fenotípica que se corresponde a los efectos genéticos, es decir

$$H^2 = \frac{V_G}{V_P} = \frac{V_A + V_D + V_I}{V_A + V_D + V_I + V_E} \tag{10.63}$$

Como las varianzas son siempre mayores o iguales a cero, entonces $H^2 \geqslant h^2$. Es decir

$$H^2 = \frac{V_A + V_D + V_I}{V_A + V_D + V_I + V_E} = \frac{V_A}{V_A + V_D + V_I + V_E} + \frac{V_D + V_I}{V_A + V_D + V_I + V_E} \quad \therefore \tag{10.64}$$

$$H^2 = h^2 + \frac{V_D + V_I}{V_A + V_D + V_I + V_E} \tag{10.65}$$

y como $\frac{V_D + V_I}{V_A + V_D + V_I + V_E} \geqslant 0$, entonces $H^2 \geqslant h^2$. Esto tiene mucho sentido intuitivo, ya que en el concepto de heredabilidad en sentido amplio tenemos en cuenta no solo los efectos aditivos de los alelos sino también los de sus combinaciones. En animales (incluido el ser humano) los gemelos monocigóticos son un buen ejemplo de individuos que comparten toda la genética ya que despreciando nuevamente la epigenética son idénticos desde el punto de vista genético.

Pero los gemelos no solo comparten la genética. Claramente, tanto los gemelos monocigóticos como los dicigóticos (también conocidos como **mellizos**) comparten ambiente en mucho mayor medida que el resto de los hermanos. Esto nos sugiere que comparando gemelos monocigóticos con dicigóticos y con otros hermanos enteros podríamos llegar a estimar la

proporción de varianzas de interacción genética (dominancia y epistasis en conjunto). Desafortunadamente, existen algunas otras complicaciones metodológicas adicionales, así como la dificultad en encontrar suficiente número de casos, por lo que se trata de una técnica con alcance bastante limitado en animales.

En plantas resulta muy sencillo ver su utilidad, así como la forma de estimarla. Supongamos que, sin pérdida de generalidad, una característica de nuestro interés se encuentra determinada por 3 loci, sin ligamiento, con dos alelos en cada uno. Supongamos que llamamos a los dos alelos alternativos para los loci A, B y C, A_1 y A_2, B_1 y B_2, C_1 y C_2 respectivamente. En vegetales es relativamente sencillo obtener **líneas puras** (*inbred lines*, en inglés) sin comprometer demasiado la viabilidad de las mismas. Supongamos que para la característica de nuestro interés tenemos dos **líneas puras** que son contrastantes para los alelos de los loci con influencia en la característica. Es decir, llamaremos $\mathbf{P_1}$ línea pura que tiene la configuración $A_1A_1/B_1B_1/C_1C_1$ y $\mathbf{P_2}$ a la línea con configuración genómica $A_2A_2/B_2B_2/C_2C_2$. Claramente, el cruzamiento de ambas líneas será $\mathbf{P_1} \times \mathbf{P_2} = \mathbf{F_1}$, cuyos individuos tendrán una configuración genómica idéntica en los loci de interés, es decir $A_1A_2/B_1B_2/C_1C_2$.

Como en la $\mathbf{F_1}$ todos los individuos son idénticos desde el punto de vista genético (al menos para los loci que tienen influencia en la característica), no existirá varianza de origen genético y toda la varianza fenotípica se puede adscribir a la varianza ambiental. Dicho de otra forma, como $\mathrm{V_G} = 0$, entonces $\mathrm{V_P} = \mathrm{V_G} + \mathrm{V_E} = 0 + \mathrm{V_E} \therefore \mathrm{V_P} = \mathrm{V_E}$. Por lo tanto, ya tenemos una estimación de la varianza ambiental. Ahora, si cruzamos entre sí a la $\mathbf{F_1}$ tenemos $\mathbf{F_1} \times \mathbf{F_1} = \mathbf{F_2}$. Como asumimos que los loci no se encontraban ligados, entonces en cada uno de ellos alcanzaremos en una sola generación el equilibrio de Hardy-Weinberg con las proporciones $\mathbf{1} : \mathbf{2} : \mathbf{1}$ (por ejemplo, de A_1A_1, A_1A_2 y A_2A_2). A su vez, los tres genotipos de cada uno de los 3 loci se combinarán al azar (porque al no estar ligados asumimos que se distribuyen en forma independiente desde el punto de vista estadístico), por lo que en la $\mathbf{F_2}$ tendremos toda la varianza genética además de la ambiental. Si llamamos $\mathbf{V_{P_{F1}}}$ a la varianza fenotípica observada en la $\mathbf{F_1}$ y $\mathbf{V_{P_{F2}}}$ a la varianza fenotípica observada en la $\mathbf{F_2}$, entonces

$$\begin{aligned} V_E &= V_{P_{F1}} \\ V_{P_{F2}} = V_G + V_E = V_G + V_{P_{F1}} &\Leftrightarrow V_G = V_{P_{F2}} - V_{P_{F1}} \end{aligned} \tag{10.66}$$

Por lo tanto, combinando los resultados de la ecuación (10.66) con la definición de **heredabilidad en sentido amplio** de la ecuación (10.63), tenemos que

$$\hat{H}^2 = \frac{V_G}{V_P} = \frac{V_{P_{F2}} - V_{P_{F1}}}{V_{P_{F2}}} \tag{10.67}$$

sería nuestro estimador del **coeficiente de determinación genética** o heredabilidad en sentido amplio. En la práctica, las estimaciones de heredabilidad en plantas suelen ser más complicadas, ya que casi siempre el fenotipo correspondiente a un genotipo es alguna suerte de media. Además, las evaluaciones se realizan normalmente en múltiples ambientes y durante muchos años, involucrando en muchos casos diseños no-balanceados, por lo que se requieren métodos de estimación más complejos, algunos de los cuales comparten la metodología **BLUP/BLUE** de uso corriente en animales (ver, por ejemplo, Schmidt et al. 2019).

Ejemplo 10.4

En un experimento para determinar la heredabilidad en sentido amplio del rendimiento en materia seca del maíz se cruzaron dos líneas puras contrastantes para rendimiento. En la $\mathbf{F_1}$ se observó una varianza fenotípica de 200 dt^2ha^2, mientras que en la $\mathbf{F_2}$ la varianza fenotípica observada fue de 1200 dt^2ha^2. Asumiendo que la característica se encuentra determinada por loci independientes, obtener un estimado de la heredabilidad en sentido amplio para la característica rendimiento de materia seca.

Usando la fórmula de la ecuación (10.67), un estimador de la heredabilidad lograda sería

$$\hat{H}^2 = \frac{V_{P_{F2}} - V_{P_{F1}}}{V_{P_{F2}}} \tag{10.68}$$

por lo que sustituyendo los valores correspondientes tenemos

$$\hat{H}^2 = \frac{[1200 - 200]\ dt^2 ha^2}{1200\ dt^2 ha^2} = \frac{1000}{1200} = 0{,}833 \tag{10.69}$$

PARA RECORDAR

- Llamamos **heredabilidad en sentido amplio** o **coeficiente de determinación genética** (y la notamos como H^2) a la proporción de la varianza fenotípica que se corresponde a los efectos genéticos: $H^2 = \frac{V_G}{V_P} = \frac{V_A+V_D+V_I}{V_A+V_D+V_I+V_E}$
- Dado que $\mathrm{H^2 = h^2 + \frac{V_D+V_I}{V_A+V_D+V_I+V_E}}$ y que $\frac{\mathrm{V_D+V_I}}{\mathrm{V_A+V_D+V_I+V_E}} \geqslant 0$, entonces $\mathrm{H^2 \geqslant h^2}$
- El estimador de la heredabilidad en sentido amplio será $\hat{H}^2 = \frac{V_G}{V_P} = \frac{V_{P_{F2}} - V_{P_{F1}}}{V_{P_{F2}}}$

10.3 Heredabilidad lograda

Una perspectiva diferente de la heredabilidad, estrechamente ligada a la selección direccional es la que se conoce como **heredabilidad lograda** o **heredabilidad realizada**. Si para una característica determinada, que supongamos tiene distribución normal con media μ_0, seleccionamos una proporción determinada de los mejores animales como progenitores de la siguiente generación, este grupo de animales seleccionados tendrá una media μ_S diferente a la de la población. Por ejemplo, si mayores valores de la característica son favorables, es decir cuanto más a la derecha en la distribución mejor el valor, entonces los animales seleccionados tendrán una media $\mu_S > \mu_0$. **A la diferencia entre la media de los individuos seleccionados y la media de la población de la que estos forman parte le llamamos diferencial de selección** y lo notamos con la letra S, es decir $S = \mu_S - \mu_0$.

En general, como el fenotipo está compuesto de aportes genéticos y ambientales, como vimos en el **modelo genético básico**, asumiendo que el ambiente en padres e hijos no se encuentra correlacionado, no toda la diferencia de los padres respecto a la media de su generación se transmitirá a los hijos. De hecho, lo único transmisible en esas condiciones son los efectos aditivos, por lo que la respuesta a la selección R, será menor que la diferencia en la generación parental. Formalmente, la respuesta a la selección es $R = \mu_1 - \mu_0$, es decir la diferencia entre la media de la población luego de la selección respecto a antes

de la selección. En la figura 10.8 podemos observar una representación de esta situación. En el eje de las abscisas tenemos el **diferencial de selección**, obviamente en la generación parental, mientras que en el eje de las ordenadas observamos la **respuesta a la selección** R. Claramente, la respuesta es menor que el diferencial, es decir $R \leqslant S$. Si $R = S$, entonces toda la diferencia en padres se observaría en hijos y el ángulo α entre el eje de las abscisas y la recta con pendiente R/S sería de 45° ya que ambos catetos del triángulo rectángulo son iguales (R y S). Cuanto más pequeño es α (más acostada está la recta roja), más pequeña es la respuesta R en proporción a S, es decir menor es R/S. Esto nos indica claramente que existe una asociación directa entre el ángulo α y la relación R/S.

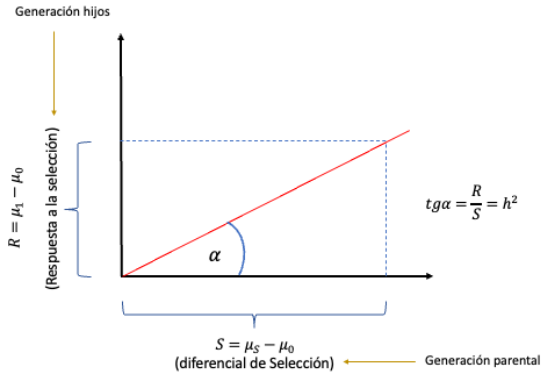


Figura 10.8. Representación gráfica de la heredabilidad lograda (o realizada). En el eje de las abscisas se marca el diferencial de selección ($S = \mu_S - \mu_0$) en la generación parental (la diferencia entre la media de los animales seleccionados y la media de la generación) mientras que en las ordenadas aparece la respuesta a la selección ($R = \mu_1 - \mu_0$), es decir la diferencia entre la media de la nueva generación (μ_1) y la generación anterior (μ_0). La pendiente de la recta es igual a la tangente del ángulo que forma con la horizontal, es decir la relación $tg\alpha = R/S = h^2$. Notemos que el eje de las abscisas representa una diferencia en la generación parental, mientras que el de las ordenadas representa una diferencia en la generación de los hijos.

Pero como vimos antes, la relación R/S es la heredabilidad de la característica. Por lo tanto, como la pendiente de la recta con ángulo α es igual a la tangente de dicho ángulo, es decir $tg\alpha$, entonces

$$tg\alpha = \frac{\text{cateto opuesto}}{\text{cateto adyacente}} = \frac{R}{S} = \frac{\mu_1 - \mu_0}{\mu_S - \mu_0} = h^2 \tag{10.70}$$

Como de esta forma la heredabilidad queda definida como la relación entre la mejora lograda (R) y la mejora pretendida (S), a esta forma de estimarla se le conoce como **heredabilidad lograda**. En general, las variaciones ambientales entre diferentes generaciones pueden introducir cambios relativamente importantes en las estimaciones de generación en generación. Más aún, las estimaciones sucesivas pueden seguir cierta tendencia que se aparte de una recta, por ejemplo, a causa de una disminución sostenida de la varianza aditiva debida al proceso de selección. Por esta razón se suele repetir el proceso durante varias generaciones. Esto puede llevar a la necesidad de tiempos enormes si el intervalo generacional de la especie es grande, como ocurre en general en muchas especies de animales grandes.

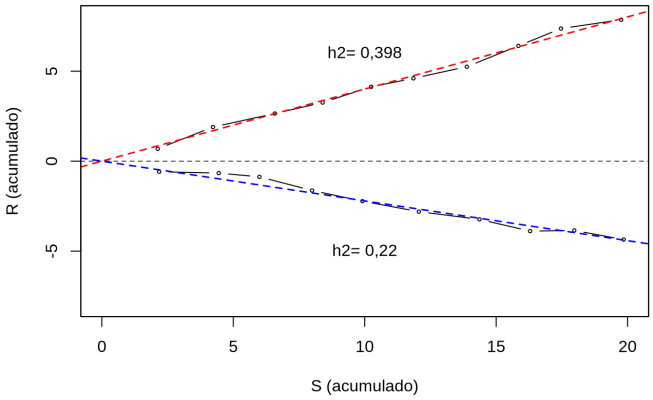


Figura 10.9. Estimación de la heredabilidad lograda a partir del diferencial de selección acumulado y de la correspondiente respuesta a la selección. A cada generación agregamos un nuevo punto como el diferencial de selección agregado al anterior (en las abscisas) y lo mismo para la respuesta a la selección (en las ordenadas). Se realiza selección simultánea en sentidos opuestos. Conectando los puntos (negros) tenemos las líneas quebradas (en negro), que representan la variación entre generaciones de la relación R/S. Con el conjunto de puntos se puede hacer un ajuste por mínimos cuadrados (incrementando la característica, línea roja a trazos, decrementando la característica, línea azul a trazos) o directamente calcular la relación $\hat{\mathrm{h}}^2 = \mathrm{tg}\alpha = \frac{\mathrm{R_{acum}}}{\mathrm{S_{acum}}}$. A fin de poder representar la divergencia en el gráfico, para la línea descendente se gráfica el valor opuesto del diferencial de selección (o sea, $-S$).

En un artículo muy interesante, Hill (1972) discute diferentes alternativas para estimar la heredabilidad a partir del concepto de heredabilidad lograda. Uno de los métodos más simples es el propuesto por Douglas S. Falconer (1960), que consiste en la regresión de la respuesta acumulada en el diferencial de selección acumulado. La idea es bien sencilla y la representamos en la figura 10.9: en cada generación de selección vamos a obtener un diferencial de selección y una respuesta a la selección, por lo que vamos acumulando las dos variables (sumando cada nuevo valor al acumulado anterior) durante el número de generaciones que deseamos. Si bien las variaciones entre generaciones son notorias en la figura (pendientes de los segmentos de recta de color negro, entre puntos), la tendencia lineal general (considerando todos los puntos) es claramente estable y consistente. Un fenómeno interesante que ocurre en ciertas características y poblaciones es la asimetría en la respuesta a la selección cuando hacemos selección en direcciones opuestas. Una representación gráfica de este fenómeno se puede ver en la figura 10.9 donde la selección apuntando a disminuir el valor de la característica (línea de regresión azul) tiene diferente valor absoluto de la pendiente (por un tema de representación gráfica, para la línea descendiente graficamos la respuesta en función de $-S$). Douglas S. Falconer (1983) presenta un ejemplo extraído de un trabajo propio presentado en 1954 en el que realizó selección divergente para el peso a las 6 semanas durante 21 (descendente) y 28 (ascendente) generaciones. Los estimados de heredabilidad que obtuvo fueron de $\hat{h}^2_A = 0{,}175 \pm 0{,}0161$ para la línea ascendente, mientras que fue $\hat{h}^2_D = 0{,}518 \pm 0{,}0231$ para la descendente. Es decir, en el caso del peso a las 6 semanas en ratones la respuesta a la selección para disminuir el peso fue mucho más efectiva que para aumentarlo.

Esto plantea algunas interrogantes y nos sugiere precaución a la hora de interpretar estos resultados. En primer lugar, suponemos que se trata de la misma característica desde el punto de vista genético la que movemos hacia arriba o hacia abajo. En ese caso, surge a primera vista una contradicción con la definición de heredabilidad como relación entre la varianza aditiva y la fenotípica ya que, al menos al comienzo del experimento, las mismas son iguales sin importar la dirección de selección. Una alternativa es considerar el promedio de las dos heredabilidades obtenidas, como una forma de suavizar las diferencias, pero sin demasiado soporte teórico. Otra alternativa es considerar la heredabilidad lograda en el sentido de nuestro interés. Por ejemplo, si queremos disminuir el peso al nacimiento de los terneros, aunque sea interesante conocer el estimado de la heredabilidad de aumentar el peso al nacimiento, mucho más útil sería el estimado de disminuir el peso. Aún en el caso de estudiar el comportamiento bajo selección en una dirección suele ser importante incluir una línea de control, es decir animales que no experimentan selección direccional a fin de controlar que no existan otros factores ambientales (como la dieta) que interfieran en los resultados.

Si bien el fenómeno de la asimetría en la respuesta a la selección se ha observado regularmente, las causas de la asimetría no son del todo claras. Douglas S. Falconer (1983) postula las siguientes:

1. Diferencial de selección que difiera entre las líneas ascendente y descendente por diferentes razones, entre ellas la coincidencia del sentido de selección natural y artificial, fertilidad incrementada en una dirección y el cambio de la varianza con la media que permitiría mayor diferencial a mayor varianza.
2. La «asimetría genética» producto de la distribución de los efectos de dominancia en un sentido (por ejemplo, hacia arriba) por los que a frecuencias intermedias de los alelos llevaría a respuestas mayores en el sentido en que los alelos tienden a ser recesivos.
3. La selección para los heterocigotas, cuando la selección en una dirección favorece a los heterocigotas en muchos loci o en pocos de gran efecto, a medida que nos acercamos a la frecuencia de equilibrio la respuesta es cada vez más lenta, mientras que alejándonos en dirección opuesta del «óptimo» avanzamos mucho más rápido hacia la fijación de alelos.
4. La depresión endogámica (algo que veremos en profundidad en otros capítulos), ya que la mayor parte de las poblaciones experimentales son pequeñas. Si existe depresión endogámica en la característica, la misma disminuirá la respuesta en un sentido y la incrementará en el otro.
5. Los efectos maternos. En algunas características como el peso de los ratones a las 6 semanas existen dos componentes, uno atribuible al ratón (el crecimiento desde el destete) y el otro casi totalmente materno (el peso al destete). Mientras que de acuerdo con Falconer no existieron diferencias en la heredabilidad de crecimiento, las diferencias se atribuyeron al peso al destete, lo que nos lleva a que la asimetría en peso a las 6 semanas refleja la asimetría en la «habilidad materna».

Si bien este método de estimar la varianza de una característica no es usual en poblaciones animales de interés comercial, el mismo sigue siendo válido para explorar la arquitectura genética de varias características y sigue siendo motivo de investigación metodológica (ver, por ejemplo, Lstibůrek et al. 2018).

Finalmente, como anticipación a lo que veremos en próximos capítulos, si manipulamos la definición de heredabilidad lograda tenemos que

$$\frac{R}{S} = h^2 \Leftrightarrow R = S\, h^2 \tag{10.71}$$

que se conoce como la **ecuación del criador** y es una de las piedras fundamentales del mejoramiento genético.

Ejemplo 10.5

Para estudiar la posibilidad de mejorar la producción de lana verde natural se condujo un estudio sobre la población original, separándola en mitades, una mitad bajo las condiciones normales de crecimiento y reproducción, la otra sujeta a selección direccional para aumentar la producción de lana. El cuadro siguiente contiene los resultados del experimento mantenido durante 7 generaciones parentales, en el cual se seleccionaron los animales reproductores basados en el peso sucio del segundo vellón (corrigiendo previamente por diferencias entre sexos). La pendiente de la regresión de acumulados para la línea control fue de cero.

Generación parental	**Diferencial selección** (S, g)	**Respuesta a la selección** (R, g)
0	715	475
1	770	485
2	745	410
3	815	405
4	790	375
5	766	395
6	825	325
TOTAL	5426	2870

Graficar la respuesta a la selección acumulada, el diferencial de selección acumulado, calcular la pendiente de la línea de regresión y graficarla. ¿Cuánto vale la heredabilidad estimada para esta característica? ¿Qué podría representar que el intercepto sea distinto de cero? ¿Es razonable el supuesto de linealidad de los acumulados? ¿Qué podría estar indicando la forma de la curva de datos reales?

Para resolverlo primero calculamos los acumulados, básicamente sumando todos los datos correspondientes hasta cada punto, lo que nos deja con el siguiente cuadro

Gener.	Dif. sel. acum. ($\sum_{i=0} S_i$, g)	Respuesta acum. ($\sum_{i=0} R_i$, g)
0	715	475
1	1485	960
2	2230	1370
3	3045	1775
4	3835	2150
5	4601	2545
6	5426	2870

La pendiente de la recta la calculamos a partir de los datos acumulados como

$$b = \frac{Cov(S_{acum}, R_{acum})}{Var(S_{acum})} = \frac{1455487\ \text{g}^2}{2873509\ \text{g}^2} = 0{,}50652 \tag{10.72}$$

La gráfica de los resultados aparece en la figura 10.10. Como la pendiente de la recta de regresión es nuestro estimado de heredabilidad, entonces tenemos que $\hat{\text{h}}^2 = \text{b} = 0{,}50652$. Claramente, en la gráfica siguiente se aprecia que el intercepto es distinto de cero. Asumiendo que la pendiente de regresión de la **línea control** fue de cero (como dice la letra), entonces no es de esperar que existan variaciones sistemáticas en el peso de vellón que no se deban a la selección, por lo que en principio descartaríamos ese factor. Al mirar el ajuste de la recta de regresión a los puntos podemos notar un leve pero consistente apartamiento hacia abajo de los puntos en los extremos. Eso pone en duda la linealidad de la curva y puede explicar el intercepto distinto de cero. Más aún, la forma de la curva real parece comenzar una tendencia hacia la horizontal a medida que pasan las generaciones, es decir una reducción sostenida en la respuesta a la selección, lo que podría indicar problemas futuros para seguir por este camino (algo que discutiremos más adelante, en próximos capítulos).

PARA RECORDAR

- Llamaremos **diferencial de selección** (S) a la diferencia entre la media de los individuos seleccionados y la media de la población de la que estos forman parte, es decir $S = \mu_S - \mu_0$.
- Llamaremos **respuesta a la selección** (R) a la diferencia entre la media de la población luego de la selección respecto a antes de la selección, es decir $R = \mu_1 - \mu_0$.
- La **heredabilidad lograda** queda definida como la relación entre la mejora lograda (R) y la mejora pretendida (S). Alternativamente, podemos pensar en la misma como la regresión de la respuesta acumulada en el diferencial de selección acumulado.
- Manipulando esta última definición llegamos a que $R = Sh^2$ conocida como la **ecuación del criador**.

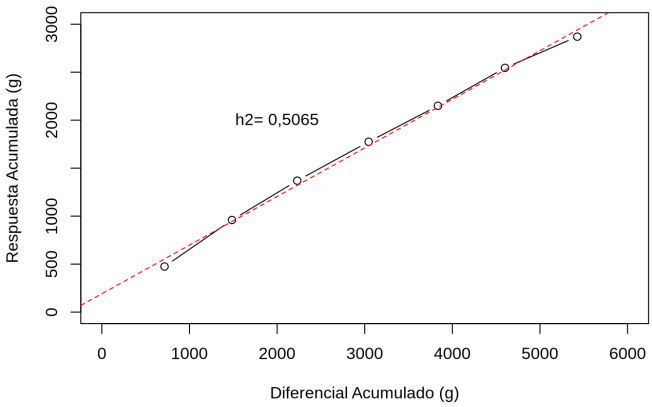


Figura 10.10. Estimación de la heredabilidad del peso de vellón (verde) sucio en ovejas verdes a través de la heredabilidad lograda. La línea roja a trazos representa la recta de regresión obtenida por mínimos cuadrados ordinarios.

10.4 Heredabilidad en poblaciones agronómicas o de laboratorio versus poblaciones naturales

Las estimaciones de heredabilidad son fundamentales, tanto para los programas de mejoramiento genético en plantas y animales, así como para el estudio de la evolución a corto y mediano de características cuantitativas en poblaciones naturales.

Las condiciones para las estimaciones de heredabilidad deben condecirse con el uso que se pretende de esta. Sin duda, en condiciones de laboratorio, o aun a campo, pero con condiciones ambientales sumamente controladas producirán por lo general estimados más altos que en condiciones naturales o sin ambientes controlados. La razón de esto es sencilla, ya que el denominador de la heredabilidad es la varianza fenotípica, uno de cuyos componentes es la varianza ambiental; en la medida en que la varianza ambiental es menor, menor será el denominador y, por lo tanto, mayor el cociente $h^2 = \frac{V_A}{V_P}$. Sin embargo, pretender basar un programa de mejora de una raza que mayormente presenta condiciones de expresión de la característica en ambientes sumamente variables a partir de las estimaciones realizadas en condiciones controladas es la receta ideal para el fracaso, o al menos para obtener logros mucho menores de los planificados. Sin embargo, para intentar entender la arquitectura genética de la característica o para imaginar el potencial evolutivo a corto y mediano plazo de una característica, posiblemente sea mejor idea obtener estimaciones en condiciones de laboratorio o muy controladas.

Por estas razones, los métodos y condiciones de estimación de la heredabilidad deben poder representar el uso pretendido. Además, bajo condiciones de selección moderada a fuerte, tanto la varianza genética como la fenotípica es posible que cambien, lo que implica cambios en los estimados de la heredabilidad. Esto lleva a la necesidad de estimar los parámetros genéticos en forma regular, particularmente en los programas de mejoramiento genético de

especies o razas con bajo número de participantes en las evaluaciones, dada la inestabilidad que introduce el bajo número de animales y registros.

En las subsecciones siguientes veremos algunos ejemplos locales en animales domésticos y plantas, pero un recurso interesante es la base de datos H^2DB, donde podemos escoger la especie de nuestro interés o el tipo de característica y obtener distintas estimaciones, la fuente de las mismas, así como diferentes gráficos que representan los resultados.

Estimaciones de heredabilidad en animales domésticos

La estimación de la heredabilidad de una característica suele ser fundamental a la hora de diseñar un programa de mejoramiento genético, o aún de entender la capacidad de evolucionar en una dirección de la misma. Por esta razón, una parte importante del trabajo de los genetistas y mejoradores suele ser la estimación de la heredabilidad en las condiciones lo más parecidas posible a las condiciones en las que se va a utilizar la estimación. Dicho de otra forma, no tiene mucho sentido obtener una estimación de la heredabilidad **en condiciones de laboratorio** para luego diseñar un programa de mejoramiento **a campo**. Resulta obvio que las estimaciones de laboratorio serán más precisas y seguramente más altas que las realizadas a campo, pero como el programa de mejoramiento involucra condiciones menos controladas los resultados obtenidos del programa de mejora no tendrán mucho que ver con nuestras expectativas. El contenido de esta sección no pretende ser una revisión de valores de heredabilidad de algunas características sino simplemente un motivador de la discusión de diferentes características de este parámetro genético tan importante.

Las estimaciones de heredabilidad en animales domésticos se pueden obtener de diferentes formas. Por un lado, como vimos en el capítulo «Parentesco y semejanza entre parientes», a partir de diferentes diseños experimentales es posible realizar estimaciones de la varianza aditiva, mientras que la varianza fenotípica se puede estimar directamente a partir de los datos, por lo que el cociente de ambos nos da una estimación de la heredabilidad ($h^2 = V_A/V_P$). Cada uno de estos diseños tiene sus ventajas y desventajas, pero algo común a todos ellos es que usualmente implican tener un gran número de animales controlados bajo condiciones similares. Por ejemplo, si pensamos en un diseño de medios hermanos, por cada padre que incluimos en el diseño debemos mantener varios hijos o hijas posteriormente. Como, a su vez, debemos incluir varios padres para obtener estimaciones razonables y que la muestra de machos sea representativa de los animales en la población, el número de animales a mantener en el experimento es necesariamente alto. Más aún, como vimos previamente, existen tensiones entre las propiedades deseables de los estimadores que se pueden obtener en los distintos diseños experimentales para estimar la varianza aditiva. En el caso de medios hermanos teníamos que la covarianza era igual a $\frac{1}{4}V_A$, por lo que era necesario multiplicar por **4** para tener un estimador insesgado, lo que inflaba el error de la estimación en la misma proporción y subsecuentemente la estimación de la heredabilidad. Por otra parte, el estimador que surgía de usar una estructura de hermanos enteros era producto de multiplicar la covarianza entre hermanos enteros por 2, pero con la consecuencia de incluir en este estimador también una parte de la varianza de dominancia y aún la del ambiente común. Nuestro estimador de la heredabilidad a partir del diseño de hermanos enteros será

$$\hat{h}^2 = \frac{2\,Cov_{P_{HE}}}{V_P} = \frac{\sigma^2_A + \frac{1}{2}\,\sigma^2_D + 2\,\sigma^2_{EC}}{V_P} = \frac{\sigma^2_A}{V_P} + \frac{\frac{1}{2}\,\sigma^2_D + 2\,\sigma^2_{EC}}{V_P} = h^2 + \frac{\frac{1}{2}\,\sigma^2_D + 2\,\sigma^2_{EC}}{V_P} \tag{10.73}$$

lo que claramente sesga el estimador hacia arriba pues el término $\frac{\frac{1}{2}\,\sigma^2_D + 2\,\sigma^2_{EC}}{V_P}$ es en general mayor a cero (es una relación de varianzas, por lo que al menos no debería ser negativo).

En vista de estas dificultades para la estimación a través de experimentos, como veremos más adelante en el curso y seguramente profundizarás en cursos siguientes, la forma corriente de hacerlo es mediante modelos estadísticos más elaborados que permiten incluir toda la información disponible de la estructura de parentesco entre todos los animales (los modelos mixtos lineales o aun los modelos mixtos lineales generalizables). Esta información se encuentra usualmente codificada en la matriz de parentesco, como vimos en el capítulo previo y, por lo tanto, puedo realizar las estimaciones a partir de todos los animales que tengan registros y sus parientes aún sin registros. De esta forma, el número de observaciones que puedo incorporar a la estimación de la heredabilidad se incrementa notoriamente (por ejemplo, todos los animales en una evaluación) y, por lo tanto, se incrementa también la precisión de la estimación. Además, bajo ciertas condiciones, los estimadores obtenidos cumplen con varias de las propiedades deseables de los estimadores.

En el cuadro siguiente se pueden apreciar estimaciones de heredabilidad para distintas características, en diferentes especies y razas dentro de las mismas, todas a partir de registros obtenidos en Uruguay. Por un lado, es fácil de apreciar que mientras que algunas características presentan heredabilidades bajas, como el **intervalo destete-concepción** en cerdos, o como el **intervalo interpartos** en bovinos de leche (raza holando en este caso), otras características poseen heredabilidades muy altas, como el **diámetro de fibras** en ovejas de la raza merino australiano. Claramente, las características que tienen una base genética compleja (donde lo aditivo no es tan importante), un ambiente extremadamente variable o interacciones entre genotipo y ambiente llevarán a menores estimaciones de la heredabilidad. Esto, a veces se confunde con el concepto de que baja heredabilidad es sinónimo de casi inexistencia de varianza aditiva (ya que es el numerador de la heredabilidad), pero esto es totalmente erróneo, como ya discutimos previamente. A la misma varianza aditiva de una característica, reducir el resto de los componentes de varianza, controlando el «ambiente» por ejemplo, llevará a un incremento concomitante en la heredabilidad. Dicho de otra forma, la baja heredabilidad de una característica está muchas veces asociada a la forma en que se registran los datos fenotípicos de la misma o a la definición imprecisa de lo que queremos medir (por ejemplo, en características reproductivas, difíciles de medir).

Por otro lado, aún en la misma característica y muy bien definida, como, por ejemplo, diámetro de fibras en ovinos, las estimaciones de heredabilidad pueden presentar diferencias muy importantes. Por ejemplo, se aprecia en el cuadro las diferencias entre la estimaciones para esta característica en las razas corriedale ($h^2 = 0{,}43 \pm 0{,}08$) y merino australiano ($h^2 = 0{,}82 \pm ?$). Si bien las razas podrían presentar diferencias notorias como las vistas, producto de diferentes bases genéticas en la característica (algo improbable) o de diferencias en la varianza genética existente (producto de la selección, por ejemplo), la causa que aparece como más probable acá tiene que ver con las condiciones en las que se realizaron las mediciones y el número de animales usados. Por otro lado, el trabajo de Ciappesoni et al. (2011) no reporta la incertidumbre de las estimaciones, por lo que resulta difícil comparar las distribuciones de las mismas. Esto resalta la importancia de las estimaciones de incertidumbre, ya que un valor puntual sin tener referencia de la distribución (dispersión, por ejemplo) no nos permite trabajar en forma confiable: no es lo mismo una heredabilidad de $h^2 = 0{,}500 \pm 0{,}025$ (desvío estándar) que $h^2 = 0{,}500 \pm 0{,}250$; en el primer caso (asumiendo

una distribución normal) el valor posible del parámetro con un 95 % de confianza será aproximadamente $0{,}450 \leqslant h^2 \leqslant 0{,}550$, mientras que en el segundo será $0{,}000 \leqslant h^2 \leqslant 1{,}000$ (es decir, casi no informativo).

Tabla 10.1. Heredabilidades para diferentes características en especies de animales domésticos, así como las correspondientes estimaciones de incertidumbre (EE=error estándar, SD=desvío estándar)

Raza	Característica	Año	h2	Incertidumbre	Tipo	Ref.
Cerdos						
Población multirracial	Número de lechones nacidos vivos	2015	0,13	0,06	EE	[1]
	Peso vivo del lechón al nacimiento		0,31	0,08		[1]
	Número de lechones destetados		0,14	0,06		[1]
	Peso vivo del lechón al destete		0,17	0,02		[1]
	Sobrevivencia durante la lactancia		0,10	0,01		[1]
	Intervalo destete-concepción		0,08	0,03		[1]
Ovinos						
Corriedale	Peso de vellón limpio	2016	0,42	0,09	SD	[2]
	Diámetro de fibra		0,43	0,08		[2]
	Presencia / Ausencia fibras meduladas		0,37	0,10		[2]
	Presencia / Ausencia fibras Kemp		0,63	0,09		[2]
	Presencia / Ausencia fibras pigmentadas		0,35	0,08		[2]
	HPG1 (ln)	2008	0,15			[3]
	HPG2 (ln)		0,28			[3]
Merino australiano	Resistencia a PGI en corderos (medida como HPG)	2012	0,25	0,03	SD	[4]
	Resistencia a PGI en ovejas periparturientas (medida como HPG)		0,08	0,03		[4]
	Peso de vellón limpio en borregos	2011	0,39			[5]
	Diámetro de fibra en borregos		0,67			[5]
	Peso de cuerpo en borregos		0,62			[5]
	Peso de vellón limpio en ovejas		0,38			[5]
	Diámetro de fibra en ovejas		0,82			[5]
	Peso de cuerpo en ovejas		0,58			[5]
Bovinos de leche						
Holando	Intervalo interparto	2012	0,07			[6]
	Producción de leche		0,3			[6]
	Días abiertos	2016	0,05			[7]
	Producción de leche		0,23			[7]
	Producción de grasa		0,21			[7]
	Producción de proteína		0,21			[7]

[1] Bell et al. 2015; [2] Sánchez et al. 2016; [3] Castells Montes 2008; [4] Goldberg, Ciappesoni y Aguilar 2012
[5] Ciappesoni et al. 2011; [6] Frioni 2012; [7] Frioni 2016

En algunos casos las características a mejorar son un poco más complejas y difíciles de especificar. En el cuadro siguiente aparecen algunas estimaciones de heredabilidad en caballos de competencia, en particular en competencias de resistencia (750 km en 15 días de competencia) en nuestro país. En primer lugar, en el caso de carreras de este tipo, si bien la ubicación en la clasificación y el tiempo total empleado parecen tener una parte importante de su base genética en común, no necesariamente la misma es total. Más aún, ambas características tienen distribuciones completamente diferentes, por lo que los modelos bajo los que se estima la heredabilidad no tienen por qué ser necesariamente equivalentes. Por esta razón es usual ensayar diferentes modelos y reportar los estimados para cada uno de ellos. En el cuadro siguiente, dado que López-Correa (2013) reportan varios resultados diferentes para diferentes modelos, nosotros aprovechamos esto para mostrar que en algunos casos las estimaciones son consistentes entre modelos (lo que aparece en el cuadro siguiente es el intervalo de las medias de los distintos modelos).

Variable	Especie/raza	h^2	Ref.
Competencias de resistencia (750 km/15 días)			
Clasificación	Caballos/Criollo	0,156(0,0510,273)	[1]
Tiempo final	Caballos/Criollo	0,10 – 0,13	[2]
Clasificación	Caballos/Criollo	0,07 – 0,10	[2]

[1] López-Correa et al. 2018; [2] López-Correa 2013 (intervalo de medias en los distintos modelos ensayados)

En otros casos las características a mejorar se encuentran «mal medidas» (en el sentido de poca precisión) o no hay forma razonable de medir directamente lo que deseamos mejorar. Usualmente es el caso de las características reproductivas, donde el conjunto de los procesos que llevan a la preñez o al nacimiento de la cría resultan imposibles de separar a nivel de establecimientos comerciales, lo que nos lleva a usar **variables poco informativas** (cría nacida, por ejemplo). Más aún, en condiciones de predios comerciales, en muchos casos se superpone el problema de la ausencia de algunos registros importantes, por lo que las variables son aún menos informativas. Por ejemplo, teniendo en cuenta los problemas anteriores, para estimar parámetros genéticos de variables reproductivas (***endpoints***) en la raza aberdeen angus, Naya, Peñagaricano y Urioste (2017) ensayaron diferentes modelos, incluyendo modelos lineales (gaussianos), de Poisson, probit (umbral), de Poisson censurados y gaussianos censurados. Los criterios tradicionales para rasgos de fertilidad incluyen el **éxito del parto** (**CS**), un rasgo binario codificado con 1 para el éxito y 0 para el fracaso, y uno algo más descriptivo, **días al parto** (**CD**), representado como el número de días desde el comienzo de la temporada de partos de un rodeo hasta la fecha de parto de la vaca. Mientras que **CD** es más descriptivo que **CS**, el mismo no permite una forma directa de representar la información de las vacas fallidas (que no tuvieron ternero). Existen diferentes alternativas para tratar este problema, alguna de las cuales incluyen **añadir un ciclo estral** a la variable (**CD+21d**), asumiendo que estas vacas estarían preñadas en el siguiente celo si se permitiese, tratar estos registros como información faltante (**CD_NA**) o considerar la CD como censurada a la derecha y modelizarla en consecuencia. Otra alternativa propuesta por Naya, Peñagaricano, and Urioste (2017) es la de modelar el **número total de celos fallidos** hasta la concepción (**FE**), como una nueva variable de respuesta no censurada de interés. La idea es de expresar en número de ciclos estrales (21 días) el tiempo en el que una vaca no está preñada, a pesar de estar en la temporada de cría; ventajas de esta variable incluyen que es cuantitativa, con el número de eventos relevantes (ciclos estrales, no días) y que permite conectar de forma continua los años en caso de haber fallado todas las oportunidades en uno.

En el cuadro siguiente se puede ver un resumen de las estimaciones de heredabilidad obtenidas por estos autores para los diferentes *endpoints* considerados, usando diferentes modelos de acuerdo con el tipo de variable.

Nombre	Modelo	Variable	h^2
FE.L	Linear	FE	0,052 (0,041 – 0,063)
FE.P	Poisson	FE	0,253 (0,221 – 0,284)
CS.L	Linear	CS	0,037 (0,031 – 0,042)
CS.Pr	Probit	CS	0,237 (0,195 – 0,277)
CD21.L	Linear	CD+21d	0,062 (0,050 – 0,074)
CD21.P	Poisson	CD+21d	0,282 (0,272 – 0,293)
CDNA.L	Linear	CD_NA	0,067 (0,053 – 0,084)
CDNA.P	Poisson	CD_NA	0,292 (0,281 – 0,304)
CD.CG	CenGaussian	CD	0,076 (0,062 – 0,090)
CD.CP	CenPoisson	CD	0,254 (0,244 – 0,264)

Claramente, en ningún caso las heredabilidades estimadas son altas o muy altas, manteniéndose siempre en el rango de bajas. Sin embargo, hay diferencias importantes entre trabajar una característica con heredabilidad del entorno de $h^2 = 0{,}05$, que de $h^2 = 0{,}25$. Por un lado, los modelos **no-lineales** mostraron sistemáticamente estimaciones más altas de heredabilidad en todos los casos que los modelos **lineales** ($h^2 < 0{,}08$ para los modelos lineales; $h^2 > 0{,}23$ para los modelos no lineales). Por otro lado, mientras que en el caso de los modelos relacionados con **FE** y el **CS**, los modelos no lineales superaron a sus homólogos lineales en el ajuste, en el caso de los modelos derivados del CD, las versiones lineales mostraron un mejor ajuste que las homólogas no lineales. Más allá de esto, la interpretabilidad del modelo es una cuestión fundamental y en el caso de los modelos **no-lineales** la misma suele verse comprometida. Finalmente, este estudio también muestra (algo que discutiremos en otros capítulos) que la **correlación genética** entre los *endpoints* considerados fue alta, pero lejos de ser igual a 1, lo que habla de que diferentes **arquitecturas genéticas** trabajarían sobre los diferentes *endpoints*.

Estimaciones de heredabilidad en plantas

Como ya vimos, en plantas los métodos de estimación pueden ser diferentes que en animales, particularmente por la posibilidad de estimar en forma relativamente la heredabilidad en sentido amplio y a su vez explotar su valor a partir de la clonación.

En el cuadro siguiente vemos algunos ejemplos de estimaciones de heredabilidad para dos especies de interés comercial, el *Eucalyptus*, especie de interés para la producción forestal y el **arándano**, una especie frutal.

Tabla 10.2. Heredabilidades para diferentes características en especies de plantas de interés comercial, así como las correspondientes estimaciones de incertidumbre (EE=error estándar, SD=desvío estándar).

Especie	Experimento	Característica	Año	h2	Incertidumbre	Ref.
Eucalyptus						
Eucalyptus	E. grandis × E. urophyla	Diámetro a la altura del pecho	2019	0,41	0,15	[1]
		Altura		0,09	0,13	
		Volumen		0,46	0,14	
		Celulosa		0,32	0,11	
		Hemicelulosa		0,33	0,09	
		Lignina insoluble		0,58	0,12	
		Lignina soluble		0,87	0,13	
		Lignina total		0,57	0,12	
		Densidad madera		0,7	0,16	
		Ángulo microfibra		0,11	0,11	
		Largo fibra		0,36	0,2	
		Ancho fibra		0,08	0,08	
Arándanos						
Arándanos	Arándanos (tetraploide)	Rendimiento	2018	0,45		[2]
		Peso fruto		0,58		
		Firmeza fruto		0,35		
		Diámetro fruto		0,25		
		pH		0,36		
		Sólidos solubles		0,33		
		Densidad brotes		0,16		

[1] Marco de Lima et al. 2019; [2] Cellon et al. 2018

Estimaciones de heredabilidad en condiciones naturales

Los métodos de estimación de la heredabilidad que hemos visto hasta ahora, tanto por análisis de la varianza como por regresión implican estructuras de parentesco conocidas y determinadas. En muchos casos esto puede significar un impedimento importante a la hora de obtener estimados de heredabilidad en condiciones naturales. Por un lado, la determinación de parentesco puede resultar difícil o aún imposible en especies donde no resulta fácil identificar las relaciones progenitor-progenie, tanto debido a que durante la estación de cría una hembra se puede aparear con diferentes machos o que en vegetales la polinización es al azar y la dispersión de semillas elevada u otras razones similares. Por otro lado, conseguir implementar un diseño experimental con determinadas estructuras de parentesco en condiciones naturales suele ser imposible, lo que nos llevará a usar las pocas observaciones que cumplan con los criterios de parentesco. Por lo tanto, las estimaciones de heredabilidad en estas condiciones suelen incluir pocos individuos y un trabajo de observación considerable.

Pese a la dificultad de generar experimentos, como veremos en «Métodos más avanzados de estimación», existen alternativas para usar la información de todos los individuos si conocemos o podemos estimar el parentesco entre los mismos. En la mayor parte de las especies de aves, por ejemplo, los huevos del nido pertenecen todos a la misma pareja de padres, por lo que se trata de hermanos enteros. En caso de características que se pueden medir en individuos juveniles (antes de abandonar el nido), resulta muchas veces posible utilizar este diseño de **hermanos enteros** que nos brinda la naturaleza para estimar a través de un modelo de ANOVA la varianza genética de la característica y de allí la heredabilidad.

En muchos casos, sin embargo, no es posible aprovecharse de la historia de vida de la especie para poder explotar determinadas relaciones de parentesco. Hasta hace poco eso constituía una importante limitante aún para especies en las que la medición del fenotipo era razonablemente posible. Sin embargo, la aparición de los métodos de identificación de parentesco a partir del ADN, en particular los métodos de genotipado y secuenciado han permitido analizar un grupo de individuos escogidos al azar, utilizando las mismas herramientas de estimación a través de métodos por máxima verosimilitud o métodos bayesianos (ver «Métodos más avanzados de estimación»).

10.5 Heredabilidad y filogenética

Como a esta altura debería quedar claro, existe una clara conexión entre la genética de poblaciones y la genética cuantitativa, una relación que se ve mucho más clara cuando ambas se consideran en un contexto evolutivo. En la «ciencia moderna», con investigadores cada día más enfrascados en sus respectivas especialidades (a su vez cada día más específicas) resulta muchas veces difícil escapar a las limitaciones propias y hacer conexiones con desarrollos de otras áreas del conocimiento. Sin embargo, cada tanto tiempo aparecen investigadores con curiosidad más amplia que su propio foco y que logran arrimar estos desarrollos externos a su área. Un caso de estos lo constituye el aporte de M. Lynch (1991), que aplicó con creatividad los desarrollos observados en la genética cuantitativa animal para comprender mejor la evolución de los caracteres cuantitativos a escala evolutiva.

La evolución de los organismos se suele representar en forma de árboles filogenéticos donde un clado ancestral se deriva en dos clados diferentes, que pueden continuar evolucionando sin dar lugar a nuevos clados o nuevamente particionarse y así sucesivamente. Esto genera una estructura que se puede representar en forma de árbol y que, por lo tanto, se conocen como **árboles filogenéticos**. Un problema típico para estudiar los caracteres cuantitativos en evolución es que uno de los supuestos tradicionales que se hacen en estadística para la mayor parte de las pruebas es que las observaciones son independientes e igualmente distribuidas. Claramente, como las especies derivan de ancestros comunes, que para cualquier característica de interés tenían determinados valores, entonces cada par de especies tiene un largo diferente de ramas hasta llegar a ese ancestro en común entre ellas (ese ancestro puede ser diferente para cada par de especies). Eso implica que han evolucionado en conjunto el tiempo desde el origen de la vida hasta el ancestro en común a ambas, o dicho de otra manera, la divergencia entre esas dos especies solo ocurre a partir del último ancestro en común. Por lo tanto, una medida intuitiva de la similaridad genética de dos especies dado que conocemos un árbol filogenético que las representa sería la proporción del largo de ramas compartido entre las dos especies dividido el largo de ramas total desde el origen (raíz del árbol) hasta las puntas (nuestras especies). Si tenemos n especies, podemos acomodar la similaridad filogenética de cada par en una matriz cuadrada de $n \times n$, simétrica. Si

los largos totales de ramas desde el origen a las puntas (especies) son iguales, entonces la diagonal estará constituida por valores de 1. En caso de que no sean iguales los valores de la diagonal pueden reflejar esta relación entre «el tiempo compartido consigo mismo» y la similaridad genética haciendo que el valor de la diagonal sea distinto de 1.

La similaridad entre la matriz que representa la similaridad filogenética y la matriz de parentesco entre especies es evidente. En principio, desde el punto de vista matemático, el diagrama de flechas que usamos para representar el pedigrí y el árbol filogenético con raíz son ambos **grafos dirigidos acíclicos** (**DAG** por sus iniciales en inglés). Los grafos son estructuras matemáticas que tienen **vértices** (o nodos), que se conectan entre sí a través de **aristas**. Cuando las aristas son **dirigidas**, es decir **apuntan desde un vértice a otro**, esto representa una dirección, jerarquía o lo que corresponda, entre los vértices. Por otra parte, **acíclico** significa que no se forman ciclos en la misma dirección, es decir que siguiendo las flechas de un vértice al siguiente nunca podemos volver al vértice inicial.

En el caso de los **árboles filogenéticos con raíz** (la raíz representa el origen de todas las especies que participan del árbol), los vértices representan especies o taxa, vivientes o ancestrales y las flechas representan las especies derivadas de una ancestral, obviamente apuntando en la dirección desde la ancestral a la descendiente. En el caso de un pedigrí, como ya vimos, los vértices son individuos y las flechas apuntan de un progenitor a su progenie. La similitud es obvia también a este nivel. En el caso de los árboles filogenéticos los largos de las aristas (ramas) suelen representar el tiempo evolutivo (por ejemplo, como función del número de sustituciones nucleotídicas) y, por lo tanto, ser variables. En el caso de los pedigrís, las aristas representan el pasaje de información de un padre a su hijo, que es siempre la mitad de su genoma, por lo tanto, la proporción $\frac{1}{2}$ es igual para todas las aristas. En el caso de un árbol filogenético dicotómico (los más usuales, digamos), cada nodo (vértice) interno genera dos descendientes y es descendiente de otros vértices (excepto la raíz), mientras que en el caso de un pedigrí cada nodo es descendiente de como máximo 2 vértices (sus padres, los dos si el pedigrí está completo, algo imposible en la práctica ya que la recursión sería infinita) pero puede emitir cualquier número de flechas (apuntando a ese número de hijos).

Esta similaridad entre pedigrís y árboles filogenéticos fue la que llevó a M. Lynch (1991) a sugerir que la misma aproximación de **modelos mixtos lineales** que usaban los genetistas cuantitativos en el área del mejoramiento genético animal se podía usar en el contexto filogenético. Ya que en el contexto del mejoramiento genético se conocían como modelos mixtos, la aproximación filogenética pasó a ser conocida como **PMM** (por Phylogenetic Mixed Model). A partir de esto, usando un enfoque por máxima verosimilitud y las ecuaciones de **BLUP/BLUE**, era posible estimar la varianza que correspondía a la filogenia (el «equivalente» del parentesco) y la que correspondía al ambiente, es decir todo lo no explicado por la filogenia.

Una diferencia práctica importante entre filogenias y pedigrís es que estos últimos suelen incluir miles y aún millones de individuos y datos y que además los nodos internos son conocidos y aún suelen tener datos. En el caso de las filogenias casi nada de esto es cierto: no se conocen los nodos internos (por lo que obviamente no tienen datos), tampoco la topología y los largos de ramas, que deben estimarse y las especies suelen ser del orden de las decenas o centenas como máximo. Teniendo todo esto en cuenta, Naya et al. (2006) propusieron el enfoque bayesiano para resolver estos problemas, lo que dio origen al **BPMM** (Bayesian PMM), que luego Hadfield (2010) implementó de forma eficiente.

A pesar de las similitudes entre pedigrís y filogenias, que nos permitirían estimar una **heredabilidad filogenética** de la característica, la varianza explicada por la filogenia no es de origen exclusivamente genético aditivo y ni siquiera genético ya que las especies derivadas suelen compartir ambiente con sus correspondientes especies ancestrales. Esto «oscurece» la interpretación de esa **heredabilidad filogenética**, pero no deja de ser un método extremadamente interesante para el análisis de caracteres cuantitativos en la evolución.

10.6 Heredabilidad en la era genómica

El advenimiento de la genómica ha tenido un profundo impacto en las estimaciones de heredabilidad. Por un lado, en el caso de las poblaciones silvestres ha permitido obviar las dificultades de tener que realizar experimentos específicos con diseños controlados para poder llegar a estimados de la misma. La posibilidad de genotipar o aún de secuenciar individuos al azar en la población permite obtener la matriz de «Parentesco genómico», lo que unido a «Métodos más avanzados de estimación» nos permite a partir de los fenotipos de los individuos seleccionados obtener una estimación razonable. Claramente, eso evita un trabajo de campo importante en el caso de animales difíciles de manejar, por ejemplo, con la identificación de las parejas reproductivas y sus respectivas crías, el desbalance en el número de descendientes, el número menor de animales a incluir, etcétera, todo lo cual reduce la calidad de las estimaciones. En la actualidad existen chips de genotipado para muchas especies, aunque la oferta es escasa para las especies no-domésticas o no de uso en experimentación. Sin embargo, como vimos antes, la **secuenciación genómica** es una alternativa que se puede considerar **libre-de-hipótesis** en el sentido de que no introduzco un sesgo al buscar solo la información conocida o aún seleccionada (como es en el caso de los **chips**). Dados los relativamente altos costos de secuenciación para muchas especies eucarióticas, existen varias alternativas como el **Low-Pass Whole Genome Sequencing** (LP-WGS) donde se reduce en forma aleatoria la cobertura del genoma, o el **Genotype By Sequencing** (GBS) donde se secuencia específicamente el producto de cortes por **enzimas de restricción**.

Por otra parte, la llegada de la genómica ha permitido definir y medir **nuevos fenotipos moleculares**. Por ejemplo, tanto con **chips de expresión** como a través de la secuenciación de transcritos (**RNAseq**) es posible determinar el nivel de expresión (transcripción) de diferentes genes en diferentes condiciones; es lo que se conoce como **transcriptómica**. En muchos casos, en genes codificantes, las diferencias entre alelos que producen diferencias en el fenotipo se deben al nivel de expresión de las proteínas que estos producen. Si asumimos que el **nivel de expresión de los transcritos** (que es ARNm) se correlaciona muy fuertemente con el nivel de proteína (es decir, después de la traducción), entonces la transcriptómica sería un **proxy** (una variable que se aproxima a la que desearíamos medir) al nivel de expresión de las proteínas, algo mucho más fácil y barato de medir que estas últimas y que puede hacerse en forma masiva (todos los genes a la vez). Como en muchos casos se conocen genes que están directamente relacionados con diferencias en el fenotipo, podemos estar interesados en conocer la heredabilidad de la expresión de los mismos. Es importante destacar que, si bien podemos conocer todo el genoma de cada individuo, eso no implica que podamos conocer o modelar razonablemente (al menos hoy) la regulación de la expresión génica, por lo que nuestros nuevos fenotipos moleculares no se diferencian tanto de los clásicos fenotipos a los que estamos acostumbrados. Como los loci de expresión cuantitativa se conocen en general como **QTL** (por sus siglas en inglés), las variantes genómicas que producen una alteración

en el nivel de expresión de los transcritos se conocen como **eQTL** (expression Quantitative Trait Loci).

En otro sentido, técnicas estándar para medir fenotipos clásicos pueden transformarse en fuente inagotable de **nuevos fenotipos** cuando exploramos los **datos brutos** generados por el equipamiento y los combinamos con fenotipos clásicos usando herramientas de aprendizaje automático. Por ejemplo, para la determinación del porcentaje de proteína y grasa en leche se utilizan en forma rutinaria espectrómetros de infrarrojos medios (**MIRS** o **FTIR**, por Fourier Transform Infra Red). Estos equipos obtienen un espectro de infrarrojos en, por ejemplo, 935 puntos de longitud de onda diferente. Es decir, el resultado de «medir» cada muestra de leche es un conjunto de 935 mediciones a longitudes de onda diferentes, que luego, usando una calibración previa del aparato con estándares conocidos y una ecuación que usa la información de algunos de esos 935 puntos, genera el resultado (predice) el porcentaje de proteína y grasa de la muestra. Sin embargo, usando esa misma información de los 935 puntos, si tenemos miles de muestras medidas y los fenotipos correspondientes para otra variable de nuestro interés, podemos buscar un modelo de predicción de la misma usando la información del espectrómetro. Por ejemplo, actualmente la medición de emisiones de metano en bovinos suele ser compleja y costosa, por lo que se reduce a unos pocos animales en condiciones experimentales. Sin embargo, si de esos pocos animales también obtenemos al mismo tiempo los espectros infrarrojos, podemos intentar desarrollar una ecuación de predicción que nos permita predecir la emisión de metano a partir de las muestras de leche. Se trata de una predicción, es decir, no es una medición específica del analito (metano en este caso), pero en la medida en que la precisión de la ecuación de predicción sea lo suficientemente alta podría ser una excelente herramienta para la reducción de las emisiones. Obviamente, cualquier otro fenotipo de nuestro interés y que en principio posea alguna relación con el espectro infrarrojo de la leche podría estudiarse en lugar del metano.

Finalmente, el acceso a los datos genómicos nos permite obtener una mucho mejor medida del parentesco entre dos individuos, ya que no es más necesario depender del parentesco estadístico esperado (ver «Parentesco genómico») y esto nos permite mejorar las estimaciones de heredabilidad. Una excelente revisión sobre la heredabilidad en el contexto de la genómica se encuentra en Visscher, Hill y Wray (2008), así como abundantes referencias.

10.7 Métodos más avanzados de estimación

Hasta aquí hemos visto la estimación de la heredabilidad a través de estructuras de parentesco conocidas, tanto en el caso de estimarla como relación varianzas recurriendo a diferentes formas del análisis de varianza (ANOVA), como en el caso de regresiones entre progenitores y progenies. Por otro lado, una alternativa que no depende de estructuras de parentesco determinadas es a través de la heredabilidad lograda, pero como ya vimos antes dicho método requiere usualmente de largos períodos de tiempo y un buen número de individuos para obtener resultados estables y con cierta precisión.

En la práctica, la mayor parte de las razas bovinas y ovinas de interés comercial que poseen una estructura de criadores suficientemente grande y estable participan de un sistema de evaluaciones genéticas a partir de los registros fenotípicos que aportan todos los participantes, así como los registros genealógicos que llevan las asociaciones y que permiten el armado de la matriz de parentesco aditivo de los animales. A partir de estos registros, si se conocen las varianzas aditivas y fenotípicas (o las residuales) es posible estimar los valores de cría de

los distintos individuos a partir de las ecuaciones **BLUE/BLUP** de modelos mixtos lineales desarrolladas por Charles Roy Henderson y que veremos en detalle en otro capítulo. Sin embargo, desafortunadamente para nosotros, lo que queremos estimar en este capítulo es justamente la heredabilidad de la característica, que como vimos es igual a $\frac{V_A}{V_P}$, por lo que estamos de alguna forma en el mismo lugar del que partimos.

Afortunadamente, existen dos enfoques conceptualmente distintos pero estrechamente relacionados para obtener los estimados de varianzas aditivas y fenotípicas a partir de las ecuaciones **BLUE/BLUP**. Conceptualmente distintos en lo que hace a la escuela estadística por detrás: el más tradicional basado filosóficamente en la escuela «frecuentista» y su aproximación por máxima verosimilitud (**ML**), el más moderno (pero cuya escuela es más antigua) basado en el enfoque estadístico «bayesiano» y su aproximación práctica basada en métodos **Monte Carlo.** Estrechamente vinculados en el hecho de que la estimación es básicamente **iterativa**, en el enfoque por **ML** avanzando sucesivamente hacia un **óptimo** (el máximo de la función de verosimilitud), en el **bayesiano** muestreando desde la distribución *a posteriori* de los parámetros de interés (en nuestro caso las varianzas aditiva y residual).

En el método usado para la estimación por máxima verosimilitud (usualmente **REML**, máxima verosimilitud restringida), una vez que el algoritmo alcanza el criterio de convergencia, además de un estimador puntual para cada una de las varianzas es posible obtener un estimador de la varianza de los estimadores (o sea, un estimador de la varianza de los estimadores de varianza), por lo que bajo ciertos supuestos (entre ellos normalidad asintótica) es posible calcular un intervalo de confianza.

En el enfoque **bayesiano**, la alternativa más sencilla y que funciona muy bien para una gran cantidad de características es el **muestreo de Gibbs** (*Gibbs sampling*), donde asumiendo determinadas distribuciones *a priori* como conjugadas de la distribución de los datos es posible obtener distribuciones *a posteriori* de la misma familia y, por lo tanto, muestrear de las mismas en forma eficiente. En cada ciclo del muestreador (iteraciones) los parámetros se van «muestreando» en forma sucesiva (algunos se pueden muestrear en bloque) condicional a los resultados obtenidos previamente para los otros parámetros. El resultado de este proceso, a partir de que se alcanza la convergencia, es una cadena de Markov muestreando de la distribución *a posteriori* de cada parámetro. Esto nos permite realizar inferencia directamente sobre las muestras obtenidas, pudiendo extraer estadísticos resumen (media, mediana, cuantiles, desvío estándar, etcétera.) pero a su vez analizar la distribución gráficamente para comprender mejor el proceso de muestreo y el alcance de la información brindada por los datos respecto a la información brindada *a priori*. Un excelente material para entender el **enfoque bayesiano** en los modelos mixtos lineales en genética animal se encuentra en el libro de Sorensen y Gianola (2002), mientras que un excelente libro general para entender los modelos lineales en genética cuantitativa es Mrode (2014).

10.8 Repetibilidad

Ciertas características fenotípicas se repiten en cada individuo, así sea en el tiempo o en el espacio. Por ejemplo, en animales podemos pesar el vellón de cada oveja todos los años, en vacas lecheras la producción de leche en cada lactancia, pero también podemos contar el número de setas en los segmentos abdominales de moscas de la especie *Drosophila* (Falconer1983) y, por lo tanto, se trataría de una repetición en el espacio en lugar del tiempo. En plantas que repiten anualmente su producción, cualquier característica que se puede medir

en los distintos ciclos claramente entra dentro de esta categoría, como el número de frutos en árboles, el peso total de los frutos y cualquier otra que siga este patrón. Es decir, existen características que tendemos a considerar como la misma, expresada varias veces en la vida del individuo.

Así como consideramos la heredabilidad como el parámetro genético central para entender el proceso de transferencia de información de padres a hijos, ahora nos enfocaremos en definir y entender un parámetro genético que hace a la **repetibilidad** de la característica de interés. Consideremos la repetición en el tiempo de una característica ya que nos resulta mucho más intuitiva. Veamos entonces qué componentes de la varianza fenotípica se corresponden con efectos que se repiten «inmutables» durante la vida del animal. Claramente, más allá de los posibles cambios epigenéticos (que despreciaremos), el genotipo de un animal no cambiará durante su vida y por lo tanto ni los efectos aditivos de los genes, ni los efectos de dominancia o epistáticos (al menos asumiendo que no hay cambios importantes en las medias poblacionales en el período en consideración). Esto nos deja con el ambiente como posibilidad a considerar en las causas de las variaciones entre medidas.

Sin duda, en la mayor parte de las características fenotípicas de interés el ambiente juega un rol importante en su desarrollo. Por ejemplo, un árbol frutal producirá diferente número de frutos o peso total de los mismos en un año «bueno» que en un año «malo». Lo mismo, una oveja producirá diferente cantidad de lana en diferentes años, producto de las variaciones entre años en factores como la alimentación y el bienestar, entre otros. Por lo tanto, nos resulta claro que las variaciones ambientales afectarán el fenotipo entre medidas en el mismo animal, pero la pregunta que nos surge es si toda la **influencia ambiental** que afecta a los animales varía de medición en medición. Antes de seguir, debemos de tener en cuenta que **en el contexto de nuestro modelo genético básico le llamamos ambiental a todo lo que no es genético**.

Volvamos a la discusión de si toda la influencia del ambiente hará variar el fenotipo de año en año. Por ejemplo, un árbol que crece en malas condiciones difícilmente llegará a desarrollar todo su potencial genético y la producción en el mismo se verá afectada durante toda la vida del árbol. Lo mismo para un cordero que crece en malas condiciones o aún para una vaquillona que sufre la pérdida definitiva de un cuarto de la ubre al comenzar su primera lactancia. En todos los casos se trata de factores ambientales de acuerdo con nuestra consideración de *ambiental* como lo *no-genético*, pero claramente estos efectos afectarán la producción de toda la vida del individuo. Sin embargo, como vimos antes, algunos factores cambian de año en año el fenotipo del animal en medidas repetidas, por lo que podemos descomponer los efectos ambientales en dos tipos: *a)* **efectos ambientales permanentes** y *b)* **efectos ambientales temporales**.

> Llamamos **efectos ambientales permanentes** a aquellos efectos en los que las diferencias no son atribuibles a la genética, pero que van a afectar todos los registros fenotípicos de una característica que se repite durante la vida del individuo.
> Por otro lado, llamamos **efectos ambientales temporales** a aquellos efectos de origen no-genético que afectan en forma particular cada una de las mediciones repetidas de la misma característica.

La repetibilidad como cociente de varianzas

De acuerdo con estas definiciones podemos partir los efectos ambientales en dos componentes, uno que se manifestará de la misma manera durante toda la vida del animal y otro que variará de observación en observación, por lo que, considerando que no hay covariación entre ambos efectos, tenemos

$$E = E_{perm} + E_{temp} \ \therefore \tag{10.74}$$

$$V_E = V_{E_{perm}} + V_{E_{temp}} \tag{10.75}$$

Ahora, si colocamos esta división de la varianza de los efectos ambientales en nuestro modelo genético básico, es decir si sustituimos $\mathrm{V_E}$ por $\mathrm{V_{E_{perm}}} + \mathrm{V_{E_{temp}}}$, tenemos que

$$V_P = V_G + V_E = V_G + V_{E_{perm}} + V_{E_{temp}} \tag{10.76}$$

Por lo tanto, a la fracción de la varianza fenotípica correspondiente a los efectos que se repiten entre diferentes mediciones en la vida del individuo la llamamos ρ y será

$$\rho = \frac{\mathrm{V_G} + \mathrm{V_{E_{perm}}}}{\mathrm{V_P}} = \frac{\mathrm{V_A} + \mathrm{V_D} + \mathrm{V_I} + \mathrm{V_{E_{perm}}}}{\mathrm{V_P}} \tag{10.77}$$

En diferentes textos es usual usar R para la repetibilidad, o aún r (Falconer (1983), pero ambos símbolos se prestan a confusión, el primero con la respuesta a la selección (o progreso genético), el segundo con cualquier coeficiente de correlación (aunque, como veremos más adelante, la repetibilidad también se puede definir como correlación), por lo que nos hemos decidido por usar ρ (la letra griega «rho» o «ro») para simbolizarla.

La repetibilidad como coeficiente de correlación intraclase

Como vimos previamente en el caso de la heredabilidad, a veces es posible definir los parámetros genéticos de diferentes formas. En el caso de la heredabilidad vimos que además de como cociente de varianzas podíamos definirlo como el coeficiente de regresión lineal, tanto del valor de cría en el fenotipo individual, como de funciones sencillas de la regresión fenotípica entre progenitores y progenie. En el caso de la repetibilidad es posiblemente aún más intuitiva otra forma matemática de definirla.

Si una característica se repite exactamente con el mismo valor una segunda vez, entonces la correlación entre pares de medidas en el individuo, es decir varios individuos medidos dos veces, será de 1. Si los valores de la segunda medida son completamente al azar respecto a la primera, entonces la correlación entre ambas medidas será de cero. Claramente, la correlación entre medidas nos indica de alguna forma cuán repetible es la característica. Como se trata de la **correlación entre medidas de la misma clase**, al coeficiente de correlación de Pearson de este tipo se le llama **coeficiente de correlación intraclase** y lo simbolizaremos con t.

Si recordamos que el coeficiente de correlación de Pearson es igual al cociente entre la covarianza y el producto de los desvíos entre las dos variables aleatorias, entonces para el caso de dos medidas fenotípicas en el mismo individuo tenemos que

$$t = \frac{Cov_{P_1P_2}}{\sqrt{V_{P_1}V_{P_2}}} = \tag{10.78}$$

$$\frac{Cov(A_1 + D_1 + I_1 + E_{perm_1} + E_{temp_1}, A_2 + D_2 + I_2 + E_{perm_2} + E_{temp_2})}{\sqrt{V_{P_1}V_{P_2}}} \tag{10.79}$$

Pero por definición, **como se trata de dos mediciones en el mismo individuo**, los efectos genéticos serán los mismos en ambas mediciones, así como los del ambiente permanente. Por lo tanto

$$A_1 = A_2 = A;\ D_1 = D_2 = D;\ I_1 = I_2 = I;\ E_{perm_1} = E_{perm_2} = E_{perm} \tag{10.80}$$

Ahora, utilizando los resultados de (10.80) en (10.79) tenemos

$$t = \frac{Cov(A + D + I + E_{perm}, A + D + I + E_{perm})}{\sqrt{V_{P_1}V_{P_2}}} \tag{10.81}$$

Como ya vimos previamente, por construcción no tenemos covarianza entre efectos aditivos y de dominancia o epistasis, ni entre estos últimos y por lo tanto, cuando no existe covarianza genético-ambiental la ecuación (10.81) se transforma en

$$t = \frac{V_A + V_D + V_I + V_{E_{perm}}}{\sqrt{V_{P_1}V_{P_2}}} \tag{10.82}$$

Claramente se trata de otra forma de definir la repetibilidad de una característica, esta vez como la correlación entre medidas repetidas del mismo individuo. Más aún, si la varianza ambiental temporal es igual en las distintas mediciones, es decir, en particular $V_{E_{temp_1}} = V_{E_{temp_2}} = V_{E_{temp}}$, entonces

$$V_{P_1} = V_G + V_{E_{perm}} + V_{E_{temp_1}} = V_G + V_{E_{perm}} + V_{E_{temp_2}} = V_G + V_{E_{perm}} + V_{E_{temp_1}} = V_P \tag{10.83}$$

y por lo tanto $\sqrt{V_{P_1}V_{P_2}} = \sqrt{V_PV_P} = \sqrt{V_P^2} = V_P$ y sustituyendo en el denominador de (10.82) tenemos que

$$t = \frac{V_A + V_D + V_I + V_{E_{perm}}}{\sqrt{V_{P_1}V_{P_2}}} = \frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P} = \rho \tag{10.84}$$

Por lo tanto, en el caso de que las varianzas fenotípicas sean iguales en las distintas mediciones consideradas, las dos definiciones de repetibilidad, como cociente de varianzas y como coeficiente de correlación intraclase coinciden.

La relación entre la heredabilidad y la repetibilidad

A esta altura es lógico preguntarnos cuál es la importancia práctica de la repetibilidad. Para eso, en primer lugar veamos cuál es la relación entre la repetibilidad, la heredabilidad en sentido amplio y la heredabilidad en sentido estricto.

Si escribimos en forma explícita los distintos componentes del numerador de la repetibilidad, recordando $V_A + V_D + V_I = V_G$ tenemos que

$$\rho = \frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P} = \frac{V_A + V_D + V_I}{V_P} + \frac{V_{E_{perm}}}{V_P} = H^2 + \frac{V_{E_{perm}}}{V_P} \qquad (10.85)$$

Como las varianzas son positivas, entonces $\frac{V_{E_{perm}}}{V_P} \geqslant 0$ y por lo tanto, de acuerdo con la ecuación (10.85) la repetibilidad es igual a la heredabilidad en sentido amplio más un número positivo (estrictamente, entre 0 y 1, ya que $V_{E_{perm}} \leqslant V_P$). Por lo tanto $\rho \geqslant H^2$.

Con igual razonamiento, recordando que $h^2 = \frac{V_A}{V_P}$, tenemos que

$$\rho = \frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P} = \frac{V_A}{V_P} + \frac{V_D + V_I + V_{E_{perm}}}{V_P} = h^2 + \frac{V_D + V_I + V_{E_{perm}}}{V_P} \qquad (10.86)$$

Nuevamente, como las varianzas son positivas, tenemos ahora que $\rho \geqslant h^2$. Más aún, como $\rho = H^2 + \frac{V_{E_{perm}}}{V_P} = h^2 + \frac{V_D+V_I+V_{E_{perm}}}{V_P}$ y $\frac{V_D+V_I+V_{E_{perm}}}{V_P} \geqslant \frac{V_{E_{perm}}}{V_P}$, entonces H^2 debe ser mayor a h^2, cosa que no hace más que verificar lo que ya sabíamos. Esto nos permite establecer firmemente la siguiente la relación entre los tres parámetros genéticos

$$\rho \geqslant H^2 \geqslant h^2 \qquad (10.87)$$

Estas relaciones son matemáticamente correctas si se cumplen los supuestos de los que partimos. Sin embargo, las estimaciones de repetibilidad pueden no establecer un límite superior a la heredabilidad si se da alguna de las siguientes condiciones (Dohm 2002):

- a) los rasgos medidos no son idénticos desde el punto de vista genético;
- b) los efectos ambientales comunes se oponen a los efectos genéticos directos;
- c) los entornos temporales de cada rasgo están correlacionados negativamente;
- d) existe una interacción significativa entre el genotipo y el ambiente;
- e) los rasgos están influidos por los efectos maternos.

En ese caso, además de no necesariamente cumplirse la relación $\rho \geqslant H^2 \geqslant h^2$, los estimados de los parámetros pueden ser incorrectos bajo los supuestos que los realizamos. Un ejemplo hipotético de *c)* podría ser una característica en la que el éxito en la expresión en el ciclo n implicase un gasto energético tal que el ciclo $n + 1$ se ve comprometido seriamente. Por ejemplo, en algunas especies de tiburones el costo energético de la reproducción es suficientemente alto como para que el ciclo natural sea de dos años pese a que la gestación es del orden del año. Si obligáramos a estas especies a reproducirse artificialmente todos los años y efectuásemos mediciones sobre las crías, posiblemente algunas características se correlacionarían negativamente entre mediciones ya que el exceso de habilidad en una gestación sería un detrimento serio para la siguiente.

Cada parámetro describe una proporción distinta de la variabilidad fenotípica explicada por diferentes efectos genéticos y ambientales y, por lo tanto, tiene sentido aplicarlos en

distintos contextos. Como vimos previamente, la heredabilidad en sentido estricto nos permite realizar una predicción del valor de cría de cada individuo a partir de su fenotipo y por lo tanto tiene más sentido su utilización en el contexto del mejoramiento genético en aquellas especies y características donde los efectos aditivos dominan la expresión fenotípica. En general, en mejoramiento genético animal la heredabilidad en sentido estricto es el parámetro genético central para la toma de decisiones (tanto de estrategia como a la hora de seleccionar individuos). En cambio, en mejoramiento genético vegetal, dependiendo de la especie puede tener mucho más sentido el uso de la heredabilidad en sentido amplio. Más aún, en especies que no tienen reproducción clonal y que repiten su producción a lo largo de la vida **es muchas veces más fácil obtener un estimado de la repetibilidad que de la heredabilidad**. Como vimos antes, para las estimas de heredabilidad es necesario recurrir a diseños experimentales con determinadas estructuras de parentesco conocidas o al menos conocer las relaciones de parentesco entre individuos y utilizar métodos iterativos y las ecuaciones **BLUE/BLUP**. En cambio, para estimar la repetibilidad alcanza con un par de mediciones por individuo y a través de estos datos calcular el coeficiente de correlación intraclase t, que como hemos visto es igual a la repetibilidad.

Esto nos deja con un método de obtener una cota superior para la heredabilidad, tanto en sentido amplio como estricto ya que ambas deben ser menores a ρ (ecuación (10.87)). Si el valor de repetibilidad obtenido es alto, la heredabilidad puede ser alta, media o baja (es decir, cualquier valor entre 0 y ρ), lo que no clarifica demasiado, pero nos da esperanza y ciertas pistas si hacemos suposiciones adicionales. Por ejemplo, si por otros estudios sabemos que se trata de una característica cuya arquitectura genética es mayormente aditiva (es decir, los efectos de dominancia y epistasis son menores) y que en general los efectos del ambiente permanente son relativamente menores (por un control estricto del ambiente durante la fase de crecimiento, por ejemplo), entonces $h^2 \approx \rho$ y como estamos en el caso de alta repetibilidad, entonces la heredabilidad también será alta. Por el otro lado, si la repetibilidad es baja, entonces necesariamente la heredabilidad será baja ya que la esta última no puede superar a la primera.

La reducción en la varianza fenotípica con varias medidas

Un concepto general que está por detrás del uso de la repetibilidad como herramienta de predicción (algo que veremos pronto) es la reducción en la varianza fenotípica producto de promediar varias observaciones de la misma característica en el mismo individuo. La idea subyacente es muy sencilla: los componentes permanentes (genéticos y ambientales) son los mismos en todas las medidas repetidas, mientras que los efectos ambientales temporales son desvíos aleatorios respecto a una media de cero; cuantas más medidas promediemos en el individuo, más cerca de cero será la media de los efectos ambientales temporales y menor será la varianza de la media de los efectos temporales. Veamos cómo funciona esta idea desde el punto de vista matemático.

En general, para aquellos caracteres que se repiten a lo largo de la vida del animal, ya sea a lo largo del tiempo (lo más usual en el contexto del mejoramiento genético animal) o aún de la anatomía del mismo, vimos que podemos descomponer nuestro modelo genético básico de tal forma de separar ahora los efectos ambientales en aquellos que son permanentes en la vida del animal, de aquellos que son temporarios

$$P = A + D + I + E_{perm} + E_{temp} \tag{10.88}$$

En términos de varianzas, asumiendo como de costumbre que las covarianzas entre componentes son iguales a cero, tenemos entonces

$$V_P = V_A + V_D + V_I + V_{E_{perm}} + V_{E_{temp}} \tag{10.89}$$

Veamos ahora qué ocurre con la varianza fenotípica del promedio de n medidas en cada animal. Para aquellos componentes que son idénticos durante la vida del animal, por definición los **permanentes**, la varianza del promedio de **n** medidas en cada animal es simplemente el correspondiente componente de la varianza (ya que no hay una reducción de la varianza por tratarse para cada animal de un valor fijo, no aleatorio). En cambio, para los efectos temporales que si son aleatorios dentro de animal (cada año, por ejemplo, el clima afecta en forma diferente la producción de cada animal), la varianza del promedio es ahora $\frac{VE_{temp}}{n}$. Esto nos lleva a que

$$V_{\bar{P}_n} = V_A + V_D + V_I + V_{E_{perm}} + \frac{V_{E_{temp}}}{n} \tag{10.90}$$

Si comparamos la varianza fenotípica de n medidas por individuo con la varianza fenotípica de la población, tenemos

$$\frac{V_{\bar{P}_n}}{V_P} = \frac{V_A + V_D + V_I + V_{E_{perm}} + \frac{V_{E_{temp}}}{n}}{V_P} \tag{10.91}$$

$$\frac{V_{\bar{P}_n}}{V_P} = \frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P} + \frac{\frac{V_{E_{temp}}}{n}}{V_P} \tag{10.92}$$

$$\frac{V_{\bar{P}_n}}{V_P} = \frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P} + \frac{1}{n}\frac{V_{E_{temp}}}{VP} \tag{10.93}$$

Pero

$$\frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P} = \rho \tag{10.94}$$

y

$$\frac{V_{E_{temp}}}{V_P} = \frac{V_P - (V_A + V_D + V_I + V_{E_{perm}})}{V_P} \tag{10.95}$$

$$\frac{V_{E_{temp}}}{V_P} = \frac{V_P}{V_P} - \frac{VA + VD + VI + V_{E_{perm}}}{V_P} = 1 - \rho \tag{10.96}$$

Sustituyendo en (10.93) el primer término de la derecha por el resultado de (10.94) y colocando en el segundo el resultado de (10.96), tenemos

$$\frac{V_{\bar{P}_n}}{V_P} = \rho + \frac{1}{n}(1-\rho) = \frac{n\rho + 1 - \rho}{n} = \frac{1 + (n-1)\rho}{n} \tag{10.97}$$

Ahora, si en lugar de verlo como la reducción de la varianza producto de las n medidas individuales, lo miramos de la forma inversa, es decir como la ganancia en precisión, tenemos una relación que veremos en gran parte de las ecuaciones para la estimación de los valores de cría y sus correspondientes precisiones en aquellos casos en que tengamos alguna forma de medidas repetidas (o **correlación intraclase** más generalmente).

$$\frac{V_P}{V_{\bar{P}_n}} = \frac{n}{1 + (n-1)\rho} \tag{10.98}$$

Como en los parámetros genéticos que hemos visto el numerador representa la varianza fenotípica, una reducción de la misma producto de promediar diferentes medidas repetidas tiene como efecto aumentar la precisión de estos a la hora de utilizarlos como predictores (del valor de cría o de la producción más probable). El efecto de la reducción de la varianza producto de promediar diferente número de observaciones se puede apreciar en la figura 10.11. Claramente, cuantas más medidas promediamos mayor será la reducción en la varianza fenotípica del promedio, lo que nos llevaría a pensar que es una buena idea esperar a tener muchas mediciones. Sin embargo, cuando las repeticiones son temporales y no espaciales, cada repetición implica esperar un nuevo ciclo entero, por lo general anual. ¿Qué sentido puede tener esperar a tener 6, 7 u 8 mediciones de peso de vellón para elegir un animal como reproductor si la vida misma del animal es de este orden de duración? De hecho, la reducción de la varianza es cada vez menor y la asíntota de la misma será igual a

$$\lim_{n\to\infty} \frac{V_{\bar{P}_n}}{V_P} = \lim_{n\to\infty} \rho + \frac{1}{n}(1-\rho) = \rho \qquad (10.99)$$

Más aún, para repetibilidades relativamente altas vemos la curva se vuelve asintótica relativamente rápido. Dicho de otra manera, cuando las repetibilidades son altas ganamos muy poco al añadir nuevas mediciones. Por ejemplo, con una repetibilidad de $\rho = 0{,}75$ tres observaciones ya están suficientemente cerca de la asíntota como para tener que pensar si justifica obtener nuevas mediciones. En la práctica, la consideración debe incluir los costos y las complejidades asociadas a esperar nuevas mediciones y una o dos observaciones suelen ser suficientes, como verás en próximos cursos. Cuando las repetibilidades son más bajas, por ejemplo $\rho < 0{,}30$, claramente el promediado de varias observaciones representa una ganancia fundamental en precisión, por lo que se suele esperar por dos o tres mediciones, raramente cuatro.

PARA RECORDAR

- Al realizar una estimación de la heredabilidad debemos tener en cuenta que los métodos y condiciones de estimación de la misma deben poder representar el uso pretendido. El denominador de la heredabilidad es la varianza fenotípica, uno de cuyos componentes es la varianza ambiental: en la medida en que la varianza ambiental es menor, menor será el denominador y por lo tanto mayor el cociente por lo que estimaciones en ambientes más controlados tenderán a estimar mayores heredabilidades. Teniendo esto en cuenta, no parecería recomendable utilizar estimaciones realizadas en laboratorios para diseñar un programa de mejora a campo.
- Nuestro estimador de la heredabilidad a partir del diseño de hermanos enteros será $\hat{h}^2 = h^2 + \frac{\frac{1}{2}\,\sigma^2_D + 2\,\sigma^2_{EC}}{V_P}$ lo que sesga el estimador hacia arriba dado que el término $\frac{\frac{1}{2}\,\sigma^2_D + 2\,\sigma^2_{EC}}{V_P}$ es en general mayor a cero.
- Las características que tienen una base genética compleja (donde lo aditivo no es tan importante), un ambiente extremadamente variable o interacciones entre genotipo y ambiente llevarán a menores estimaciones de la heredabilidad.
- Es importante tener en cuenta que una **baja heredabilidad no es reflejo de una baja varianza aditiva**, la baja heredabilidad de una característica está muchas veces asociada a la forma en que se registran los datos fenotípicos de la misma o a la definición imprecisa de lo que queremos medir.

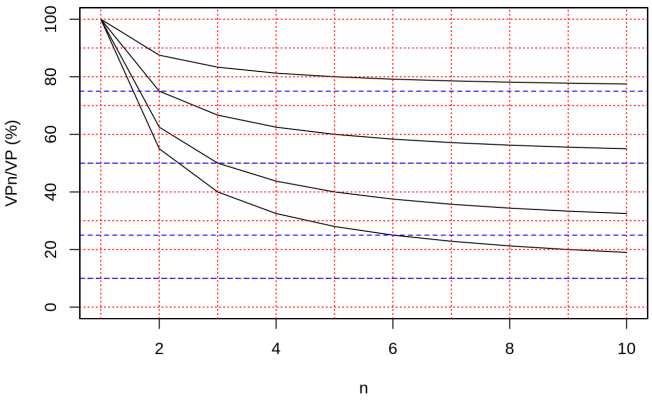


Figura 10.11. Reducción en la varianza fenotípica producto de promediar n observaciones. En las abscisas aparece en número de observaciones fenotípica promediadas mientras que en las ordenadas aparece la relación $\frac{V_{Pn}}{V_P} = \frac{1+r(n-1)}{n}$ (como porcentaje). Se ilustra el comportamiento de la relación para distintos valores de repetibilidad. Mientras que a valores altos de la repetibilidad tenemos poca reducción de la varianza y por lo tanto poca ventaja de incorporar más observaciones, a valores bajos de la **repetibilidad** (por ejemplo, $r = 0{,}10$) la reducción es sustancial hasta cerca de **4** medidas. Las líneas azules a trazos representan las asíntotas correspondientes a las distintas repetibilidades.

- Las estimaciones de heredabilidad a campo suelen incluir pocos individuos y un trabajo de observación considerable debido a los requisitos que sabemos conlleva esta estimación.
- La aparición de los **métodos de identificación de parentesco a partir del ADN**, en particular los métodos de genotipado y secuenciado han permitido analizar un grupo de individuos escogidos al azar, utilizando las mismas herramientas de estimación a través de métodos por máxima verosimilitud o métodos bayesianos.
- Los **árboles filogenéticos** constituyen estructuras donde un clado ancestral se deriva en dos clados diferentes, que pueden continuar evolucionando sin dar lugar a nuevos clados o nuevamente particionarse y así sucesivamente.
- Alternativas más avanzadas en la determinación de las varianzas aditivas y fenotípicas pueden ser la aproximación por máxima verosimilitud (**ML**; a través del método **REML** usualmente) perteneciente a la escuela frecuentista o el método **Monte Carlo** (**muestreo de Gibbs**) que sigue el enfoque estadístico bayesiano.

- Recordando que en el contexto de nuestro modelo genético básico le llamamos **ambiental a todo lo que no es genético** podemos descomponer los efectos ambientales en dos tipos:

a) **efectos ambientales permanentes** serán aquellos en los que las diferencias no son atribuibles a la genética, pero que van a afectar todos los registros fenotípicos de una característica que se repite durante la vida del individuo.

b) **efectos ambientales temporales** serán aquellos efectos de origen no-genético que afectan en forma particular cada una de las mediciones repetidas de la misma característica.

- Al igual que con la heredabilidad, podemos expresar la repetibilidad (que representaremos con la letra ρ) como un **coeficiente de varianzas** tal que $\rho = \frac{V_G + V_{E_{perm}}}{V_P} = \frac{V_A + V_D + V_I + V_{E_{perm}}}{V_P}$

10.9 La repetibilidad como herramienta en la predicción

En muchos casos, para producciones que se repiten a lo largo de la vida del individuo resulta interesante conocer de antemano que se puede esperar en un nuevo ciclo productivo del mismo. Por ejemplo, en el caso del peso de vellón en lanares, o de la producción de leche en razas lecheras, la producción de frutos en árboles, etcétera. Como ya vimos antes, en estos casos es interesante plantearse una regresión lineal como predictor de los nuevos valores, basados en los valores anteriores obtenidos para cada animal.

En el caso de que poseamos o usemos un solo registro previo, entonces resulta claro que si la varianza fenotípica del previo y del que vamos a predecir es la misma, entonces el coeficiente de la regresión será igual a

$$b_{P_2 P_1} = \frac{Cov_{P_1 P_2}}{V_{P_1}} = \frac{Cov_{P_1 P_2}}{\sqrt{V_{P_1} V_{P_2}}} = \rho \tag{10.100}$$

Por lo tanto, considerando que la regresión la hicimos con valores que son desvíos de la media poblacional, la predicción de un nuevo registro (x_2) a partir del registro anterior (x_1), a la que llamamos **producción más probable** y denotamos como PMP sería

$$\mathrm{PMP}(\mathrm{x}_2) = \bar{\mathrm{x}} + (\mathrm{x}_1 - \bar{\mathrm{x}})\rho \tag{10.101}$$

Es decir, se trata de un simple predictor lineal de la próxima producción basado en el desvío del presente registro respecto a la media poblacional, $(\mathrm{x}_1 - \bar{\mathrm{x}})$. En la práctica, dadas las diferencias sistemáticas que suelen existir entre sexos, por ejemplo, se suelen corregir los registros previamente por esas diferencias sistemáticas. Además, como es posible que existan diferencias de tratamientos entre años o grupos de manejo, entonces es posible realizar las comparaciones respecto a la media del **grupo contemporáneo** en lugar de la media general.

De acuerdo con lo que vimos antes y al concepto mismo de repetibilidad, cuando disponemos de más registros previos de la misma característica en la vida de un individuo más

nos acercamos al verdadero valor de los efectos permanentes y por lo tanto se incrementa la información acerca de las predicciones futuras. Esto nos lleva a considerar la alternativa de desarrollar la regresión no en el último registro (o alguno en particular), sino en el promedio de los mismos. Supongamos que tenemos entonces n registros de la misma característica en cada individuo y queremos desarrollar un predictor lineal con los mismos para el próximo registro (o sea, el registro $n+1$).

Por definición, el coeficiente de regresión para los registros (como desvíos de la media) será

$$b_{P_{n+1}P_{\bar{n}}} = \frac{Cov_{P_{\bar{n}}P_{n+1}}}{V_{P_{\bar{n}}}} \tag{10.102}$$

Pero como vimos antes, en la ecuación (10.98), por ejemplo, la varianza del promedio de n medidas fenotípicas en el mismo individuo es

$$\frac{V_P}{V_{\bar{P}_n}} = \frac{n}{1+(n-1)\rho} \therefore V_{\bar{P}_n} = \left[\frac{1+(n-1)\rho}{n}\right] V_P \tag{10.103}$$

A su vez, como la covarianza entre la media de n efectos temporarios y el próximo efecto temporario es cero, la covarianza del promedio de n medidas fenotípicas con una nueva medida es

$$Cov_{P_{\bar{n}}P_{n+1}} = Cov(A+D+I+E_{perm}+\bar{E}_{temp}, A+D+I+E_{perm}+E_{temp_{n+1}}) = V_A+V_D+V_I+V_{E_{perm}} = \rho\, V_P \tag{10.104}$$

La última línea de la ecuación anterior surge de que
$\rho = \frac{V_A+V_D+V_I+V_{E_{perm}}}{V_P} \Leftrightarrow V_A+V_D+V_I+V_{E_{perm}} = \rho\, V_P$
Sustituyendo el resultado de la ecuación (10.104) en el numerador de la ecuación (10.102) y el resultado de la ecuación (10.112) por el denominador de (10.102), tenemos ahora que

$$b_{P_{n+1}P_{\bar{n}}} = \frac{\rho\, V_P}{\left[\frac{1+(n-1)\rho}{n}\right] V_P} = \frac{n\rho}{1+(n-1)\rho} \tag{10.105}$$

Por lo tanto, en el caso de utilizar el promedio de n medidas de la misma característica como predictor de la producción más probable en el próximo registro, el coeficiente regresión será el de la ecuación (10.105) y la producción más probable para el registro $n+1$ será

$$PMP(x_{n+1}) = \bar{x} + (x_{n+1}-\bar{x})\left[\frac{n\rho}{1+(n-1)\rho}\right] \tag{10.106}$$

De hecho, podemos escribir el factor que multiplica al desvío del próximo registro como

$$\frac{n\rho}{1+(n-1)\rho} = \rho\frac{n}{1+(n-1)\rho} = \rho\left[\frac{n}{1+(n-1)t}\right] \tag{10.107}$$

con t el **coeficiente de correlación intraclase.** En este caso particular $t = \rho$, pero vimos antes y lo veremos con más detalle luego que en otros casos $t = r_{A_1A_2}h^2$ (cuando discutimos la varianza del estimador de la heredabilidad por regresión).

Claramente, la ecuación (10.101) es un caso particular de la ecuación (10.106), cuando $n = 1$, es decir cuando la predicción está basada en un solo registro, ya que si $n = 1$ entonces

$$PMP(x_{n+1}) = \bar{x} + (x_{n+1} - \bar{x}) \left[\frac{n\rho}{1 + (n-1)\rho} \right] = \bar{x} + (x_2 - \bar{x}) \left[\frac{(1)\rho}{1 + (1-1)\rho} \right] = \tag{10.108}$$

$$\bar{x} + (x_2 - \bar{x}) \left[\frac{\rho}{1+0} \right] = \bar{x} + (x_2 - \bar{x})\rho = PMP(x_2) \tag{10.109}$$

Hasta ahora vimos cómo el incremento en el número de registros nos permite incrementar la precisión en la predicción de una nueva producción. Sin embargo, el mismo criterio se aplica a la predicción del valor de cría a partir de una regresión. Como vimos en la sección «La heredabilidad como regresión», el coeficiente de regresión que usábamos para predecir el valor de cría de un animal a partir de un registro fenotípico individual era la heredabilidad, es decir

$$\hat{A}_i = h^2(x_i - \bar{x}) \tag{10.110}$$

Esto venía de la ecuación (10.5), la que ahora, si consideramos que poseemos la información de n medidas en el individuo y usamos su promedio se transforma en

$$b_{A\bar{P}} = \frac{Cov(A,\bar{P})}{V_{\bar{P}}} = \frac{Cov(A,A + D + I + E_{perm} + \bar{E}_{temp})}{V_{\bar{P}}} = \frac{V_A}{V_{\bar{P}}} \tag{10.111}$$

Pero como vimos en la ecuación (10.112), $V_{\bar{P}_n} = \left[\frac{1+(n-1)\rho}{n} \right] V_P$, por lo que sustituyendo esto en el denominador de la ecuación (10.111) tenemos ahora

$$b_{A\bar{P}_n} = \frac{V_A}{V_{\bar{P}}} = \frac{V_A}{\left[\frac{1+(n-1)\rho}{n} \right] V_P} = \frac{V_A}{V_P} \left[\frac{n}{1 + (n-1)\rho} \right] = h^2 \left[\frac{n}{1 + (n-1)\rho} \right] \tag{10.112}$$

Por lo tanto, con este coeficiente de regresión, el predictor del valor de cría de un individuo a partir del promedio de n registros fenotípicos en el mismo será

$$\hat{A}_i = h^2 \left[\frac{n}{1 + (n-1)\rho} \right] (x_i - \bar{x}) \tag{10.113}$$

y la precisión de la estimación, r_{AC}, estará dada ahora por

$$r_{AC} = h \sqrt{\frac{n}{1 + (n-1)\rho}} \tag{10.114}$$

en lugar de h, que era cuando teníamos un solo registro. Obviamente, sustituyendo $n = 1$ en la ecuación (10.105), es decir un solo registro, obtenemos el mismo resultado $r_{AC} = h$

PARA RECORDAR

- El concepto de **repetibilidad** se desprende del interés de conocer de antemano (para producciones que se repiten a lo largo de la vida del individuo) qué se puede esperar en un nuevo ciclo productivo del mismo.
- La **predicción de un nuevo registro** (x_2) a partir del registro anterior (x_1), a la que llamamos **producción más probable** (**PMP**) sería $PMP(x_2) = \bar{x} + (x_1 - \bar{x})\rho$ que se trata de un predictor lineal de la próxima producción basado en el desvío del presente registro respecto a la media poblacional ($(x_1 - \bar{x})$) o al grupo contemporáneo.
- Cuando tenemos más de un registro, podemos calcular la **PMP** para el registro $n+1$: $PMP(x_{n+1}) = \bar{x} + (x_{n+1} - \bar{x}) \left[\frac{n}{1+(n-1)t}\right]$ siendo t el **coeficiente de correlación intraclase**.
- El predictor del **valor de cría** de un individuo a partir del promedio de n registros fenotípicos en el mismo será $\hat{A}_i = h^2 \left[\frac{n}{1+(n-1)\rho}\right] (x_i - \bar{x})$ y la precisión de la estimación, r_{AC}, estará dada ahora por $r_{AC} = h\sqrt{\frac{n}{1+(n-1)\rho}}$

10.10 Evolucionabilidad

Un tema importante en términos evolutivos es la capacidad de evolucionar de una característica y que por lo tanto ha adoptado el extraño término de **evolucionabilidad** en castellano (como traducción de *evolvability*). En principio, resulta tentador pensar en la heredabilidad de una característica como su potencial de evolucionar, particularmente a partir de la **ecuación del criador** $R = h^2 S$. En un artículo de gran profundidad conceptual, Houle (1992) plantea que la heredabilidad no es una buena medida de la capacidad de evolucionar de una característica. En lugar de proponer estandarizar la varianza aditiva entre la varianza fenotípica, propone en su lugar la estandarización a partir de la media de la característica usando el **coeficiente de variación aditivo**, es decir el desvío aditivo dividido la media de la característica $CV_A = \frac{\sigma_A}{\bar{x}}$.

De esta manera, de acuerdo con Houle (1992), características que presentan sistemáticamente baja heredabilidad, como las directamente asociadas al **fitness** (por ejemplo, las reproductivas), tendrían una gran **evolucionabilidad**, mayor a las morfológicas (que tienen usualmente heredabilidad alta) cuando se realiza la estandarización por la media. Sin embargo, utilizar un coeficiente de variación solo tiene sentido para características en las que la media tiene un sentido biológico intrínseco y además la respuesta a la selección natural o artificial es una función de $h\,\sigma_A$, por lo que no solo depende de la varianza aditiva sino también de la heredabilidad.

En resumen, si bien la heredabilidad no es una buena medida de la capacidad de evolucionar de una característica, la alternativa planteada por Houle (1992) también presenta problemas importantes y su campo de aplicación es restringido.

PARA RECORDAR

- La capacidad de evolucionar de una característica y ha tomado el término de **evolucionabilidad** y se ha planteado que la heredabilidad no es una buena medida de la misma. En su lugar se ha propuesto la estandarización a partir de la media de la característica usando el **coeficiente de variación aditivo**, es decir el desvío aditivo dividido la media de la característica $CV_A = \frac{\sigma_A}{\bar{x}}$.

10.11 Conclusión

En este capítulo presentamos un concepto de suma importancia para la teoría de genética cuantitativa y el mejoramiento genético: la *heredabilidad*. Como vimos en los capítulos anteriores no siempre resulta fácil estimar los componentes de la varianza fenotípica en la realidad, ya que algunos de ellos requerirían de condiciones demasiado costosas o incluso inmedibles en la actualidad (podemos pensar en los componentes epistáticos o algunos casos respecto a la variabilidad ambiental). Afortunadamente, la heredabilidad es una medida que teniendo en cuenta la componente genético-aditiva y la varianza fenotípica total nos permite tener una cuantificación de la similitud entre parientes para un carácter de interés. La selección (ya sea natural o artificial) solo puede operar sobre componentes hereditarios, por lo que resulta claro que para poder realizar un análisis racional del proceso de selección artificial es necesario poseer una medida que nos permita estimar en qué medida los valores fenotípicos observados en una generación se ven reproducidos en la siguiente; todo esfuerzo de selección será vano si la variabilidad fenotípica observada no se explica por un componente hereditario. Por lo tanto, ahora sí nos encontramos prontos para utilizar los conocimientos teóricos y prácticos desarrollados en este capítulo, combinado con nuestro conocimiento de los procesos selectivos, para llevar nuestra atención al proceso de selección artificial.

10.12 Actividades

10.12.1 Control de lectura

1. Define el concepto de heredabilidad, describiéndolo matemática y conceptualmente.
2. A partir de la observación de fenotipos es posible predecir la heredabilidad de un carácter de interés. ¿Cómo se logra esto? Ten en cuenta los posibles errores del método descrito.
3. Define el concepto de repetibilidad, describiéndolo matemática y conceptualmente.
4. ¿Cuál es la diferencia conceptual entre repetibilidad y heredabilidad? ¿Estos parámetros son referentes a un individuo o una población?
5. Es posible hacer una regresión de los valores fenotípicos de una progenie (o un conjunto de progenies) en *i)* los valores fenotípicos de un progenitor, o *ii)* del promedio de los progenitores. ¿Qué información se puede obtener en cada uno de estos procedimientos?

10.12.2 ¿Verdadero o falso?

1. Todos los componentes genéticos del fenotipo son heredables de padres a hijos.
2. La heredabilidad de una característica se define como $h = \sqrt{\frac{V_A}{V_P}}$ (o lo que es equivalente, $h^2 = \frac{V_A}{V_P}$).
3. La estimación de la heredabilidad a partir de caracteres fenotípicos por métodos de regresión es dependiente de qué carácter se desea estudiar.
4. La repetibilidad de una característica será siempre mayor o igual a la heredabilidad de la misma.
5. La producción más probable (PMP) de una población nos posibilita predecir un nuevo registro de una característica basándonos en un registro previo de la misma.

Soluciones

1. **Falso.** Según el modelo genético básico podemos descomponer la variabilidad fenotípica en un componente de efectos genéticos aditivos, de dominancia y ambiental. Los primeros dos son componentes que refieren a la genética de los individuos, pero solo el primero es heredable (en tanto los genotipos de un progenitor no son heredables por su progenie).

2. **Falso.** El símbolo utilizado para denotar a la heredabilidad es h^2, pero no se trata de una cantidad elevada al cuadrado. Matemáticamente, la heredabilidad se define como $h^2 = \frac{V_A}{V_P}$ (y no la raíz de esta fracción).

3. **Verdadero.** Esto se ve claramente al tener en cuenta que el coeficiente de Pearson para una recta de regresión entre heredabilidad y carácter fenotípico posee una magnitud de h (proporcional a la heredabilidad, h^2). Esto implica que la estimación de heredabilidad poseerá un error de estimación que depende de la heredabilidad del carácter estudiado.

4. **Verdadero.** Teniendo en cuenta la definición en términos de varianza de ambos parámetros (y recordando que no existen las varianzas negativas) podemos notar que la repetibilidad es mayor a la heredabilidad en una magnitud $\frac{V_D+V_I+V_{Ep}}{VP}$. Es decir, a

lo sumo la repetibilidad puede ser igual a la heredabilidad de una característica, pero nunca menor a esta.

5. **Falso.** Si bien la PMP nos permite predecir un registro de una característica basándonos en registros anteriores, es importante recordar que este no es un parámetro poblacional, sino que se calcula para cada individuo de una población.

10.12.3 Ejercicios

Ejercicio 10.1

(Inspirado en J. Felsenstein 2019)

Una población de tamanduás (*Tamandua tetradactyla*) posee un largo promedio de lengua de 63,2 cm, con una varianza de 8,4 cm^2. Adicionalmente, la pendiente de la regresión de los valores fenotípicos padre-hijo es de 0,43.

a. ¿Cuál es la heredabilidad del carácter fenotípico «largo de lengua»?

b. Si tomamos un solo registro de un macho y su lengua mide 70 cm, ¿cuál es su valor de cría para esta característica?

Solución

a. Recordemos que la pendiente de la regresión de los valores fenotípicos de una progenie en uno de sus progenitores equivale a $\frac{1}{2}h^2$, por lo que $0{,}43 = \frac{1}{2}h^2 \Rightarrow h^2 = 0{,}43 \times 2 = 0{,}86$.

b. El estimador del valor de cría (VC) cuando se cuenta con un solo registro es

$$\hat{A}_i = h^2(x_i - \bar{x})$$

donde $\bar{x}$ es la media poblacional para el carácter fenotípico de interés y x_i es el valor del registro del individuo.

Tenemos entonces

$$\hat{A}_i = 0{,}86(70 - 63{,}2)$$

$$\hat{A}_i \approx 5{,}85$$

Ejercicio 10.2

Un conjunto de investigadores se encontraban estudiando parámetros genéticos relacionados a la resistencia de la fibra de cáñamo (*Cannabis sativa*) con fines industriales. Sin embargo, víctimas de su propio desorden, obtuvieron resultados, pero no recuerdan a qué corresponde cada uno. En un cuaderno de laboratorio encontraron que un parámetro estimado era de 0,562 y otro de 0,734.

Otro grupo de investigación, más ordenado, estimó para esa población que la covarianza padre-hijo es de 0,142.

a. ¿Cuál de los valores encontrados por el grupo de investigación corresponde a la heredabilidad y cuál a la repetibilidad?

b. ¿Cuál es la estimación para la varianza fenotípica para la característica «resistencia de la fibra» en esta población?

Solución

a. Como podemos recordar, la repetibilidad es el límite superior de la heredabilidad, es decir, en un caso extremos ambas serían iguales si $\frac{V_D+V_I+V_{Ep}}{VP} = 0$. Por lo tanto, la estimación de 0,562 corresponde a la heredabilidad y la de 0,734 a la repetibilidad de la característica.

b. Para lograr una estimación de la varianza fenotípica de la característica, podremos valernos de dos datos: la heredabilidad y la covarianza que existe entre padres e hijos para ella.

Recordemos que $h^2 = \frac{V_A}{V_P}$ y que $COV_{PH} = \frac{1}{2}V_A$ por lo que

$$V_A = COV_{PH} \times 2 = 0{,}142 \times 2 = 0{,}284 \tag{10.115}$$

$$V_P = \frac{V_A}{h^2} = \frac{0{,}284}{0{,}562} \approx 0{,}505 \tag{10.116}$$

Ejercicio 10.3

(Tomado de Hartl 2020)

Demuestra que en una población en equilibrio de Hardy-Weinberg:

a. La heredabilidad en sentido estricto de una característica determinada completamente por un alelo autosomal recesivo con frecuencia q equivale a $2q/(1+q)$, lo cual implica que $h^2 \approx 0$ para $q \approx 0$. (**Pista:** si $V_E = 0$, entonces $V_G = V_A + V_D$)

b. La heredabilidad de una característica determinada por un solo alelo autosomal dominante con frecuencia q equivale a $2(1-q)/(2-q)$, lo cual implica que $h^2 \approx 1$ para $q \approx 0$.

Solución

a. Para un alelo recesivo $d = a$, y por lo tanto $\alpha = a + (q - p)d = 2qa$. Por lo tanto, $V_A = 2pq\alpha^2 = 8pq^3a^2$ y $V_D = 4p^2q^2a^2$. Luego, notemos que $h^2 = \frac{V_A}{V_A+V_D}$, expresión que queda reducida a $2q/(1 + q)$, como se quería demostrar.

b. En este caso tenemos que $d = -a$. Por lo tanto,
$V_A = 8p^3qa^2$ y $V_D = (2pqd)^2 = 4p^2q^2a^2$.
Por lo tanto, la expresión $h^2 = \frac{V_A}{V_A+V_D}$ se reduce a $2(1 - q)(2 - q)$

Ejercicio 10.4

El peso adulto de las ratas de laboratorio es un carácter de gran importancia debido a que puede afectar los resultados experimentales. En este sentido, se han estimado los siguientes valores de cría para los individuos Ratín ($VC = -5{,}12$), Ratón ($VC = 12{,}16$) y Ratún ($VC = -1{,}28$) pertenecientes a una población cuya media es 24 g. Los registros fenotípicos utilizados para estas estimaciones fueron de 16, 43 y 22 g respectivamente. ¿Cuál es la heredabilidad de la característica?

Solución

Recordemos que $VC = (x_i - \bar{x})h^2$ por lo que $h^2 = \frac{VC}{(x_i-\bar{x})}$, entonces podremos comprobar:

$$h^2 = \frac{-5{,}12}{(16 - 24)} = 0{,}64 \tag{10.117}$$

$$h^2 = \frac{12{,}16}{(43 - 24)} = 0{,}64 \tag{10.118}$$

$$h^2 = \frac{-1{,}28}{(22 - 24)} = 0{,}64 \tag{10.119}$$

Ejercicio 10.5

Un productor de porcinos se ha propuesto refugar dos cerdas que actualmente se encuentran en producción por falta de espacio en las parideras. Su criterio será basarse en el tamaño de camada que logra destetar cada cerda y refugar los animales que tengan las camadas más pequeñas al momento del destete. A su vez, se ha logrado estimar para la población que la varianza genética es de 1,2 individuos^2, la varianza genética no aditiva es de 0,4 individuos^2, la varianza ambiental permanente es de 2,1 individuos^2 y la varianza ambiental total es de 5,4 individuos^2. La media de esta población es de 9,9 lechones destetados. Por otra parte las candidatas a ser refugadas son

- Cerda 0837: 3 partos con camadas de 13, 9 y 8 lechones destetados
- Cerda 3759: 5 partos con camadas de 8, 5, 14, 0 y 9 lechones destetados
- Cerda 0193: 5 partos con camadas de 2, 9, 11, 8 y 8 lechones destetados
- Cerda 4132: 2 partos con camadas de 10 y 8 lechones destetados
- Cerda 2845: 1 parto con 11 lechones destetados

Con esta información,

a. ¿Qué cerdas refugarías?

b. ¿Cuál sería una estimación de la heredabilidad de la característica?

Solución

a. Para poder evaluar a estos animales como productores y realizar una elección de refugo, debemos evaluarlos en cuanto a su producción. Para ellos calcularemos la producción más probable (PMP) según la ecuación (10.106) tal que

$PMP(x_{n+1}) = \bar{x} + \frac{n\rho}{1+(n-1)\rho}(x_{n+1} - \bar{x})$

Pero previamente deberemos calcular la repetibilidad

$$\rho = \frac{\sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{EP}}}{\sigma^2_{\mathrm{P}}} = \frac{1{,}2 + 2{,}1}{1{,}2 + 5{,}4} = 0{,}50 \tag{10.120}$$

Ahora podemos calcular la PMP de cada cerda y tomar la decisión de refugo:

Cerda 0837

$$\mathrm{PMP} = 9{,}9 + \frac{3 \times 0{,}5}{1 + (3-1) \times 0{,}5} \times (10 - 9{,}9) = 10 \text{ lechones destetados} \tag{10.121}$$

Cerda 3759

$$\mathrm{PMP} = 9{,}9 + \frac{5 \times 0{,}5}{1 + (5-1) \times 0{,}5} \times (7{,}2 - 9{,}9) = 7{,}7 \text{ lechones destetados} \tag{10.122}$$

Cerda 0193

$$PMP = 9{,}9 + \frac{5 \times 0{,}5}{1 + (5-1) \times 0{,}5} \times (7{,}6 - 9{,}9) = 8 \text{ lechones destetados} \tag{10.123}$$

Cerda 4132

$$PMP = 9{,}9 + \frac{2 \times 0{,}5}{1 + (2-1) \times 0{,}5} \times (9{,}0 - 9{,}9) = 9{,}3 \text{ lechones destetados} \tag{10.124}$$

Cerda 2845

$$PMP = \mu + R(\bar{x}_i - \mu) = 9{,}9 + 0{,}5 \times (11 - 9{,}9) = 10{,}5 \text{ lechones destetados} \tag{10.125}$$

Entonces, las cerdas a ser refugadas serán la 3759 y 0193, debido a que sus PMP son las menores.

b. Como sabemos, la heredabilidad (h^2) puede calcularse como

$$h^2 = \frac{\sigma^2_{\mathrm{A}}}{\sigma^2_{\mathrm{P}}} \tag{10.126}$$

Entonces, utilizando la información de esta población podemos despejar la varianza aditiva (^2_A) ya que la varianza genética total es la suma de la varianza genético-aditiva y no aditiva. A su vez, la varianza fenotípica, es la suma de la varianza genética total y varianza ambiental total, por lo tanto

$$h^2 = \frac{1{,}2 - 0{,}4}{1{,}2 + 5{,}4} \approx 0{,}12 \tag{10.127}$$

Capítulo 11

Selección artificial I

¡Mire esta vaca, Serafín! Musa inspiradora de miles de composiciones escolares... ¡Y ahora es acusada de traficante de colesterol por el naturismo apátrida! Nos da su leche, su carne, su cuero. ¡Lo quiero ver a usté haciéndose una campera de zapayitos![1]

Inodoro Pereyra, el renegáu

Alcanza con observar las diferentes formas de los organismos vivientes para comprender, a la luz de la teoría de la evolución, el rol que la selección natural ha jugado en la adaptación de los organismos al ambiente. En lo que llevamos avanzado del presente material, hemos intentado ponderar el rol de las distintas fuerzas aleatorias y determinísticas en la evolución de los organismos. En el capítulo «Selección natural» estudiamos, a través de modelos muy simples y estilizados, cómo la selección natural puede cambiar las frecuencias de los alelos en un locus o apenas en un par de loci. Claramente, la selección natural llamó la atención de Charles Darwin[2] de tal manera que la experiencia como naturalista a bordo durante la expedición del *Beagle* le permitió formular las bases de la teoría evolutiva a partir de la selección natural. Sin embargo, la comprensión del mecanismo subyacente fue en gran parte a través de la observación del proceso de selección artificial que realizaban y realizan los mejoradores de diferentes razas de animales domésticos o de producción (y de los colombófilos).

En efecto, alcanza con observar la variedad de razas de perros para entender que la fuerte presión de selección por desarrollar determinadas características puede tener una respuesta concomitante en relativamente pocas generaciones. De hecho, la mayor parte de las razas de perros que conocemos en la actualidad surgieron a mitad del siglo XIX, en plena revolución industrial. Pese a tratarse de animales de la misma especie, resulta difícil imaginarse el apareamiento de un gran danés con un chihuahua (en realidad artificialmente es posible lograr crías de una hembra de gran danés con un macho Chihuahua, al revés es imposible la gestación). Todos los cambios y la diversidad observada desde el ancestro lobo domesticado se han obtenido como resultado de ese esfuerzo por modelar esta especie de acuerdo con nuestros gustos, costumbres y necesidades.

[1] Inodoro Pereyra, el renegáu, es un personaje de Roberto Fontanarrosa.

[2] Charles Darwin (12 febrero de 1809 – 19 abril de 1882) fue un naturalista inglés, fundador de la teoría de la selección natural. En el segundo viaje del *HMS-Beagle*, con el objetivo primario de cartografiar la costa sudamericana, Darwin viajó como naturalista, recogiendo observaciones geológicas y diferentes especímenes biológicos. Su obra más famosa, *On the Origin of Species*, fue publicada por primera vez en 1859.

Lo mismo es posible de observar en cada especie que tiene alguna importancia económica o cultural en nuestra vida diaria. En las diferentes razas bovinas vemos animales completamente especializados en la producción de leche, mientras que otros se especializan en la producción de carne. Las diferencias morfológicas son notables, pero aún más lo son en las características de interés productivo. El tamaño de los pollos parrilleros en los EE. UU. creció más de 4 veces desde 1957 a 1991 (Havenstein et al. 1994), en gran parte debido a la selección especializada para carne o producción de huevos.

En el presente capítulo vamos a comenzar a estudiar la selección artificial para características métricas. Para ello usaremos un escenario sencillo, esencialmente basado en lo que se conoce como selección por truncamiento, en el que solo los individuos que sobrepasen un valor umbral de la característica (asumiendo que los mejores se corresponden con valores altos, al revés si los mejores son los valores bajos) son considerados como reproductores y progenitores de la siguiente generación.

OBJETIVOS DEL CAPÍTULO

☐ Definir el concepto de diferencial de selección direccional, buscando comprender la relación entre la media poblacional en un momento dado y la obtenida luego de seleccionar una fracción de individuos.

☐ Introducir la *ecuación del criador*, que nos permite explorar la conexión entre estas medias fenotípicas y la heredabilidad de la característica seleccionada.

☐ Introducir el concepto de intensidad de selección, discutiendo cómo se relaciona con la proporción de individuos seleccionados.

☐ Obtener otra perspectiva del diferencial de selección direccional a través de la identidad de Robertson-Price.

11.1 Factores de corrección

El interés en mejoramiento genético suele estar puesto en elegir los «mejores» individuos para que se reproduzcan y usualmente dependemos de la información fenotípica para poder realizar la selección de los «mejores» individuos. Podríamos imaginar que para una característica fácil de medir, como lo son el peso de vellón sucio o limpio, la producción de leche diaria, el peso a los 18 meses en bovinos de carne o muchas otras, alcanza con medir y elegir a los individuos que poseen los valores «mejores» en la característica. Sin embargo, como se puede deducir de nuestro modelo genético básico, si bien el fenotipo es lo único con lo que contamos usualmente, no es de por sí una excelente referencia sobre el valor reproductivo de los individuos. Claramente, los efectos genéticos no-aditivos y el ambiente tienen una influencia usualmente importante en la manifestación fenotípica de la característica. Desafortunadamente, en general nos resulta imposible conocer los efectos de dominancia y

epistasis, así como los efectos ambientales que han influido en la vida del individuo, por lo que en ese sentido no podemos reducir la incertidumbre.

Sin embargo, existen una serie de efectos que se manifiestan siempre de la misma manera y que por lo tanto afectan los registros fenotípicos en forma sistemática. En mamíferos, por ejemplo, existen en general importantes diferencias en la expresión de características métricas asociadas al tamaño de los individuos. Si volvemos a mirar con atención la figura 10.1 del capítulo «Parámetros genéticos: heredabilidad y repetibilidad» veremos que antes de deducir la relación entre alturas de padres e hijos, Galton (1886), multiplica la altura de las mujeres por el coeficiente 1,08 a fin de que las mismas sean comparables. Es decir, Galton reconoce algo que todos observamos diariamente: que existen **diferencias sistemáticas** en las alturas de hombres y mujeres, siendo estas últimas más bajas. Eso no quiere decir que todas las mujeres sean más bajas que todos los hombres, ni que determinada mujer sea más baja que determinado hombre. Simplemente establece una relación entre la media de un grupo y la media del otro grupo. Claramente esto se aplica a muchas otras características, tanto en seres humanos como en otras especies. En algunas características un sexo poseerá registros promedialmente superiores al otro y en otras será a la inversa.

En este punto podríamos plantearnos si tiene algo de sentido lo que estamos discutiendo. Al ser características con distribuciones aleatorias (por todos los efectos que tienen esas distribuciones, genéticos y ambientales), es altamente probable que cualquier factor que utilicemos para separar los individuos (por ejemplo, el color de pelo, color de ojos, etcétera) genere diferencias, algunas de ellas estadísticamente significativas. ¿Qué hace entonces particular a los factores que estamos discutiendo en esta sección? Lo particular de los factores que estamos analizando es precisamente su carácter **sistemático**. Es decir, no importa si analizamos la diferencia de peso de vellón entre machos y hembras en Salto, Paysandú, Río Negro o en Tacuarembó (utilizando siempre un gran número de individuos para reducir la incertidumbre en las estimaciones), la diferencia observada en las medias de ambos sexos será similar en todos ellos. No importa si analizamos esas diferencias este año, el año pasado o tres años hacia atrás, las diferencias entre sexos serán similares.

Otras diferencias sistemáticas no están relacionadas al sexo de los individuos. Por ejemplo, hasta la cuarta o quinta lactancia las vacas producirán en cada lactancia un poco más que en la anterior; esto se debe al crecimiento que experimenta la vaca con el paso del tiempo y se trata de diferencias que son sistemáticas. Lo mismo ocurrirá con el peso de los borregos, que se verá afectado por el sexo del animal, por el hecho de haber transcurrido la gestación como mellizo o trillizo y aún por haber sido criado en los primeros meses como único, mellizo o trillizo; se trata en todos los casos de diferencias sistemáticas entre las clases consideradas y que por lo tanto se repetirán en signo y magnitud durante un intervalo de tiempo razonable (no sería razonable suponer que **esas mismas diferencias** se observaran hace dos siglos, por ejemplo, aunque sí el signo de estas).

Un punto importante es el qué hace al carácter genético o ambiental de los efectos y su consideración como efectos ambientales sistemáticos. Claramente, el sexo está determinado genéticamente, por lo que resulta llamativo que lo incluyamos en el mismo lugar que otros efectos ambientales. En cierta forma, lo mismo podría pensarse para la calidad de mellizos, que en ovinos está parcialmente controlada por los alelos presentes en el gen de fecundidad Booroola (**FecB**). El punto aquí es que para nuestros objetivos, mejorar la característica métrica de interés, tanto el sexo como la calidad mellicera son genéticamente irrelevantes. Para reproducir la especie necesitaremos individuos de ambos sexos y si no consideramos las correlaciones entre características (algo que veremos en el próximo capítulo) el hecho de

que sean mellizos o no es irrelevante para nuestro mejoramiento en el peso de cada borrego. Lo mismo, las diferencias entre producción de diferentes lactancias en la vaca tienen un correlato con la curva de crecimiento de los animales, la que se encuentra determinada en gran parte a nivel genético. Sin embargo, no tendría sentido para nosotros elegir todas las vacas que serán reproductoras entre las de quinta lactancia; debemos poder corregir por esas diferencias sistemáticas (esperables).

Esto nos lleva directamente a la pregunta fundamental ¿Por qué son tan importantes para nosotros las diferencias sistemáticas? La respuesta es relativamente sencilla: **porque podemos corregir fácilmente por estas diferencias** (ya que son siempre del mismo signo y magnitud), tratando de forma más justa a los diferentes candidatos a la selección. Tradicionalmente se aplicaron factores de corrección como forma de corregir por esas diferencias sistemáticas, calculándose los mismos previamente y luego ajustando los datos antes de procesarlos en la evaluación. Afortunadamente, desde hace tiempo, los modelos lineales mixtos usados en las evaluaciones permiten «corregir» directamente por los efectos sistemáticos, estimando sus efectos al mismo tiempo que los efectos genéticos y otros efectos (como el del grupo contemporáneo), por lo que no es ya más necesaria la precorrección. Sin embargo, a fines didácticos, el cálculo de los mismos sigue siendo interesante porque muestra más de cerca los problemas de comparar cosas que tienen diferentes bases y las formas de corregir por esas diferencias que no son relevantes a nuestros fines, por lo que incluimos esta sección.

¿Cómo funcionan los factores de corrección? Su funcionamiento es sencillo, pero cuenta con dos fases completamente diferentes:

1. **Cálculo de los factores de corrección** (que se hace en forma regular, cada determinado tiempo) **a partir de la mayor cantidad posible de registros** de la «misma» gran población.
2. **Aplicación del factor de corrección correspondiente a cada registro** que se pretenda utilizar en una evaluación genética.

Factores de corrección aditivos

Hasta aquí hemos venido discutiendo en términos generales acerca de los efectos sistemáticos, mencionando que, al tratarse de diferencias con el mismo signo y magnitud por un plazo razonable de tiempo, resultaba sencillo corregir por ellas. Sin embargo, no mencionamos hasta ahora cómo corregirlas. Lo primero a tener en cuenta a la hora de corregir por diferencias sistemáticas entre determinadas clases es entender las diferencias en las distribuciones de las clases (o niveles) del factor en consideración para la característica de interés. Para esto, un histograma o un gráfico de densidad de la distribución suele ser extremadamente útil, acompañado de los estadísticos descriptivos usuales (media, mediana, varianza, cuantiles).

En la figura 11.1 se representa una situación clásica para una característica métrica de interés. Por lo que se aprecia en la figura, los dos sexos presentan una distribución diferente ya que las medias de ambas distribuciones se encuentran desplazadas una respecto a la otra. En este caso, por otra parte, **las dos distribuciones tienen aproximadamente la misma varianza**, por lo que la forma de hacerlas coincidir parece obvia: o desplazamos la distribución de machos a la izquierda hasta que coincida con la de hembras, o desplazamos la de hembras

hasta que coincida con la de machos. En la figura nos hemos decidido por esta última opción, en forma totalmente arbitraria. Para ello calcularemos un factor de **corrección aditivo** para cada clase (o nivel) del factor, en este caso **sexo**. Para que el método sea general a más de dos niveles o clases, vamos a definir siempre una clase como la **clase de referencia**, clase **cuyos registros ajustados permanecerán idénticos** a pesar del ajuste. En el caso de la figura, elegimos la categoría de los **machos** como **clase de referencia**. Si llamamos $\bar{y}_r$ a la media de la clase de referencia, entonces los factores para la clase i se calcula como

$$c_i = \bar{y}_r - \bar{y}_i \tag{11.1}$$

En el caso de la clase de referencia, $\bar{y}_i = \bar{y}_r$, por lo que la ecuación anterior nos deja con

$$c_{i=r} = \bar{y}_r - \bar{y}_i = \bar{y}_r - \bar{y}_r = 0 \tag{11.2}$$

Es decir, para calcular los distintos factores de corrección aditivos alcanza con restarle a la media de las clases de referencia las medias de cada una de las clases.

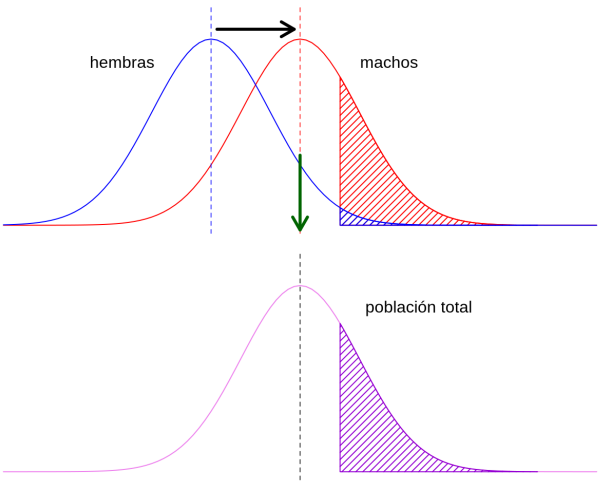


Figura 11.1. Factor de corrección aditivo. La característica de interés tiene una distribución diferente en hembras y machos ya que, si bien su varianza es igual, las medias para ambos sexos son diferentes. Si definimos un umbral de selección determinado, si no corregimos por esta diferencia casi ninguna hembra resultaría seleccionada (la pequeña porción rayada en azul en la parte superior de la figura). Como la diferencia entre las distribuciones es apenas en la media, alcanza con desplazar la media de hembras hasta alcanzar la media de machos (podríamos haber hecho lo contrario), categoría que usaremos como referencia. Esto produce que las dos distribuciones ahora coincidan (parte inferior de la figura) y que el umbral de selección brinde las mismas posibilidades de ser elegidos a machos y hembras.

El paso final, luego de haber calculado los distintos factores de corrección será aplicarlos sobre los distintos registros. Para eso alcanza con sumar al registro fenotípico el valor del

factor de corrección correspondiente a la clase a la que pertenece el registro. Es decir, si un registro fenotípico j pertenece a la clase i, su valor ajustado será

$$y_j = y_{i,j} + c_i \tag{11.3}$$

Notemos que esta notación desapareció el subíndice i luego de la corrección ya que todas las clases serán comparables en pie de igualdad luego de ajustadas.

Ejemplo 11.1

Supongamos que tenemos un grupo de vacas lecheras, y queremos seleccionar como reproductora a la que produce mayor volumen de leche por lactancia.

Como información sabemos que, a nivel poblacional para producción de leche, distintas subclases de lactancia presentan diferente media e igual varianza:

N.º Lactancia	Media leche 305d (litros)	Varianza leche 305d (litros^2)
1	3897	$1{,}3 \times 10^6$
2	4481	$1{,}3 \times 10^6$
3	4740	$1{,}3 \times 10^6$

¿Cuál de las siguientes vacas seleccionaríamos? Calcula la producción de leche corregida para cada vaca, planteado como subclase de referencia a la lactancia 2.

Vaca	N.º Lactancia	Leche 305d (litros)
$I801$	1	4000
$J805$	1	3500
$K709$	2	4900
$L713$	2	6000
$M607$	3	5600

Lo primero que tenemos que hacer es calcular los factores de corrección para cada clase (número de lactancia), lo que hacemos eligiendo una clase de referencia, restándole luego a la media de dicha clase la media de cada clase, es decir (si elegimos la segunda lactancia como referencia)

$$\begin{aligned} \mathbf{a_1} &= \bar{y}_r - \bar{y}_1 = 44813897 = 584 \\ \mathbf{a_2} &= \bar{y}_r - \bar{y}_2 = 44814481 = 0 \\ \mathbf{a_3} &= \bar{y}_r - \bar{y}_3 = 44814740 = 259 \end{aligned} \tag{11.4}$$

A partir de estos factores de corrección (poblacionales) debemos corregir cada registro individual, sumándole el factor de corrección que le corresponda. Por lo tanto, para cada una de las 5 vacas, la producción corregida será igual a

$$y^*_{I801} = y_{I801} + a_1 = 4000 + 584 = 4584 \text{ litros} \tag{11.5}$$

$$y^*_{J805} = y_{J805} + a_1 = 3500 + 584 = 4084 \text{ litros} \tag{11.6}$$

$$y^*_{K709} = y_{K709} + a_2 = 4900 + 0 = 4900 \text{ litros} \tag{11.7}$$

$$y^*_{L713} = y_{L713} + a_2 = 6000 + 0 = 6000 \text{ litros} \tag{11.8}$$

$$y^*_{M607} = y_{M607} + a_3 = 5600 + (-259) = 5341 \text{ litros} \tag{11.9}$$

Seleccionaríamos a la vaca **L713**, la cual presenta la mayor producción de leche corregida.

Factores de corrección multiplicativos

Como discutimos antes, los factores de corrección aditivos se pueden aplicar cuando las diferencias en las distribuciones de los fenotipos de las clases son exclusivamente en las medias de cada una de ellas. En esos casos alcanza con sumar un valor (el **factor de corrección aditivo**) a cada uno de los registros de la clase para que todos los registros queden expresados en términos equivalentes respecto al factor en consideración. Sin embargo, en muchos casos no solo las medias difieren sino que también lo hacen las dispersiones de cada clase. Un ejemplo típico es el de la producción de leche por lactancia. Mientras que en las primeras lactancias la producción es promedialmente menor que en las segundas lactancias, la variación también es menor en las primeras respecto a las segundas. Esto se extiende a su vez a las terceras lactancias, como puede verse en el ejemplo de la figura 11.2, en la parte superior de la misma. Si utilizáramos un umbral de selección sin corregir, casi ninguna vaca de segunda lactancia sería elegida y ninguna de las vacas en primera lactancia lo sería. Más aún, si aplicásemos un factor de corrección aditivo tomando como referencia la 3.ª lactancia, solamente desplazaríamos las curvas de 1.ª y 2.ª lactancia hacia la derecha, pero al quedar con menor variación que la curva de 3.ª tampoco estaríamos siendo justos con las vacas en 1.ª o 2.ª lactancia (que se elegirían en menor proporción que las de 3.ª). Por lo tanto, hace falta aplicar alguna corrección que al mismo tiempo que desplace las medias para hacerlas coincidir con la media de 3.ª lactancia, también expanda las varianzas de las distribuciones. Nuevamente, si llamamos $\bar{y}_r$ a la media de la clase de referencia, ahora el factor de corrección para la clase i estará dado por

$$c_i = \frac{\bar{y}_r}{\bar{y}_i} \tag{11.10}$$

En el caso de la clase de referencia, $\bar{y}_i = \bar{y}_r$, por lo que la ecuación anterior nos deja con

$$c_{i=r} = \frac{\bar{y}_r}{\bar{y}_i} = \frac{\bar{y}_r}{\bar{y}_r} = 1 \tag{11.11}$$

o lo que es lo mismo, un factor multiplicativo de $c_r = 1$ para la clase de referencia.

Nuevamente, el paso final, luego de haber calculado los distintos factores de corrección será aplicarlos sobre los distintos registros. Para eso alcanza con multiplicar el valor del registro

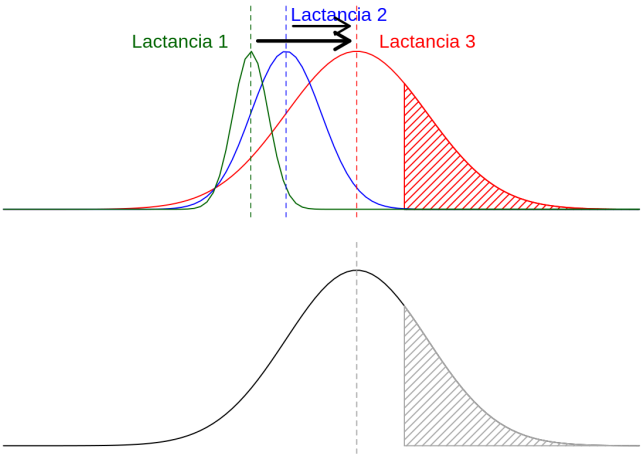


Figura 11.2. Factor de corrección multiplicativo La característica de interés, producción de leche a 305 días, posee una distribución diferente en cada lactancia (1.ª, 2.ª y 3.ª) y así como crece las medias de la característica también lo hacen las varianzas. Si definimos un umbral de selección determinado, si no corregimos por esta diferencia casi ninguna hembra segunda y ninguna de primera lactancia resultaría seleccionada (la pequeña porción rayada en azul en la parte superior de la figura y la no existente rayada en verde). Como la diferencia entre las distribuciones es en ambos parámetros (medias y varianzas), ambos en el mismo sentido, es posible que si definimos la 3.ª lactancia como categoría de referencia, al multiplicar los registros de las otras dos categorías por su correspondiente **factor de corrección multiplicativo** estemos corrigiendo ambos parámetros, como se ve en la figura de abajo.

fenotípico por el valor del factor de corrección correspondiente a la clase a la que pertenece el registro. Es decir, si un registro fenotípico j pertenece a la clase i, su valor ajustado será

$$y_j = y_{i,j} \times c_i \tag{11.12}$$

Igual que para el factor de corrección aditivo, observar que en esta notación desapareció el subíndice i luego de la corrección ya que todas las clases serán comparables en pie de igualdad luego de ajustadas.

Ejemplo 11.2

Supongamos que tenemos un grupo de terneros, y queremos seleccionar al de mayor peso al destete.

Como información sabemos que, a nivel poblacional para peso al destete, distintas subclases de edad de la madre presentan **diferente media** y **diferente varianza** (creciente con la edad de la madre):

Edad de la madre (años)	Media peso al destete (kg)
3	157
4	166
5	169

¿Cuál de los siguientes terneros seleccionaríamos? Calcula el peso al destete corregido para cada ternero, planteado como subclase de referencia a la edad de la madre de **4** años.

Ternero	Edad de la madre (años)	Peso al destete (kg)
$P101$	3	132
$Q108$	3	145
$R114$	4	140
$T123$	5	150

Lo primero que debemos hacer es nuevamente calcular los factores de corrección, en este caso multiplicativos. Para ello vamos a **dividir** la media de la clase de referencia entre el valor de la media de cada clase de **edad de la madre**, es decir

$$m_3 = \frac{\bar{y}_r}{\bar{y}_3} = \frac{166}{157} = 1{,}06 \tag{11.13}$$

$$m_4 = \frac{\bar{y}_r}{\bar{y}_4} = \frac{166}{166} = 1{,}0 \tag{11.14}$$

$$m_5 = \frac{\bar{y}_r}{\bar{y}_5} = \frac{166}{169} = 0{,}98 \tag{11.15}$$

Luego, para tener los registros corregidos por el efecto de la edad de la madre alcanza con multiplicar el valor de cada registro por su correspondiente factor de corrección, es decir

$$y^*_{P101} = y_{P101} \times m_3 = 132 \times 1{,}06 = 140 \text{ kg} \tag{11.16}$$

$$y^*_{Q108} = y_{Q108} \times m_3 = 145 \times 1{,}06 = 154 \text{ kg} \quad (11.17)$$

$$y^*_{R114} = y_{R114} \times m_4 = 140 \times 1 = 140 \text{ kg} \quad (11.18)$$

$$y^*_{T132} = y_{T132} \times m_5 = 150 \times 0{,}98 = 147 \text{ kg} \quad (11.19)$$

Seleccionaríamos al ternero **Q108**, el cual presenta el mayor peso al destete corregido.

Un detalle práctico final a la hora de observar las diferencias entre factores aditivos y multiplicativos es la que tiene que ver con la relación que existe en el elemento neutro de las operaciones de adición y multiplicación y los correspondientes valores de las clases de referencia para factores de corrección aditivos y multiplicativos, respectivamente. Llamamos clase de referencia a aquella a la que íbamos a referir todos los registros y que asumíamos como que no cambiaba luego de ajustados los registros. Es decir, luego de aplicar la operación correspondiente al tipo de corrección, los registros quedaban intactos. En la adición, el único elemento que sumado a otro deje sin cambiar el valor de este último es el cero, neutro de la suma. Por lo tanto, **el factor de corrección aditivo correspondiente a la clase de referencia debe ser cero**. De la misma manera, el único elemento que al multiplicar a un número por este deja sin cambio al último es el uno, el neutro de la multiplicación. Por lo tanto, **el factor de corrección multiplicativo correspondiente a la clase de referencia debe ser uno**. Esto tan sencillo puede ser una ayuda importante si poseemos una tabla de factores de corrección pero no sabemos si se trata de factores de corrección aditivos o multiplicativos.

Factores de corrección lineales

En algunos casos existe un gran número de categorías posibles o aún infinitas, por lo que calcular un factor de corrección aditivo para cada una de ellas resulta impráctico o aún imposible. Ejemplo de esto podría ser la necesidad de corregir pesos de acuerdo con el crecimiento del animal. Por ejemplo, si definimos una característica **peso al destete (a los 205 días)**, lo ideal sería pesar a todos los animales exactamente cuando cumplen 205 días. Sin embargo, en la práctica resulta impracticable esta idea ya que los animales suelen parir en un período extendido de tiempos y deberíamos estar pesando todos los días. La resultante de esto es que algunos animales serán pesados exactamente cuando tienen 205 días, pero unos cuantos lo serán con menos días y otros varios con más de 205 días. ¿Cómo corregimos entonces por estas diferencias en el número de días?

Si asumimos que en intervalos cortos de tiempo, por ejemplo, hasta un par de semanas después de los 205 días, el comportamiento de la evolución del peso con los días se corresponde a un crecimiento lineal, entonces podríamos ajustar una recta (por mínimos cuadrados, por ejemplo) de los pesos como función de los días desde el nacimiento, si tuviésemos esos varios datos para cada animal. **La pendiente de esta recta será la ganancia diaria**, que llamaremos b_{PE}, por ejemplo, en kg/día, que es el elemento central en la corrección lineal que vamos a aplicar.

En caso de no tener esa posibilidad, si tenemos el peso al nacimiento (P_N), además del peso a la edad cercana a los 205 días (P_d), por ejemplo a los d días, podemos estimar esa pendiente de ganancia diaria como

$$b_{PE} = \frac{(P_d - P_N)}{d} \tag{11.20}$$

Si llamamos b_{PE} a la pendiente de regresión del peso en la edad (cualquiera sea el método para obtenerla) para el animal en particular que queremos corregir, es decir a **la ganancia diaria esperada para ese animal**, entonces un registro de peso P_d obtenido a los d días se corregirá a 205 días de acuerdo con la siguiente ecuación

$$P_{205} = P_N + b_{PE} \times 205 \tag{11.21}$$

Dicho de otra forma, el registro corregido a 205 días será igual al peso al nacimiento sumado a la ganancia diaria esperada multiplicada por los 205 días establecidos como referencia. En el caso de que $d = 205$, entonces $b_{PE} = \frac{(P_d - P_N)}{205}$, por lo que al sustituir en la ecuación (11.21) tenemos

$$P_{205} = P_N + b_{PE} \times 205 = P_N + \frac{(P_d - P_N)}{205} \times 205 = P_N + P_d - P_N = P_d \tag{11.22}$$

PARA RECORDAR

- Existen una serie de efectos o factores que se manifiestan siempre de la misma manera y que por lo tanto afectan los registros fenotípicos en forma sistemática, es por ello que tenemos la posibilidad de corregir por estas diferencias con el fin de hacer sus medidas comparables.
- El funcionamiento de los **factores de corrección** consta de dos fases:

1) Cálculo de los factores de corrección a partir de la mayor cantidad posible de registros de la «misma» gran población.
2) Aplicación del factor de corrección correspondiente a cada registro que se pretenda utilizar en una evaluación genética.

- Los **factores de corrección aditivos** los utilizamos cuando las distribuciones de las características a corregir presentan **diferente media, pero tienen aproximadamente la misma varianza**. Para aplicarlos se calcula **un factor de corrección para cada clase del factor** donde se define siempre una **clase de referencia** cuyos registros ajustados permanecerán idénticos a pesar del ajuste. Llamamos $\bar{y}_r$ a la media de la clase de referencia, entonces los factores para la clase i se calculan como $c_i = \bar{y}_r - \bar{y}_i$
- En el caso de la clase de referencia de los factores aditivos, $\bar{y}_i = \bar{y}_r$, por lo tanto: $c_{i=r} = \bar{y}_r - \bar{y}_i = \bar{y}_r - \bar{y}_r = 0$. En el caso de los registros del resto de las clases, teniendo un registro fenotípico j que pertenece a la clase i, su valor ajustado será $y_j = y_{i,j} + c_i$

- En los casos en los que no solo las medias difieren, sino que también lo hacen las dispersiones de las clases, debemos recurrir a **factores de corrección multiplicativos**. Si denominamos $\bar{y}_r$ a la media de la clase de referencia, el factor de corrección para la clase i será $c_i = \frac{\bar{y}_r}{\bar{y}_i}$
- Para la clase de referencia de los factores multiplicativos, $\bar{y}_i = \bar{y}_r$, por lo que $c_{i=r} = \frac{\bar{y}_r}{\bar{y}_i} = \frac{\bar{y}_r}{\bar{y}_r} = 1$. En el caso de los registros del resto de las clases, teniendo un registro fenotípico j que pertenece a la clase i, su valor ajustado será $y_j = y_{i,j} \times c_i$
- Los **factores de corrección lineales** son aquellos a los que recurrimos cuando el número de categorías a corregir es tal que se convierte en impráctico o imposible la determinación de un factor para cada una de ellas.
- Tomando como ejemplo la ganancia de peso de terneros, podríamos ajustar una recta de los pesos como función de los días desde el nacimiento (si tuviésemos varios datos para cada animal). En caso de no contar con varios registros, podremos realizar una estimación utilizando el peso al nacimiento P_N y un peso cercano a los 205 días (P_d; como ejemplo), con d días podremos determinar: $b_{PE} = \frac{(P_d - P_N)}{d}$
- Siendo b_{PE} la pendiente de la recta de la regresión del peso en la edad para el animal que queremos corregir, la misma representará la ganancia media diaria esperada y será el elemento central en la corrección lineal que vamos a aplicar. Entonces, un registro de peso P_d obtenido a los d días se corregirá a 205 días de acuerdo con $P_{205} = P_N + b_{PE} \times 205$

11.2 La respuesta a la selección y su predicción

Cualquiera que haya salido al campo puede fácilmente reconocer las diferencias entre las razas de ganado bovino, por ejemplo, entre vacas lecheras y vacas destinadas a la producción de carne. No solo son claramente observables las diferencias en color (en nuestro país las razas más abundantes en ganado de carne son la hereford y aberdeen angus, mientras que en leche es holando), sino también en el tamaño y en el desarrollo morfológico del animal. Aun para aquellos que nunca salieron al campo, resultan obvias las diferencias entre razas de perros, tamaños, colores, formas de la cabeza, etcétera. Tanto en ganado bovino como en cánidos domésticos, las diferencias que observamos no fueron producto de la evolución actuando libremente, sino que mayormente fueron guiadas por el ser humano y moldeadas de acuerdo con sus intereses. Vista la diversidad y el poco tiempo (en términos evolutivos) en el que la humanidad realizó la selección artificial, resulta bastante llamativo el éxito obtenido.

Las bases de la selección artificial tradicional son muy sencillas y razonables: si una característica se hereda, entonces eligiendo en cada generación a los «mejores» individuos como reproductores (padres de la siguiente generación) es de esperar que la característica vaya «mejorando». Este razonamiento ha estado con nosotros desde el inicio de la domesticación, tanto en plantas, animales, como en microbiología (mejores fermentos, etcétera). Por ejemplo, si obtener mayor cantidad de lana por oveja parece una buena idea, entonces también lo parece elegir (seleccionar) a los mejores animales (aquellos con más lana) como reproductores y progenitores de la próxima generación. En principio alcanza con repetir este proceso tantas veces como sea necesario para llegar a nuestro objetivo.

Veamos cómo opera esta idea concretamente. Supongamos que tenemos una característica fenotípica de nuestro interés y que deseamos individuos con valores altos de la misma, es decir, cuanto más a la derecha de la media mejor. Para ser objetivos, definimos un determinado valor **umbral**, que indica que individuos son aceptables para reproducirse y seleccionamos a los que superen dicho umbral. Una representación de esto puede verse en la parte superior de la figura 11.3, donde el área rayada en rojo denota los individuos seleccionados en la generación parental. La diferencia entre la media de los individuos seleccionados $\bar{x}_S$ y la media de la población $\bar{x}_0$ será igual al diferencial de selección, es decir $S = \bar{x}_S - \bar{x}_0$. Aunque puede parecer trivial, es un error frecuente confundir el valor umbral a partir del cual vamos a elegir los individuos que serán reproductores con la media de la característica para los reproductores seleccionados; en el caso de que estemos seleccionando para valores altos de la característica, el valor umbral es el mínimo de la misma para que los individuos sean considerados como reproductores, por lo que la media será en general mayor que el umbral (en el peor de los casos será igual).

> **Diferencial de selección** (*S*) es la diferencia entre la media de los individuos seleccionados como progenitores de la siguiente generación respecto a la media de la población contemporánea. En selección por truncamiento (selección masal), la media de los animales seleccionados debe tener un valor más favorable (más alto si estamos seleccionado para valores altos) que el **umbral de selección**. El diferencial de selección, al tratarse de una resta de medias de la característica, tendrá las mismas unidades que la característica.
> **Umbral de selección** es un valor fijo a partir del cual elegimos a los individuos que serán reproductores de la siguiente generación, es decir, aquellos individuos que tengan un fenotipo igual o más favorable que el valor umbral.

Como ya vimos en el capítulo «Parámetros genéticos: heredabilidad y repetibilidad», no todas las diferencias que observamos en los padres respecto a la media será observada en la progenie, un fenómeno que ya había sido observado por Francis Galton y descrito como **regresión hacia la mediocridad**. También vimos previamente que **la pendiente de la regresión del fenotipo de los hijos en el promedio del fenotipo de los padres es igual a la heredabilidad** (h^2). Más aún, como adelanto de este capítulo vimos en la sección «Heredabilidad lograda» que la pendiente de la recta que relacionaba la respuesta a la selección R en función del diferencial de selección S era $\frac{R}{S} = h^2$. Dado que la regresión lineal del valor fenotípico de los hijos en el promedio de los padres (seleccionados) es una función lineal, entonces la media de la regresión es igual a la regresión de la media. Dicho de otra manera, si usamos el coeficiente de regresión para, a partir del valor fenotípico de los padres seleccionados, obtener el valor fenotípico **predicho** de los hijos y luego le hacemos la media a ese valor, sería lo mismo que predecir la media de la generación de hijos ($\bar{x}_1$) usando el coeficiente de regresión ($b_{O\bar{P}} = h^2$) sobre la media de los padres seleccionados ($\bar{x}_0$). Como el coeficiente de esta regresión lo habíamos obtenido de datos centrados, tanto en la variable dependiente como en la independiente (no tiene intercepto), la predicción debemos hacerla en los mismos términos. Es decir, en términos matemáticos, la predicción será a partir de $\bar{x}_S - \bar{x}_0 = S$

Como muestra la figura 11.3, la media de la nueva generación será $\bar{x}_1$ y por lo tanto, la respuesta a la selección será la diferencia entre la media de la generación de los progenitores

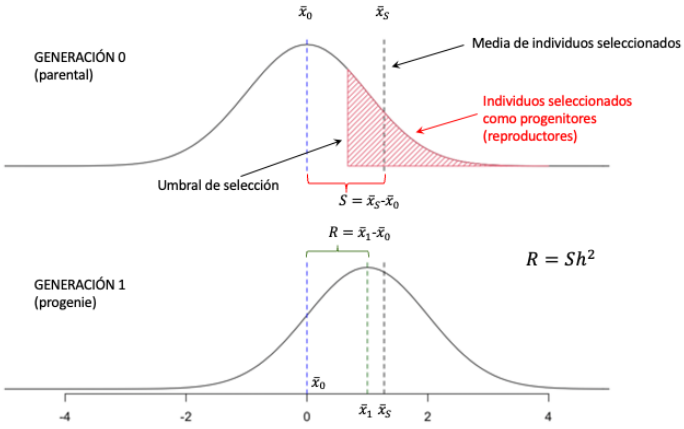


Figura 11.3. Respuesta a la selección. En la parte superior aparece la distribución de valores fenotípicos centrados de una característica cuantitativa en la generación parental. Si realizamos **selección masal**, luego de definido un valor **umbral**, los individuos que superen (en este caso) dicho valor son elegidos para reproducirse (área rayada en rojo). Claramente, la media de la característica debe ser (en este caso) mayor al valor umbral. La diferencia entre la media de los seleccionados y la media de su generación es el diferencial de selección $S = \bar{x}_S - \bar{x}_0$. La descendencia de los animales seleccionados constituirá la siguiente generación (distribución de abajo). La respuesta a la selección será la diferencia de medias entre la población en la generación 1 y en la generación 0, es decir $R = \bar{x}_1 - \bar{x}_0$. La relación entre la respuesta y lo elegido como reproductores es igual a la heredabilidad de la característica, es decir $\frac{R}{S} = h^2$ y por lo tanto $R = S\,h^2$, que se conoce como la ecuación del criador.

y la media de los descendientes, es decir $R = \bar{x}_1 - \bar{x}_0$. Por lo tanto, usando como criterio de predicción de la media de la generación de los hijos ($\bar{x}_1$) respecto a la media de la generación parental ($\bar{x}_0$) el desvío de la media de los animales seleccionados respecto a la media de su generación ($S = \bar{x}_S - \bar{x}_0$), tenemos

$$(\bar{x}_1 - \bar{x}_0) = R = (\bar{x}_S - \bar{x}_0) b_{O\bar{P}} = (\bar{x}_S - \bar{x}_0) h^2 \;\; \therefore \tag{11.23}$$

$$R = S\,h^2 \tag{11.24}$$

que es uno de los resultados más importantes de la genética cuantitativa. La ecuación (11.24) es central para el mejoramiento genético y es conocida como la **ecuación del criador**. El mensaje de la misma es muy claro: **cuando la heredabilidad es alta, la respuesta a la selección basada en el fenotipo individual será muy eficiente** y **la media de la nueva generación será cercana a la media de los animales seleccionados como reproductores**. Por otro lado, **cuando la heredabilidad es muy baja, la respuesta a la selección será pobre** y **la media de la nueva generación será casi igual a la de la generación anterior**. Desafortunadamente, como lo plantea Houchmandzadeh (2014), el alcance de la linealidad implícita en la ecuación del criador se encuentra restringida al caso en que tanto en los genotipos como los factores ambientales tienen distribución normal. Pese a esto, dado que para muchas características de interés agronómico parece ser un supuesto razonable, seguiremos con esta forma particular de la relación entre diferencial de selección y respuesta.

Otra cosa que seguramente te llamó la atención es el hecho de que pese a que la distribución de los individuos seleccionados como progenitores (el área rayada en rojo en la figura 11.3) no es una distribución normal, ni siquiera simétrica, hemos dibujado para los descendientes de estos nuevamente una distribución normal.

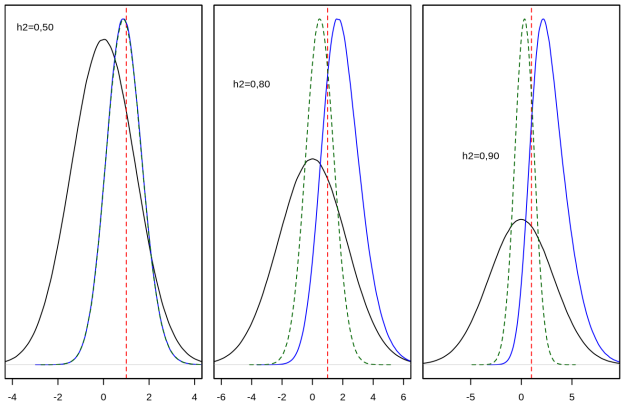


Figura 11.4. Distribución fenotípica de la población (en negro), de los fenotipos seleccionados como reproductores (área rayada en rojo) y de los genotipos seleccionados (azul), así como de los efectos ambientales (verde) en los individuos seleccionados. De izquierda a derecha, heredabilidades de $h^2 = 0{,}50$, $h^2 = 0{,}80$ y $h^2 = 0{,}90$. Se aprecia claramente que, pese a que la selección es por truncamiento en los fenotipos, los genotipos **de los individuos seleccionados** se distribuyen en casi todo el continuo de valores y con una distribución aproximadamente normal a heredabilidades no demasiado altas.

Por ejemplo, la densidad de los individuos que están cerca del valor umbral es mucho mayor que la densidad de los individuos que están hacia valores extremos a la derecha en la distribución. Entonces, la pregunta sería ¿por qué razón los descendientes tendrían pese a esto una distribución **aproximadamente** normal? La razón es intuitiva: como **genotipo y ambiente son independientes** (y a su vez los componentes genéticos son ortogonales entre sí), los individuos que están cerca del umbral (límite inferior de los seleccionados) pueden provenir de muchas combinaciones de genotipos y ambientes. Por ejemplo, muy buen genotipo y ambiente mediocre, relativamente buen genotipo y ambiente, o excelente ambiente y mediocre genotipo. En cambio, a medida que nos desplazamos hacia la derecha en los fenotipos seleccionados, ambos factores, genéticos y ambientales deben ser muy buenos para que su suma sea excelente. Esto nos deja con una distribución de genotipos seleccionados (y en particular los valores aditivos) que tiene una distribución mucho más simétrica que la zona roja rayada, como se puede apreciar en la figura 11.4. En ella se grafican para tres valores de heredabilidad diferentes las distribuciones de los efectos genotípicos y ambientales de los animales seleccionados. Claramente, pese a que realizamos selección por truncamiento, tanto los efectos de los genotipos como los ambientales tienen una distribución continua y aproximadamente normal. Además, para generar los descendientes de esta distribución los apareamientos serán al azar y en este caso lo único relevante son los valores aditivos, que tienen una distribución bastante razonable y a la que se le sumarán el resto de los efectos aleatorios (dominancia, epistasis y ambiente), por lo que no es difícil imaginar que el resultado será aproximadamente normal.

Como en la ecuación del criador la respuesta a la selección es directamente proporcional al diferencial de selección y a la heredabilidad, incrementos en cualquiera de las dos llevan a incrementar la respuesta a la selección. Sin embargo, el significado de ambas es conceptualmente distinto. Por un lado, mientras que el diferencial de selección se refiere a una decisión que podemos adoptar casi a nuestra voluntad en una población reducida (rodeo, majada, etcétera), la heredabilidad se trata de un parámetro genético poblacional que requiere de mucho esfuerzo para ser modificado en el sentido deseado. Como vimos previamente, la heredabilidad es igual a la proporción de la varianza fenotípica que corresponde a los efectos aditivos, es decir, a la varianza aditiva. Por lo tanto, la única forma de incrementarla, dada la arquitectura genética de una característica, es reducir la varianza ambiental. En muchas características esto es posible a nivel del predio, con manejo más controlados, mejores registros que permitan de alguna manera «corregir» por diferencias, etcétera, pero en general suele ser algo complicado y costoso.

Por otro lado, el diferencial de selección es una función de la varianza de la característica y de la proporción de individuos que decidamos quedarnos como reproductores. Cuanta más varianza fenotípica tenga la característica en la población en la que estoy haciendo la selección, mayor podrá ser el diferencial de selección a la misma proporción de individuos seleccionados. En general, la idea es mover la media lo más hacia el extremo deseado que sea posible, pero en casi cualquier distribución razonable eso suele implicar cada vez menos individuos que cumplen con los criterios de inclusión (que superen el umbral, por ejemplo).

En la figura 11.5 se puede apreciar la relación existente entre el diferencial de selección y la varianza de la distribución a partir de un valor de umbral prefijado.

En la parte superior de la figura podemos apreciar la gráfica de una característica con distribución normal que por comodidad hemos centrado restándole el valor de la media, por lo que ahora la media será cero y con desvío estándar igual a 1 unidad de la característica (por ejemplo, kg en peso de vellón, micras en diámetro, etcétera). Si definimos un umbral de selección, por ejemplo de $u = 0{,}6745$, todos los individuos que superen ese valor en su fenotipo serán considerados como reproductores, es decir, progenitores de la siguiente generación. Con ese umbral, en la distribución normal estándar (media igual a cero y desvío estándar igual a 1) quedan $25\,\%$ de individuos en el área seleccionada (rayada en rojo) y la media de la característica para esos individuos es de $\bar{x}_S = 1{,}27$. Como la media de la característica para toda la población en nuestro caso es de $\bar{x}_0 = 0$ (porque la centramos), entonces el diferencial de selección será de $S = \bar{x}_S - \bar{x}_0 = 1{,}27 - 0 = 1{,}27$ unidades de la característica. Por otra parte, en la parte inferior de la figura se representa la situación para la misma característica, pero en una población cuyo desvío estándar fenotípico es de $2/3$ (menor a 1 que era el valor de la distribución en la parte superior). Ahora, utilizando el mismo umbral de selección que utilizamos en la población de arriba, es decir $u = 0{,}6745$, ahora tendremos solamente el $16\,\%$ de los individuos como reproductores en el área rayada en rojo y el valor de la media para los mismos será $\bar{x}_S = 1{,}02$, por lo que el diferencial de selección será ahora de $S = \bar{x}_S - \bar{x}_0 = 1{,}02 - 0 = 1{,}02$ unidades de la característica.

En conclusión, pese a que en ambos casos el valor de umbral de selección utilizado es el mismo, en el primer caso la varianza mayor hace que una proporción mayor de los individuos cumpla con el criterio de selección y además, en el caso de la distribución normal, que el diferencial de selección sea mayor.

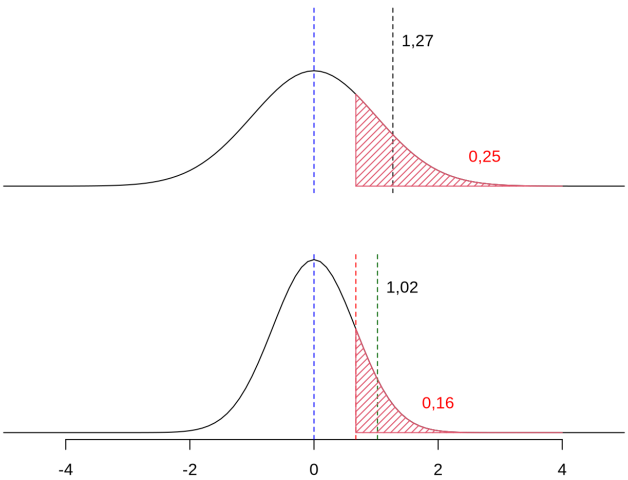


Figura 11.5. La relación entre el diferencial de selección y la varianza fenotípica de la característica. En la parte superior se ilustra la distribución fenotípica de una característica respecto a la media de la misma (es decir, con media de cero) y con desvío estándar de 1 unidad de la característica. Si decidimos hacer selección por truncamiento, eligiendo arbitrariamente un umbral $u = 0{,}6745$ (el percentil 75 de la distribución normal estándar), los individuos que pasen ese valor serán consideramos como reproductores. En este caso, 25 % de los individuos cumplen esa condición y la media de los individuos seleccionados es de $\bar{x}_S = 1{,}27$, por lo que el diferencial de selección será $S = \bar{x}_S - \bar{x}_0 = 1{,}27 - 0 = 1{,}27$. En la parte inferior, la población (para la misma característica) tiene un desvío de 2/3, inferior al de la población en la parte superior; esto implica que ahora solo el 16 % de los individuos serán seleccionados como reproductores. Además, el diferencial de selección será ahora de $S = \bar{x}_S - \bar{x}_0 = 1{,}02 - 0 = 1{,}02$, inferior al de la población de la gráfica superior, pese a que el valor de umbral es el mismo en ambos casos.

Ejemplo 11.3

En la figura 11.6 se presentan diferentes situaciones de proporción de selección representadas gráficamente, suponiendo características de distribución normal. Determinar para cada una de ellas la correspondiente intensidad de selección

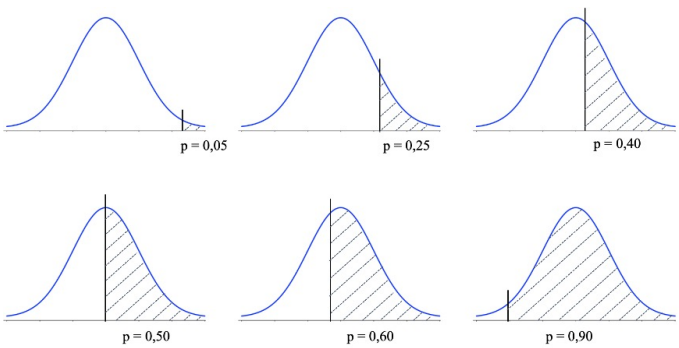


Figura 11.6. Proporción de selección representada gráficamente, suponiendo características de distribución normal

Asumiendo normalidad, podemos usar la relación $i = \frac{z}{p}$, donde p es la proporción de animales seleccionados. A partir de esta relación podemos generar una tabla con la proporción de animales seleccionados (expresada en %) y la consiguiente intensidad de selección i:

Proporción %	i
1	2,6652
2	2,4209
3	2,2681
4	2,1543
5	2,0627
6	1,9854
7	1,9181
8	1,8583
9	1,8043
10	1,7550
15	1,5544
20	1,3998
25	1,2711
30	1,1590
35	1,0583
40	0,9659
50	0,7979
60	0,6439
70	0,4967
80	0,3500
90	0,1950
100	0,0000

Por lo tanto, el primer gráfico de la línea superior corresponde a seleccionar el 5 % de los mejores animales y de acuerdo con nuestra tabla eso equivale a una intensidad de selección de $i = 2{,}0627$. De la misma forma, para $p = 0{,}25$ (es decir, 25 %) la intensidad será de $i = 1{,}2711$, para $p = 0{,}40$ de $i = 0{,}9659$, para $p = 0{,}50$ de $i = 0{,}7979$, para $p = 0{,}60$ de $i = 0{,}6439$, para $p = 0{,}80$ de $i = 0{,}3500$ y finalmente $p = 0{,}90$ será de $i = 0{,}1950$.

PARA RECORDAR

- Cuando estamos trabajando en seleccionar animales a partir de un umbral, es muy importante **no confundir el umbral a partir del cual los animales son seleccionados con la media de los seleccionados.**
- El **diferencial de selección** (S) es la diferencia entre la media de los individuos seleccionados como progenitores de la siguiente generación respecto a la media de la población contemporánea; al tratarse de una resta de medias de la característica, tendrá las mismas unidades que la característica.
- El **umbral de selección** es un valor fijo a partir del cual elegimos a los individuos que serán reproductores de la siguiente generación, es decir, aquellos individuos que tengan un fenotipo igual o más favorable que el valor umbral.
- Recordando la **ecuación del criador** ($R = Sh^2$), cuando la heredabilidad es alta, la respuesta a la selección basada en el fenotipo individual será muy eficiente y la media de la nueva generación será cercana a la media de los animales seleccionados como reproductores. Por otro lado, cuando la heredabilidad es muy baja, la respuesta a la selección será pobre y la media de la nueva generación será casi igual a la de la generación anterior.
- Como en la ecuación del criador la **respuesta a la selección es directamente proporcional al diferencial de selección y a la heredabilidad**, incrementos en cualquiera de las dos llevan a incrementar la respuesta a la selección. El significado de ambas es conceptualmente distinto: mientras que el diferencial de selección se refiere a una decisión que podemos adoptar casi a nuestra voluntad en una población reducida, la heredabilidad se trata de un parámetro genético poblacional que requiere de mucho esfuerzo para ser modificado en el sentido deseado.

11.3 Diferencial de selección e intensidad de selección

Como vimos antes (figura 11.5), el diferencial de selección para un determinado umbral de selección es una función de la varianza de la característica. Esto nos deja con una medida de la **presión de selección** que es dependiente de los parámetros de la distribución fenotípica de la población de interés. Es decir, el significado de un determinado umbral de selección es completamente dependiente de la distribución fenotípica de la característica en la población en la que se piensa aplicar.

Por otra parte, cuando queremos comparar la presión de selección en dos características diferentes, el diferencial de selección deja de tener sentido: ¿cómo podría comparar la presión de selección en peso sucio de vellón (PVS, en kg) con la presión de selección en diámetro de fibras (DF, μm)? ¿Es acaso más importante un esfuerzo de 100 gramos en PVS que $2\mu m$ en DF? Claramente, resulta imposible de comparar cuando ambas características tienen distintas unidades. Aun teniendo las mismas unidades, si se trata de características distintas es

esperable que las distribuciones sean diferentes y por lo tanto el diferencial de selección sea incomparable en términos del esfuerzo de selección o presión de selección que implican.

Todo lo anterior sugiere que para poder comparar la presión de selección de una manera razonable es necesario estandarizar de alguna forma el diferencial de selección. Como vimos antes, una de las dificultades para comparar esta presión usando el diferencial de selección era la dependencia de este último en la varianza de la característica: a un determinado valor del diferencial de selección, cuanto menor era la varianza de la característica mayor era el esfuerzo que implicaba ya que debía quedarme con menos individuos como reproductores (pese a que el diferencial de selección podía ser menor en valor absoluto). Esto nos sugiere que a una presión de selección determinada hay una relación inversa entre el diferencial de selección y la dispersión de la distribución. Por otro lado, el desvío estándar de una característica posee las mismas unidades que la característica. Por lo tanto, dividiendo el diferencial de selección entre el desvío estándar fenotípico de la misma estamos corrigiendo los dos problemas en forma simultánea, es decir

$$i = \frac{S}{\sigma_P} \tag{11.25}$$

La **intensidad de selección** es una medida de la presión de selección que es independiente de las unidades de la característica y que **es igual al diferencial de selección estandarizado**, es decir, el diferencial de selección dividido entre el desvío estándar fenotípico de la característica.

Claramente, tanto el diferencial de selección como el desvío estándar fenotípico tienen las unidades de las características, por lo que al dividir ambas cantidades el resultado es adimensional (es decir, no tiene unidades). Por otra parte, cuanto mayor la varianza de una característica, mayor es su desvío estándar (que es la raíz cuadrada de la varianza) y dado un valor fijo de umbral una mayor parte de la población superará ese umbral (aumentando el diferencial de selección). Sin embargo, ese efecto se ve contrarrestado en la **intensidad de selección** ya que el diferencial de selección se divide entre el desvío estándar fenotípico, lo que compensa exactamente el incremento del diferencial de selección.

Normalmente, en especies de mamíferos domésticos de interés productivo (bovinos, ovinos, equinos, caprinos, porcinos, etcétera) es posible mantener una presión de selección mucho más alta en machos que en hembras. Esto se debe a que las hembras deben hacerse cargo de la preñez, quedando «bloqueadas» durante todo ese período a los fines de nuevos eventos reproductivos. Es decir, una vez preñada la hembra deberá continuar con la gestación hasta que nazcan las crías y recién a partir de ese momento (en general algo después) estará disponible para volver a ser gestante. En cambio, para los machos la preñez de la hembra no significa ninguna restricción a su capacidad de fecundar los gametos de otras hembras. Claramente, el factor limitante de la parte reproductiva de la mayor parte de los sistemas de producción animal es la capacidad de alojar a las hembras gestantes, ya que los machos, luego de la fecundación, dejan de tener importancia para el nacimiento de las crías. Estas

diferencias nos llevan a que un macho pueda aparearse con muchas hembras y por lo tanto necesitaremos muchos menos machos que hembras.

En términos de la proporción de animales seleccionados, lo anterior puede expresarse de la siguiente manera. Usualmente, la proporción de hembras que debo retener es la relación entre el número de hembras que tengo en la categoría de entrada a la población (rodeo, majada, piara, etcétera) y el número de las mismas que generé como reemplazos. El número de hembras que tengo en la categoría de entrada es función de la estructura de edades de la población y de la capacidad de animales que tiene el establecimiento, mientras que el número de reemplazos generados es función del número de hembras reproductivas en la población y de la eficiencia reproductiva de las mismas. Por otro lado, la proporción de machos retenidos como reproductores (en el caso de que los mismos salgan del establecimiento) suele obtenerse sobre una base mucho más sencilla: el número de machos requeridos es una función del número de hembras que un solo macho puede fecundar por estación de cría, mientras que el número de reemplazos generados es (en general) el mismo que el de hembras generadas.

Todo lo anterior nos lleva a la necesidad de poder manejar de forma integrada las diferencias en intensidad de selección que corresponde a machos y hembras. Como el aporte genético a la nueva generación vendrá la mitad por el lado de machos y la mitad por el lado de hembras, la intensidad de selección a nivel de la población será el promedio de la intensidad de selección en machos y en hembras.

$$i = \frac{i_M + i_H}{2} \tag{11.26}$$

En el caso de que en ambos sexos la varianza fuese diferente, la ecuación anterior implica

$$i = \frac{i_M + i_H}{2} = \frac{\frac{S_M}{\sigma_{P_M}} + \frac{S_H}{\sigma_{P_H}}}{2} = \frac{1}{2}\left[\frac{S_M}{\sigma_{P_M}} + \frac{S_H}{\sigma_{P_H}}\right] \tag{11.27}$$

Es decir, cada diferencial de selección es dividido entre su correspondiente desvío estándar fenotípico y luego ambos promediados.

Por otra parte, **si la característica tiene el mismo desvío estándar en machos que en hembras**, es decir $\sigma_{P_M} = \sigma_{P_H} = \sigma_P$, entonces sustituyendo la definición de la ecuación (11.25) en la ecuación (11.26) tenemos ahora

$$i = \frac{i_M + i_H}{2} = \frac{\frac{S_M}{\sigma_P} + \frac{S_H}{\sigma_P}}{2} = \frac{\frac{S_M + S_H}{\sigma_P}}{2} = \frac{S_M + S_H}{2\sigma_P} = \frac{\frac{S_M + S_H}{2}}{\sigma_P} = \frac{\bar{S}}{\sigma_P} \tag{11.28}$$

PARA RECORDAR

- Cuando queremos comparar la presión de selección en dos característica diferentes, el **diferencial de selección deja de tener sentido**, por lo que debemos recurrir a una variable adimensional que llamaremos **intensidad de selección.**
- La **intensidad de selección** es una medida de la presión de selección que es **independiente de las unidades de la característica** y que es igual al diferencial de selección estandarizado, es decir, el diferencial de selección dividido entre el desvío estándar fenotípico de la característica ($i = \frac{S}{\sigma_P}$).

- Cuando queremos comparar la presión de selección en dos característica diferentes, el **diferencial de selección deja de tener sentido**, por lo que debemos recurrir a una variable adimensional que llamaremos **intensidad de selección.**
- La **intensidad de selección** es una medida de la presión de selección que es **independiente de las unidades de la característica** y que es igual al diferencial de selección estandarizado, es decir, el diferencial de selección dividido entre el desvío estándar fenotípico de la característica ($i = \frac{S}{\sigma_P}$).

11.4 Intensidad de selección y proporción seleccionada

La definición de intensidad de selección que hemos dado, es decir, el diferencial de selección estandarizado, nos deja con una medida de presión de selección que es sencilla de obtener si conocemos tanto el diferencial de selección como el desvío fenotípico de la característica. Sin embargo, en muchos casos no podemos darnos el lujo de determinar un valor dado de umbral para la característica ya que podría comprometer la sostenibilidad del sistema o implicar una presión de selección mucho menor que la posible y deseable. Por ejemplo, si fijamos un umbral tan alto que restrinja el número de hembras por debajo de las necesarias para generar reemplazos de la primera categoría en el rodeo, entonces comenzaremos un proceso de reducción en el tamaño del mismo, a menos que compensemos de otra forma. Esto nos lleva a la posibilidad de que el determinante para la presión de selección posible sea la proporción de individuos que debo quedarme como reproductores.

En general, para distribuciones arbitrarias esto implicaría obtener la media de los individuos que quedan en la proporción seleccionada, lo que puede resultar algo complejo de derivar. Para distribuciones conocidas una alternativa adecuada a los tiempos es recurrir a algún software que permita hacer los cálculos basados en las distribuciones o aún realizar simulaciones (R Core Team 2021) y obtener una buena estimación de esta media y del desvío estándar de la característica. Claramente, esto implica conocer la distribución, así como un estimado razonable de sus parámetros. Afortunadamente, la mayor parte de las características productivas siguen una distribución normal, con media μ y varianza σ_P^2, es decir $N(\mu,\sigma^2)$. El diferencial de selección es igual a la media de los seleccionados menos la media de la población, por lo que al restarle la media de la población a los todos los datos los mismos quedan centrados en cero (ya que si $\mu = x_0$, entonces la nueva media al restarle x_0 será $\mu = x_0 - x_0 = 0$). Esta operación solo desplaza la distribución centrándola en $\mu = 0$, es decir, cada valor será un desvío de la media, pero no cambia su varianza. Sin embargo, en la definición de intensidad de selección que vimos, i $= \frac{S}{\sigma_P}$. Ahora, si a esta distribución centrada en cero la dividimos entre el desvío estándar fenotípico, es decir σ_P, tenemos que los datos tienen ahora distribución $N(0,1)$.

Dicho de otra forma, haciendo las operaciones clásicas que haríamos para obtener la intensidad de selección a partir del diferencial de selección, lo que estamos haciendo es obtener el apartamiento para la media de los individuos seleccionados en una distribución normal estandarizada, es decir $N(0,1)$. Esto que puede parecer trivial es de suma importancia ya que la distribución normal estandarizada nos permite obtener varios resultados importantes sin tener que hacer cuentas específicas para cada combinación de x_0 y σ_P^2.

Como muestra la figura 11.7, **en una distribución normal estandarizada la media de los individuos seleccionados es exactamente la intensidad de selección.** Es decir, para estandarizar una distribución normal $N(x_0, \sigma_P^2)$ primero le restamos x_0 a todas las observaciones y luego dividimos esto entre el desvío estándar de la distribución (σ_P). La primera operación deja la media de los individuos seleccionados exactamente igual al diferencial de selección ya que $\mathrm{x_S} - \mathrm{x_0} = \mathrm{S}$, mientras que la segunda transforma el diferencial de selección en la intensidad de selección ya que $\frac{\mathrm{S}}{\sigma_P} = \mathrm{i}$. Ahora, basados en las propiedades de la distribución normal estándar, la media de la región seleccionada es igual a la altura del punto de truncamiento dividido entre el área bajo la curva a partir del límite de truncamiento y hasta $+\infty$.

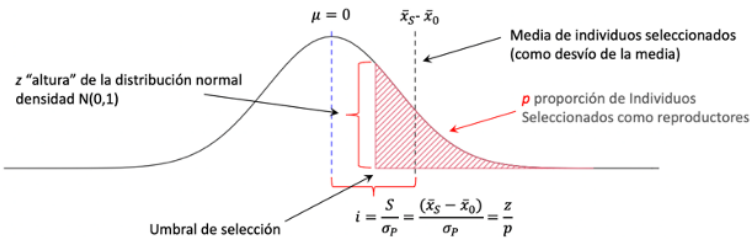


Figura 11.7. La intensidad de selección en una distribución normal estandarizada. Al haber estandarizado la distribución fenotípica, es decir, restado a los valores la media original y luego dividido entre el desvío estándar, ahora la media de los individuos seleccionados será directamente i, ya que $\frac{(\bar{\mathrm{x}}_\mathrm{s} - \bar{\mathrm{x}}_0)}{\sigma_P} = \frac{\mathrm{S}}{\sigma_P} = \mathrm{i}$. Esto nos permite explotar la relación existente en la distribución normal de la media de la región seleccionada como la relación entre la altura en el punto de truncamiento z (que es la densidad de la normal $N(0,1)$ en ese punto) y la proporción seleccionada p (que es el área bajo la curva entre el punto de truncamiento y $+\infty$), es decir $\mathrm{i} = \frac{\mathrm{z}}{\mathrm{p}}$.

El resultado obtenido, $\mathrm{i} = \mathrm{z/p}$, es muy importante porque nos dice que, a una proporción determinada de individuos seleccionados en una característica con distribución normal, cualquiera sea esta, le corresponde siempre la misma intensidad de selección. Además, como esta intensidad de selección es siempre la misma para una proporción dada es posible construir una tabla, como la que sigue, con los pares de valores a la precisión deseada. Más aún, es trivial calcular el valor exacto para cualquier proporción en la mayoría de las hojas de cálculo o en software estadístico como R Core Team (2021).

Tabla 11.2. Intensidad de selección en función de la proporción de individuos seleccionados como reproductores en una característica con distribución normal

Proporción seleccionada (%)	Intensidad		Proporción seleccionada (%)	Intensidad
100	0		50	0,7979
99	0,0269		49	0,8139
98	0,0494		48	0,8301
97	0,0701		47	0,8464
96	0,0898		46	0,8629
95	0,1086		45	0,8796
94	0,1267		44	0,8964
93	0,1444		43	0,9135
92	0,1616		42	0,9307
91	0,1785		41	0,9482
90	0,195		40	0,9659
89	0,2113		39	0,9838
88	0,2273		38	1,002
87	0,2432		37	1,0205
86	0,2588		36	1,0392
85	0,2743		35	1,0583
84	0,2897		34	1,0777
83	0,3049		33	1,0974
82	0,32		32	1,1175
81	0,335		31	1,138
80	0,35		30	1,159
79	0,3648		29	1,1804
78	0,3796		28	1,2022
77	0,3943		27	1,2246
76	0,409		26	1,2476
75	0,4237		25	1,2711
74	0,4383		24	1,2953
73	0,4529		23	1,3202
72	0,4675		22	1,3459
71	0,4821		21	1,3724
70	0,4967		20	1,3998
69	0,5113		19	1,4282
68	0,5259		18	1,4578
67	0,5405		17	1,4886
66	0,5552		16	1,5207
65	0,5698		15	1,5544
64	0,5846		14	1,5898
63	0,5993		13	1,6273
62	0,6141		12	1,667
61	0,629		11	1,7094
60	0,6439		10	1,755
59	0,6589		9	1,8043
58	0,674		8	1,8583
57	0,6891		7	1,9181
56	0,7043		6	1,9854
55	0,7196		5	2,0627
54	0,7351		4	2,1543
53	0,7506		3	2,2681
52	0,7662		2	2,4209
51	0,782		1	2,6652

En la figura 11.8 se muestra la curva de intensidad de selección en función de la proporción de individuos seleccionados. A la izquierda se grafica esta relación para todo el intervalo posible de proporciones (es decir entre cero y uno), expresado como porcentaje y en la misma se aprecia el carácter aproximadamente sigmoidal de la relación. Claramente, a medida que nos vamos acercando hacia el cero la pendiente se hace más empinada. Por esta razón, a la izquierda se grafica solamente la región entre 0 % y 10 %, una zona que es usual en

machos de mamíferos domésticos. Por ejemplo, si seleccionamos el 2 % de los machos como reproductores, tendremos entonces (de acuerdo con la gráfica) aproximadamente una intensidad de selección igual a 2,4; como referencia, el valor «exacto» usando la expresión «dnorm(qnorm(0,98))/0,02» en R Core Team (2021) es de 2,420907.

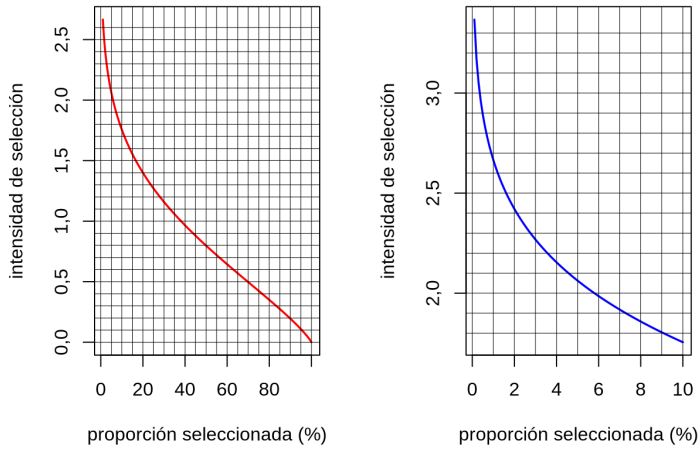


Figura 11.8. Intensidad de selección en función de la proporción de individuos seleccionados como reproductores en una característica con distribución normal. Mientras que la gráfica de la derecha representa todo el intervalo posible de selección, la de la derecha representa la región de más intensidad de selección, típica de machos en la mayor parte de las especies de mamíferos domésticos.

Un tema relativamente menor en la práctica, pero relevante en algunos casos es el del tamaño poblacional y su efecto en el cálculo de la intensidad de selección. La obtención de la intensidad de selección a partir de las propiedades de la distribución normal asume precisamente eso, una distribución normal de la característica. En particular, con muy pocos individuos, el efecto de seleccionar una determinada proporción nunca será continuo, como sí lo asume la distribución normal. Por ejemplo, si de 10 individuos voy a seleccionar solo 1, la proporción de elegidos es $p = 1/10 = 0{,}1$. Por otro lado, si tengo que elegir 100 individuos de una población que tiene 1000 individuos, también la proporción será de $p = 100/1000 = 0{,}100$. Sin embargo, mientras que la «resolución» de este último caso es de $1/1000$ (si hubiese tenido que elegir 1 individuo más la proporción sería de $p = 101/1000 = 0{,}101$), en el primer caso la «resolución» es de $1/10$ (si hubiese elegido un individuo más sería $p = 2/10 = 0{,}2$). Claramente, el **tamaño de los pasos** en la discontinuidad es de $1/n$. Para corregir este efecto, Bulmer (1980) propone una «corrección» de la proporción a utilizar en la distribución normal (que llamaremos p*) y que para seleccionar k individuos entre n individuos disponibles es igual a

$$p^* = \frac{\left(k + \frac{1}{2}\right)}{\left(n + \frac{k}{2n}\right)} \tag{11.29}$$

En la figura 11.9 se puede apreciar el efecto del número total de individuos de la población en la corrección necesaria para la proporción de individuos seleccionados. Cada una de las

curvas de colores representa una proporción diferente de individuos seleccionados, desde 10 % (curva en rojo) hasta 20 % (curva en azul). Mientras que a partir de los 100 individuos las diferencias entre la proporción realmente seleccionada (p) y el equivalente efectivo de proporción (p^*) a considerar para calcular la intensidad de selección es insignificante, a valores bajos del número de individuos (n) puede ser bastante importante. Por ejemplo, para $k = 1$ y $n = 10$, mientras que $p = 0{,}10$, $p^* \approx 0{,}1492$. Sin embargo, para la misma proporción en $n = 100$ individuos (por lo tanto, $k = 10$), tenemos que $p^* \approx 0{,}1049$. El comportamiento asintótico se puede apreciar claramente en la figura ya que a medida que aumenta el n las curvas de colores tienden a las líneas punteadas que les «corresponden», es decir $\lim_{n\to\infty} p^* = p$.

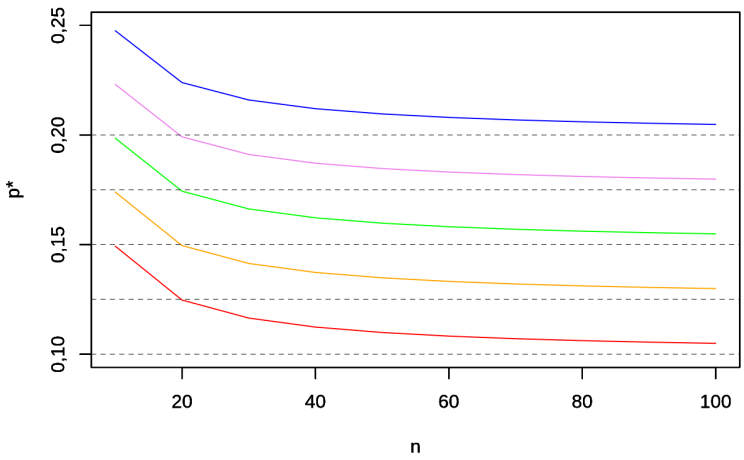


Figura 11.9. Ajuste de la proporción de individuos seleccionados para un número bajo de individuos. De abajo hacia arriba, las curvas representan proporciones reales de 0,100, 0,125, 0,150, 0,175 y 0,200, como lo indican las líneas a trazos correspondientes. A medida que el número total de individuos crece la diferencia entre la proporción real *p* (línea a trazos) y la proporción ajustada p^* (curvas de color) tiende a cero.

Hasta ahora hemos solo considerado el caso de la selección contra determinados individuos (aquellos que quedan por debajo del umbral), pero no necesariamente este tipo de selección es óptimo a largo plazo. Si la presión de selección es muy alta, en la medida en que elegimos los «mejores individuos» basados en nuestra característica de interés, estamos reduciendo en forma sustancial la variabilidad genética disponible, lo que puede atentar contra el futuro genético a mediano o largo plazo. Una alternativa más general es utilizar una función continua (por ejemplo) de nuestra característica de interés para determinar la probabilidad de que el individuo quede entre los reproductores. Esto es lo mismo que establecer el **fitness del individuo** como **función de la característica**. Con esto en mente, ahora existe una probabilidad de permanecer como reproductor para cada valor fenotípico estandarizado y que será el «peso» que tendrá cada fenotipo en el cálculo de i. Es decir, la intensidad de selección será la media ponderada (por la función **fitness**) de los fenotipos estandarizados.

¿De dónde sale la relación $i = \frac{z}{p}$?

Una de las relaciones más usadas para conocer la intensidad de selección en mejoramiento genético animal es la que establece $i = \frac{z}{p}$, es decir, en selección por truncamiento, la intensidad es la relación de la ordenada de una distribución $\sim N(0,1)$ en el punto de truncamiento, dividido la proporción de animales seleccionados (ver figura 11.7). ¿Pero de dónde sale esta relación?

Para entenderla, recordemos que la intensidad de selección era igual al diferencial de selección dividido entre el desvío estándar fenotípico, lo que nos referirá a una distribución $\sim N(0,1)$ y por lo tanto la intensidad de selección será la media de la fracción seleccionada a partir del punto de truncamiento. En el caso de la distribución normal, la función de densidad acumulada es[3]

$$\Phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{(u-\mu)^2}{2\sigma^2}} du \tag{11.30}$$

por lo que la proporción p de individuos seleccionados a partir de un valor X_u en una $\sim N(0,1)$ será

$$p = \frac{1}{\sqrt{2\pi}} \int_{X_u}^{\infty} e^{-\frac{u^2}{2}} du = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X_u} e^{-\frac{u^2}{2}} du = 1 - \Phi_{0,1}(X_u) \tag{11.31}$$

Pero la media es la integral en el intervalo de interés del producto de x por la densidad, que ahora será la densidad original de la normal, pero dividida por la integral de la misma en el intervalo (la proporción de individuos con los que nos quedamos), para que integre a 1. Calculemos la primitiva del numerador

$$\frac{1}{\sqrt{2\pi}} \int x e^{-\frac{x^2}{2}} dx. \text{ Ahora, si } u = -\frac{x^2}{2} \Rightarrow \frac{du}{dx} = -x \therefore dx = -\frac{1}{x} du \tag{11.32}$$

$$\frac{1}{\sqrt{2\pi}} \int x e^{-\frac{x^2}{2}} dx = -\frac{1}{\sqrt{2\pi}} \int x e^{u} \frac{du}{x} = -\frac{1}{\sqrt{2\pi}} \int e^{u} du = \tag{11.33}$$

$$-\frac{e^{u}}{\sqrt{2\pi}} = -\frac{e^{\frac{x^2}{2}}}{\sqrt{2\pi}} \tag{11.34}$$

Ahora, de acuerdo con la definición que dimos recién, si dividimos este último resultado (evaluado en el intervalo de interés) entre la proporción de individuos, obtenemos

$$i = \frac{\frac{1}{\sqrt{2\pi}} \int_{X_u}^{\infty} u e^{-\frac{u^2}{2}} du}{\frac{1}{\sqrt{2\pi}} \int_{X_u}^{\infty} e^{-\frac{u^2}{2}} du} = \frac{-\frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}\Big|_{u=X_u}^{\infty}}{\frac{\operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)}{2}\Big|_{u=X_u}^{\infty}} = \frac{\frac{e^{-\frac{X_u^2}{2}}}{\sqrt{2\pi}}}{\frac{\operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)}{2}\Big|_{u=X_u}^{\infty}} = \frac{z}{p} \tag{11.35}$$

[3] La igualdad se sostiene debido a que se cumple que $\Phi(x) = 1 - \Phi(-x)$, hecho que la persona interesada puede consultar en libros de texto que traten estadística básica.

En las ecuaciones anteriores $erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$ es la **función de error** (o **función de error de Gauss**).

Resumiendo, la intensidad de selección es la ordenada de la función de densidad $\sim N(0,1)$ en el punto de truncamiento, dividida entre la proporción de la distribución que es mayor a ese punto de truncamiento (la proporción de individuos seleccionados).

La relación entre diferentes formas de calcular la intensidad de selección

Vimos antes dos formas diferentes de calcular la intensidad de selección, $i = \frac{S}{\sigma_P}$ y $i = \frac{z}{p}$. La pregunta que nos hacemos entonces es si estas definiciones son exactamente equivalentes. Recordemos que, desde el punto de vista conceptual, la intensidad de selección es la media de los individuos seleccionados, pero expresada en relación al desvío estándar, o sea, independiente de las unidades con que se mide la característica. Ahora, la primera de las dos ecuaciones claramente se ajusta a esta definición en palabras, que no asume ninguna forma de la distribución de valores de la característica, al menos en forma explícita. En cambio, la segunda ecuación invoca explícitamente una distribución $\sim N(0,1)$.

La definición de intensidad de selección como el diferencial de selección estandarizado es simple y precisa, pero requiere del conocimiento del diferencial de selección, lo que requiere usualmente del conocimiento de las medias de la población y de los individuos seleccionados, además del desvío estándar fenotípico de la característica en la población. En cambio, la definición de la intensidad de selección como la relación $i = z/p$ es apenas una definición operacional que se cumple bajo el supuesto de normalidad. Por el contrario, lo único que es necesario conocer es la proporción de animales que se seleccionarán, lo que usualmente resulta trivial de calcular para estructuras poblacionales típicas de la producción animal. En resumen, el uso de una u otra dependerá de la información disponible y de que se cumpla o no el supuesto de normalidad para la característica de interés.

Ejemplo 11.4

En la siguiente figura se presentan diferentes situaciones de diferencial e intensidad de selección representados gráficamente, suponiendo características de distribución normal:

PARA RECORDAR

- En muchos casos no podemos darnos el lujo de determinar un valor dado de umbral para la característica ya que podría comprometer la sostenibilidad del sistema o implicar una presión de selección mucho menor que la posible y deseable. Esto nos lleva a la posibilidad de que el determinante para la presión de selección posible sea la **proporción de individuos que debo quedarme como reproductores**.
- La proporción $i = z/p$ nos indica que a una **proporción determinada de individuos seleccionados** en una **característica con distribución normal**, cualquiera sea esta, le **corresponde siempre la misma intensidad de selección**. Además, como esta intensidad de selección es siempre la misma para

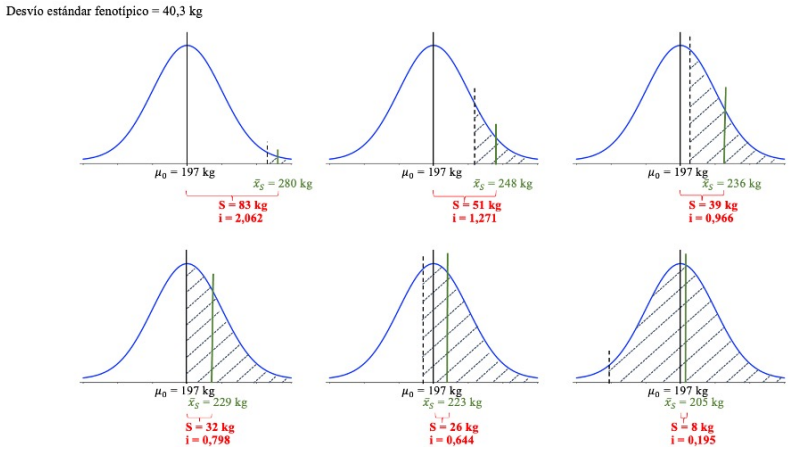


Figura 11.10. Diferencial e intensidad de selección representados gráficamente, suponiendo características de distribución normal

una proporción dada **es posible construir una tabla con los pares de valores a la precisión deseada.**

- Si la **presión de selección es muy alta**, en la medida en que elegimos los «mejores individuos» basados en nuestra característica de interés, **estamos reduciendo en forma sustancial la variabilidad genética disponible**, lo que **puede atentar contra el futuro genético a mediano o largo plazo.** Una alternativa más general es utilizar una función continua de nuestra característica de interés para determinar la probabilidad de que el individuo quede entre los reproductores. Esto **es lo mismo que establecer el fitness del individuo como función de la característica.**
- Respecto a las dos formas diferentes de calcular la intensidad de selección que mencionamos ($i = S/\sigma P$ y $i = z/p$), el uso de una u otra dependerá de la información disponible y de que se cumpla o no el supuesto de normalidad para la característica de interés.

11.5 Intervalo generacional

En general, el progreso genético medido en términos de generaciones nos permite desarrollar los conceptos principales sobre los mecanismos y procesos que intervienen en el mejoramiento genético o en el estudio de las variaciones en las frecuencias génicas. Sin embargo, su utilidad resulta reducida cuando tenemos que pensar su significado en los términos a los que estamos acostumbrados a referenciarnos los seres humanos. Por ejemplo, un productor rural debe pagar los impuestos todos los años, un productor ganadero debe entorar todos los años y un productor lechero idealmente debe elegir dosis de semen cada año para las vacas en producción. Es decir, solemos manejarnos con nuestro marco de referencia temporal a la hora de decidir y eso requiere que podamos expresar el progreso genético en dicho marco. Más aún, como veremos más adelante, diferentes estrategias de selección suelen implicar diferentes intervalos de tiempo para observar el progreso genético,

por lo que debemos poder «ajustar» nuestros resultados para tener en consideración estos efectos.

El ciclo de vida de diferentes organismos implica que cuando una generación alcanza la madurez para reproducirse, la anterior ya desapareció o al menos no se encuentra más en su fase reproductiva. Ejemplo de esto serían las plantas anuales, especies de peces anuales y varios otros organismos. En estos casos, pasar del progreso generacional al anual es muy sencillo: alcanza con dividir el progreso genético generacional (es decir, expresado en las unidades de la característica que avanzamos **por generación**) entre el intervalo (en años) entre dos generaciones (años/generación), es decir

$$R_a = \frac{R}{IG} \tag{11.36}$$

Las unidades de R_a serán entonces las de la característica pero **por año** (o sea, divididas entre año). En el caso de plantas anuales, el intervalo de tiempo entre generaciones será igual a 1 año, por lo que los valores de R_a y R serán iguales, aunque expresados en distintas unidades. En animales, un ejemplo de especies anuales son las *Cynolebias*, peces que pertenecen a la familia Rivulidae y al orden Ciprinodontiformes. Estos peces habitan en charcos o bañados y su distribución incluye Argentina, Brasil, Paraguay y Uruguay. El ciclo de vida de estos organismos comienza como huevos que son puestos por la generación parental en primavera, los que son enterrados en el fondo barroso. Al llegar el verano el calor hace que los charcos se sequen y los adultos mueran, mientras que los huevos continúan incubándose en el suelo. Con las primeras lluvias de abril y mayo los peces nacen y se desarrollan rápidamente hasta la primavera, donde ya maduros sexualmente comenzarán una nueva puesta.

Hasta acá resulta inmediata la relación entre progreso genético generacional y progreso genético anual, ya que el intervalo generacional es trivial de calcular, por lo que también la aplicación de la ecuación (11.36) es trivial. Sin embargo, en poblaciones en las que las generaciones se superponen resulta mucho menos intuitivo cuál es el intervalo generacional ya que en cada momento del tiempo hay individuos de diferentes generaciones reproduciéndose. Esto nos lleva a la necesidad de tener una definición operativa del intervalo generacional.

Llamamos intervalo generacional a la edad promedio de los padres cuando nacen sus hijos. Acá es importante tener en cuenta que promedio se refiere tanto al promedio entre los dos sexos, como al promedio dentro de cada progenitor (ya que cada progenitor puede tener más de un hijo en distintos momentos).

Esto nos deja con una definición muy clara pero que en general no será trivial de calcular, excepto en situaciones bastante estilizadas o simplificadas.

Lo primero que se deriva de nuestra definición es que el intervalo generacional será el promedio de los intervalos generacionales de ambos sexos, es decir

$$IG = \frac{IG_m + IG_h}{2} = \frac{IG_m}{2} + \frac{IG_h}{2} \tag{11.37}$$

Si bien el número de machos no es, por lo general, igual al de hembras, los aportes genéticos de ambos sexos son similares (desde el punto de vista del genoma nuclear) ya que los hijos reciben la mitad de sus genomas de cada progenitor. Por otra parte, el promedio de las edades a las que nacen los hijos deberá tener en cuenta el aporte relativo en cada categoría de edad. Por ejemplo, en seres humanos, es posible encontrar madres que tuvieron descendencia a la edad de 55 años, pero claramente son casos bastante excepcionales y este aporte es proporcionalmente mucho menor que el de las madres de 25 años de edad. Por lo tanto, para cada sexo, si j representa una edad determinada, la edad promedio será

$$IG_m = \frac{1}{n_m} \sum_j n_{j_m} \times j \tag{11.38}$$

$$IG_h = \frac{1}{n_h} \sum_j n_{j_h} \times j \tag{11.39}$$

con

$$\sum_j n_{j_m} = n_m \tag{11.40}$$

$$\sum_j n_{j_h} = n_h \tag{11.41}$$

el número total de machos y hembras, respectivamente. Es decir, el intervalo generacional será el promedio de edades de los padres al nacer los hijos, ponderado el promedio por la proporción de hijos en cada una de las edades.

Ejemplo 11.5

Un rodeo vacuno que produce sus propios reemplazos posee las siguientes cantidades de machos y hembras en las distintas categorías de edades (en años, al nacimiento de los hijos)

Edad en años (al nacimiento de los hijos)	Hembras	Machos
3	250	15
4	220	10
5	200	5
6	170	–
7	160	–
TOTAL	1000	30

Calcular el intervalo generacional de machos, el de hembras y el intervalo generacional promedio.

En machos hay solo 3 categorías de edad, por lo que el intervalo generacional de machos sería, aplicando las ecuaciones (11.39) y (11.41), tenemos que

$$n_m = \sum_j n_{j_m} = 15 + 10 + 5 = 30 \tag{11.42}$$

y

$$IG_m = \frac{1}{n_m} \sum_j n_{j_m} \times j = \frac{3 \times 15 + 4 \times 10 + 5 \times 5}{30} = \frac{110}{30} = 3{,}6667 \text{ años} \tag{11.43}$$

En hembras, usando el mismo razonamiento y las mismas ecuaciones, tenemos

$$n_h = \sum_j n_{j_h} = 250 + 220 + 200 + 170 + 160 = 1000 \tag{11.44}$$

y

$$\begin{aligned} IG_h &= \frac{1}{n_h} \sum_j n_{j_h} \times j \\ &= \frac{3 \times 250 + 4 \times 220 + 5 \times 200 + 6 \times 170 + 7 \times 160}{1000} \\ &= \frac{4770}{1000} = 4{,}77 \text{ años} \end{aligned} \tag{11.45}$$

Por lo tanto, ahora aplicando la ecuación (11.37), tenemos que el intervalo generacional del rodeo será igual a

$$IG = \frac{IG_m + IG_h}{2} = \frac{3{,}6667 + 4{,}77}{2} = \frac{8{,}4367}{2} \sim 4{,}22 \text{ años} \tag{11.46}$$

Es importante destacar que **la estructura de edades** en el rodeo, majada o población **que nos importa es la de los individuos activos en términos reproductivos** (y que sean exitosos, es decir que generan descendencia). En general, para que una estructura de edades sea estable en una población es necesario que se cumplan algunas cosas. Por ejemplo, si consideramos a todos los individuos de una **población cerrada a los ingresos**, cada categoría de edad debe tener el mismo número o menos de individuos que la edad que la antecede, ya que de lo contrario los individuos aparecerían mágicamente o no se trataría de una estructura de edades estabilizada. La estructura de edades de los individuos que se reproducen sí puede ser arbitraria, con más individuos reproductivos en edades más avanzadas, por ejemplo, si impedimos que los más jóvenes se reproduzcan o aún si los más jóvenes tienen menor éxito reproductivo que los mayores (y lo tenemos en cuenta en el modelo). Por lo general, para simplificar, usaremos ejemplos y ejercicios que muestran el número total de individuos en cada categoría de edad, asumiendo que **refugamos** a los individuos que no se reproducen o que directamente no están (por muerte, por ejemplo).

PARA RECORDAR

- En el caso del ciclo de vida de algunos organismos, cuando una generación alcanza la madurez para reproducirse, la anterior ya desapareció o al menos no se encuentra más en su fase reproductiva. En estos casos, pasar del progreso generacional al anual es muy sencillo: alcanza con dividir el progreso genético generacional entre el intervalo (en años) entre dos generaciones (años/generación), es decir $R_a = \frac{R}{IG}$
- En poblaciones en las que las generaciones se superponen resulta mucho menos intuitivo cuál es el intervalo generacional ya que en cada momento del tiempo hay individuos de diferentes generaciones reproduciéndose.
- Llamamos **intervalo generacional a la edad promedio de los padres cuando nacen sus hijos**. Es importante tener en cuenta que promedio se refiere tanto al promedio entre los dos sexos, como al promedio dentro de cada progenitor (ya que cada progenitor puede tener más de un hijo en distintos momentos).
- El intervalo generacional poblacional lo calcularemos como el intervalo generacional promedio entre los sexos: $IG = \frac{IG_m + IG_h}{2} = \frac{IG_m}{2} + \frac{IG_h}{2}$. Por su parte, para calcular el IG de cada sexo, si j representa una edad determinada, la edad promedio será $IG_m = \frac{1}{n_m} \sum_j n_{j_m} \times j$; $IG_h = \frac{1}{n_h} \sum_j n_{j_h} \times j$; con $\sum_j n_{j_m} = n_m$ y $\sum_j n_{j_h} = n_h$
- Debemos siempre recordar que la estructura de edades en el rodeo, majada o población que **nos importa es la de los individuos activos en términos reproductivos** (y que sean exitosos, es decir que generan descendencia).

11.6 Medidas de la respuesta

En este capítulo hemos manejado hasta ahora el progreso genético ocurrido en el transcurso de una generación y su relación con el diferencial de selección y la heredabilidad. Claramente, cuando presentamos la ecuación del criador como forma de estimar la «Heredabilidad lograda», es decir $\hat{h}^2 = \frac{R}{S}$, necesitamos tener buenas medidas de la respuesta a la selección y de diferencial de selección.

El diferencial de selección teórico es simplemente la diferencia entre la media de la población seleccionada y la media de la población de referencia, es decir $S = \mu_S - \mu_0$, por lo que su medición no presenta, en principio mayores dificultades. Alcanza con medir los individuos en ambos grupos, la población seleccionada y la población en general (o al menos un estimado razonable de esta última) y calcular la diferencia. Sin embargo, ese diferencial de selección teórico funcionaba en nuestras derivaciones si asumíamos que todos los padres seleccionados dejaban el mismo número de descendientes. En los casos en los que existen diferencias notorias en la descendencia que dejaron los padres seleccionados es necesario obtener una mejor estimación del **diferencial de selección efectivo**, lo que puede obtenerse ponderando la medida fenotípica de cada progenitor por el número de descendientes que aportaron a la nueva generación. Esto tiene en cuenta no solo el efecto de la fertilidad de los reproductores sino la viabilidad de los descendientes. Por ejemplo, si existen diferencias en la fertilidad o viabilidad asociadas a fenotipos más extremos, pese a que nuestro esfuerzo en la **selección artificial** es hacia los mismos, el efecto de la **selección natural** contrarrestará, de alguna manera, este esfuerzo. Un ejemplo de esto podría ser en la producción lechera, donde los fenotipos más productores (extremos) suelen tener peores resultados reproductivos (vinculado esto a la partición energética de base genética).

Por otra parte, como ya vimos en la sección dedicada a la «Heredabilidad lograda», el resultado del progreso genético observado en cada generación sucesiva puede ser diferente de los otros, esencialmente debido a los efectos del azar en la respuesta, sumados a las variaciones aleatorias de los efectos del diferencial de selección efectivo (por diferencias en el número de progenies dejadas debido al azar). Esto nos lleva a la necesidad de buscar algo de estabilidad, lo que podemos hacer a partir del promedio de varias generaciones, o del ajuste de una regresión de la respuesta acumulada en el diferencial de selección acumulado (como vimos previamente). Sin embargo, en estas estimaciones estamos pasando por alto un factor que puede ser determinante: la evolución del «ambiente». La mejora en la producción puede y suele venir de la genética, del «ambiente», o de ambos (y de sus interacciones). Cuando miramos la respuesta a la selección como solo de origen genético nos estamos olvidando de la posible existencia de mejoras en las condiciones de producción (alimentación, alojamiento, salud, manejo, etcétera), por lo que estaríamos sobrestimando la respuesta. Como $\hat{h}^2 = \frac{R}{S}$, una sobrestimación del numerador nos lleva a una sobrestimación de la heredabilidad también. Una alternativa para esto es el uso de datos en una población similar pero en la que no se realizó selección en el mismo período de tiempo. Eso nos permitiría tener un testigo y usar además esa información para descontarla del progreso obtenido en la población bajo selección.

11.7 Progreso genético generacional y anual

Como vimos previamente en la ecuación (11.36), el progreso genético anual es simplemente el resultado de dividir el progreso genético que se observa en una generación entre el intervalo de tiempo que le toma a una generación sustituir a la siguiente (en términos reproductivos), es decir, el **intervalo generacional**. En las ecuaciones (11.39) y (11.41) encontramos una forma de calcular el intervalo generacional cuando las generaciones se solapaban, mientras que previamente habíamos obtenido distintas formulaciones para el progreso genético generacional (por ejemplo, en la ecuación del criador $\mathrm{R} = \mathrm{S}\ \mathrm{h}^2$, ecuación (11.24)).

Por lo tanto, poniendo todo junto, tenemos una primera expresión algo más explícita del progreso genético anual

$$\mathrm{R_a} = \frac{\mathrm{R}}{\mathrm{IG}} = \frac{\mathrm{S\ h^2}}{\mathrm{IG}} \tag{11.47}$$

En ciertos casos, como vimos antes, es difícil obtener el diferencial de selección ya que implica conocer tanto la media de la población como la media de los individuos seleccionados. Sin embargo, para características fenotípicas con distribución normal, la intensidad de selección es mucho más fácil de obtener a partir de la proporción de individuos seleccionados. Recordando además que la relación entre diferencial de selección e intensidad de selección es $i = \frac{S}{\sigma_P} \therefore S = i \times \sigma_P$, sustituyendo en la ecuación (11.47) tenemos ahora

$$R_a = \frac{S\ h^2}{IG} = \frac{i\ \sigma_P\ h^2}{IG} \tag{11.48}$$

Ahora, como la intensidad de selección de la población es igual al promedio de machos y hembras y el intervalo generacional de la población también es igual al promedio de dicho intervalo en machos y hembras, poniendo todo junto tenemos que la ecuación (11.48) se transforma en

$$\mathrm{R_a} = \frac{\left(\frac{i_m + i_h}{2}\right)\ \mathrm{h^2}\ \sigma_P}{\left[\frac{\mathrm{IG_m + IG_h}}{2}\right]} \tag{11.49}$$

A su vez, $\mathrm{h} = \frac{\sigma_A}{\sigma_P}$, entonces, sustituyendo en el numerador de la ecuación (11.48) tenemos

$$\mathrm{R} = \mathrm{i\ h^2}\ \sigma_P = \mathrm{i\ h\ h}\ \sigma_P = \mathrm{i\ h}\ \frac{\sigma_A}{\sigma_P}\sigma_P\ \therefore \tag{11.50}$$

$$\mathrm{R} = \mathrm{i\ h}\ \sigma_A \tag{11.51}$$

lo que nos deja con el siguiente equivalente de la ecuación (11.49)

$$\mathrm{R_a} = \frac{\left(\frac{i_m + i_h}{2}\right)\ \mathrm{h}\ \sigma_A}{\left[\frac{\mathrm{IG_m + IG_h}}{2}\right]} \tag{11.52}$$

Recordando que h era la precisión en el caso de selección individual basada en el fenotipo y que también era la correlación entre el valor de cría y la fuente de información, $\mathrm{r_{AC}}$ (en este caso el fenotipo del individuo), entonces, podemos generalizar la ecuación (11.51) a

$$\mathrm{R} = \mathrm{i\ r_{AC}}\ \sigma_A \tag{11.53}$$

o llevar la ecuación (11.52) a una de sus formas más explícitas que usaremos en este capítulo

$$\mathrm{R_a} = \frac{\left(\frac{i_m + i_h}{2}\right)\ \mathrm{r_{AC}}\ \sigma_A}{\left[\frac{\mathrm{IG_m + IG_h}}{2}\right]} \tag{11.54}$$

PARA RECORDAR

- En los casos en los que existen **diferencias notorias en la descendencia que dejaron los padres seleccionados** es necesario obtener una mejor estimación del diferencial de selección efectivo, lo que puede obtenerse **ponderando la medida fenotípica de cada progenitor por el número de descendientes que aportaron a la nueva generación.**
- Por ejemplo, si existen diferencias en la fertilidad o viabilidad asociadas a fenotipos más extremos, **pese a que nuestro esfuerzo en la selección artificial** es hacia ellos, **el efecto de la selección natural contrarrestará**, de alguna manera, este esfuerzo.
- Cuando miramos la respuesta a la selección como solo de origen genético nos estamos olvidando de la posible existencia de mejoras en las condiciones de producción, por lo que estaríamos sobrestimando la respuesta. Una alternativa para esto es el uso de datos en una población similar pero en la que no se realizó selección en el mismo período de tiempo.
- Podemos expresar la respuesta anual (R_a) a la selección como

$$\mathrm{R_a} = \frac{\left(\frac{i_m + i_h}{2}\right) \mathrm{r_{AC}}\ \sigma_A}{\left[\frac{\mathrm{IG_m} + \mathrm{IG_h}}{2}\right]}$$

11.8 Cambio en las frecuencias alélicas bajo selección artificial

Hasta ahora hemos visto el proceso de selección de diferentes maneras, pero sin conectarlas entre sí. Mientras que en el capítulo «Selección natural» analizamos diferentes formas de selección y sus implicancias en el cambio de frecuencias alélicas, en este capítulo vimos el proceso de selección desde la perspectiva del cambio en la distribución de los valores fenotípicos resultantes. Es momento, pues, de combinar las dos aproximaciones para intentar entender los cambios en las frecuencias alélicas que ocurren como resultante del proceso de selección basado en la información fenotípica disponible.

Abundemos un poco en lo anterior antes de avanzar en los cálculos. Mientras que en el capítulo «Selección natural» trabajamos sobre el **fitness** específico de cada genotipo en un locus determinado, en combinación con las frecuencias respectivas, en el presente capítulo supusimos que la característica de interés puede estar gobernada por uno o más loci y que no contamos con la información correspondiente (ni siquiera asumimos saber cuántos loci gobiernan la característica). Este hecho nos llevó a tener que trabajar con la información que contamos, básicamente la distribución fenotípica de la característica y el conocimiento sobre los parámetros genéticos más relevantes, en particular las varianzas fenotípicas y aditivas o la heredabilidad. Mientras que el primer enfoque nos permite derivaciones más profundas sobre el cambio de las frecuencias alélicas, excepto a partir de información genómica y en situaciones muy particulares, su aplicación no resulta para nada práctica o realizable. En cambio, la segunda aproximación (condicional a que se cumplan los supuestos) es totalmente practicable y ha sido la base del mejoramiento genético animal durante mucho tiempo y en particular durante la segunda mitad del siglo XX.

Veamos entonces ahora cómo podemos conectar las dos aproximaciones. Para ello vamos a suponer al comienzo una situación muy estilizada, de una característica que es gobernada mayormente por un locus con dos alelos, donde además el modo de acción es de dominancia completa del alelo A_1. La situación se ilustra en la figura 11.11. Como nos encontramos en una situación de dominancia total de uno de los dos alelos (el A_1), entonces los genotipos A_1A_1 y A_1A_2 serán indistinguibles desde el punto de vista de los fenotipos para la característica, pero ambos serán diferentes del genotipo A_2A_2. Por lo tanto, desde el punto de vista de los genotipos, nuestra población la podemos descomponer en dos distribuciones, las que poseen al menos un alelo A_1 (en rojo en la figura) y las que no tienen ningún alelo A_1. De acuerdo con el modelo genético básico, en la escala fenotípica (genotípica) transformada, los genotipos homocigotos A_1A_1 y A_2A_2 tenían valores respectivos de a y $-a$, por lo que la diferencia entre ellos es de $a-(-a)=a+a=2a$. Obviamente, el genotipo A_1A_2 pertenece a la misma clase fenotípica que el A_1A_1, por lo que compartirá distribución de valores con este. En nuestro caso, asumiremos adicionalmente que no existe interacción genotipo ambiente y que la varianza residual es igual en ambas clases fenotípicas (y centrada en cero), por lo que la amplitud de las distribuciones respecto a la media es similar para ambas clases, como se muestra en la figura 11.11.

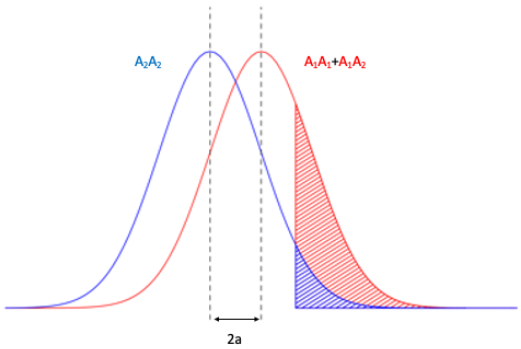


Figura 11.11. Cambio en las frecuencias alélicas como producto de la selección por umbral. En este caso asumimos que el alelo A_1 es completamente dominante, por lo que existen solo dos fenotipos para este locus, lo que se ve enmascarado por los efectos ambientales y del resto de los loci dando la distribución continua para la característica. La diferencia entre las medias de los genotipos $A_1A_1 + A_1A_2$ y A_2A_2 será igual a $a-(-a)=2a$. En la figura se representa la situación en que $q^2=\frac{1}{2}$

Claramente, si fijamos un umbral de selección (como hemos hecho en este capítulo), la proporción de individuos seleccionados de la distribución A_1A_1/A_1A_2 es mayor que la de los individuos A_2A_2. Este es el caso particular en que $q^2=\frac{1}{2}$, por lo que hay la misma proporción en las dos clases fenotípicas, pero lo mismo se aplicaría a cualquier frecuencia si consideramos la relación entre las frecuencias de las dos clases antes de la selección y en los individuos seleccionados. Eso es así debido a que la distribución de la clase A_2A_2 está más a la izquierda respecto al umbral y por lo tanto solo los valores más extremos respecto a la media de la clase alcanzan el umbral.

En principio, no tenemos una forma de conocer la frecuencia del alelo A_2 entre los seleccionados, pero podemos recurrir al concepto de regresión que hemos utilizado antes para

determinar esta frecuencia a partir de la diferencia de la media de los individuos seleccionados. Esta media era, por definición, el **diferencial de selección** S, por lo que la nueva frecuencia del alelo A_2 será igual a

$$q_1 = q + b_{qP}S \tag{11.55}$$

Es decir, la nueva frecuencia será igual a la frecuencia anterior q, más la regresión de la frecuencia en el diferencial de selección. Para determinar el coeficiente de selección vamos a necesitar calcular la covarianza entre las frecuencias de los alelos (q) en los genotipos y los valores fenotípicos. Para ello vamos a emplear la siguiente tabla

Genotipo	Frecuencia	Alelo q	$G = P$
A_1A_1	p^2	0	a
A_1A_2	$2pq$	$\frac{1}{2}$	d
A_2A_2	q^2	1	$-a$

En ella tenemos las frecuencias de cada genotipo (asumiendo equilibrio de Hardy-Weinberg), la frecuencia del alelo q en cada genotipo (no hay ninguno en A_1A_1, hay uno de dos en A_1A_2 y hay dos de dos en A_2A_2), así como el valor genotípico (observar que los heterocigotas se especifican como d, por lo que no estamos asumiendo ningún modo particular de relación entre alelos). La covarianza podemos calcularla como el producto cruzado de la proporción de alelos q en cada genotipo con el valor genotípico, ponderado por la frecuencia de cada genotipo. En la definición que hemos usado de covarianza hacíamos este producto cruzado de los desvíos respecto a las medias respectivas, pero más práctico en este caso resulta hacer los productos de las columnas 2, 3 y 4 de la tabla y a eso restarle el producto de las medias de la frecuencia alélica por la media fenotípica (genotípica). Del capítulo «El modelo genético básico», recordarás que la media era $M = a(p - q) + 2pqd$ (ecuación (8.35)). Por lo tanto, la covarianza será

$$Cov_{qP} = [a \times 0 \times p^2 + d \times \frac{1}{2}2pq - a \times 1 \times q^2] - qM = pqd - q^2a - qM \tag{11.56}$$

Sustituyendo el valor de M en la ecuación anterior, recordando que $1-2q = [p+q]-2q = p - q$, tenemos entonces que

$$\begin{aligned} \mathrm{Cov}_{qP} &= pqd - q^2a - q[a(p - q) + 2pqd] \\ &= pqd - q^2a - pqa + q^2a - 2pq^2d \\ &= pq[a + d(1 - 2q)] = pq[a + d(q - p)] \\ &\therefore \mathrm{Cov}_{qP} = -pq\alpha \end{aligned} \tag{11.57}$$

con $\alpha = a + d(q - p)$ el efecto de sustitución alélica (ecuación (8.51)). Por lo tanto, usando este resultado para la covarianza, con la varianza fenotípica igual σ^2_P, el coeficiente de regresión será simplemente

$$b_{qP} = -\frac{pq\alpha}{\sigma^2_P} \tag{11.58}$$

Por otra parte, teniendo en cuenta que el diferencial de selección podemos escribirlo como el producto entre la intensidad de selección y el desvío estándar fenotípico, es decir $S = i\sigma_P$, tenemos por lo tanto que la ecuación (11.55) la podemos escribir ahora como

$$q_1 = q + b_{qP} i\sigma_P = q - \frac{pq\alpha}{\sigma_P^2} i\sigma_P \tag{11.59}$$

y por lo tanto, el cambio de frecuencia Δq esperable en una generación producto de la selección será igual a

$$\Delta q = q_1 - q = [q - \frac{pq\alpha}{\sigma_P^2} i\sigma_P] - q = -ipq\frac{\alpha}{\sigma_P} \tag{11.60}$$

La interpretación de esta ecuación es bastante sencilla: el cambio en la frecuencia del alelo A_2 será negativo ya que estamos seleccionando en contra del mismo. La intensidad del cambio será directamente proporcional a la intensidad de selección i, al producto de las frecuencias de los alelos (pq) y al efecto de sustitución alélica estandarizado entre el desvío fenotípico de la característica.

La pregunta que nos hacemos ahora es cómo se relaciona esto con los cambios de frecuencia esperados bajo selección que vimos en capítulos previos (en especial en el capítulo «Selección natural»). Para ello vamos a realizar una pequeña digresión para recordar como derivábamos el cambio de frecuencia en una generación bajo selección.

Por ejemplo, en el caso de que el modo de acción génica fuese de **dominancia completa**, con un coeficiente de selección s contra el recesivo, teníamos la siguiente tabla

Genotipo	A_1A_1	A_1A_2	A_2A_2	**Total**
Frecuencia inicial	p^2	$2pq$	q^2	1
Fitness	1	1	$1-s$	
Contribución	p^2	$2pq$	$q^2(1-s)$	$1-sq^2$

Por lo tanto, la frecuencia en la siguiente generación (es decir, después de la selección) sería

$$q_1 = \frac{q^2(1-s)+pq}{1-sq^2} \tag{11.61}$$

y el cambio de frecuencia, por lo tanto

$$\Delta q = q_1 - q = \frac{q^2(1-s)+pq}{1-sq^2} - \frac{q(1-sq^2)}{1-sq^2} = \frac{q^2(1-s)+(1-q)q-q(1-sq^2)}{1-sq^2} = \tag{11.62}$$

$$= \frac{q^2 - sq^2 + q - q^2 - q + sq^3}{1-sq^2} = \frac{-sq^2+sq^3}{1-sq^2} \quad \therefore \tag{11.63}$$

$$\Delta q = -\frac{sq^2(1-q)}{1-sq^2} \tag{11.64}$$

En el caso

Genotipos	A_1A_1	A_1A_2	A_2A_2	**Total**
Frecuencia inicial	p^2	$2pq$	q^2	1
Fitness	1	$1 - rac12s$	$1 - s$	
Contribución	p^2	$2pq - pqs$	$q^2(1 - s)$	$1 - sq$

El nuevo total de frecuencias (para relativizar) el resultado de cada alelo será

$$p^2 + 2pq - pqs + q^2(1 - s) = 1 - pqs - q^2s = 1 - s[pq + q^2] = 1 - s[q(p + 1)] = 1 - sq \tag{11.65}$$

La nueva frecuencia del alelo A_2, q_1, será por lo tanto

$$q_1 = \frac{pq - \frac{1}{2}pqs + q^2(1 - s)}{1 - sq} = \frac{q - q^2 - \frac{1}{2}s[q - q^2] + q^2 - q^2s)}{1 - sq} = \tag{11.66}$$

$$= \frac{q - \frac{1}{2}s[q - q^2] - q^2s)}{1 - sq} = \frac{q[1 - \frac{1}{2}s(1 - q) - qs]}{1 - sq} = \tag{11.67}$$

$$= \frac{q[1 - \frac{1}{2}s + \frac{1}{2}qs - qs]}{1 - sq} = \frac{q[1 - \frac{1}{2}s - \frac{1}{2}qs]}{1 - sq} = \tag{11.68}$$

$$= \frac{q[1 - \frac{1}{2}s - \frac{1}{2}qs]}{1 - sq} \tag{11.69}$$

y entonces, el cambio de frecuencia correspondiente será igual a

$$\Delta q = q_1 - q = \frac{q[1 - \frac{1}{2}s - \frac{1}{2}qs]}{1 - sq} - q = \frac{q[1 - \frac{1}{2}s - \frac{1}{2}qs] - q[1 - sq]}{1 - sq} = \tag{11.70}$$

$$= \frac{q - \frac{1}{2}qs - \frac{1}{2}q^2s - q + sq^2}{1 - sq} = \frac{-\frac{1}{2}qs + \frac{1}{2}q^2s}{1 - sq} = -\frac{\frac{1}{2}sq(1 - q)}{1 - sq} \tag{11.71}$$

En general, cuando las frecuencias del alelo recesivo son bajas, podemos considerar una buena aproximación el hecho de que el denominador será cercano a 1 (en ambos casos, sin dominancia y sobre todo con dominancia completa), lo que nos deja solo con el numerador de las correspondientes fórmulas.

La idea ahora es entonces igualar los resultados obtenidos para el cambio de frecuencia en los distintos modos génicos en una generación de selección con el resultado obtenido a partir del diferencial (intensidad) de selección.

Cuando no hay dominancia, como vimos antes, el cambio de frecuencia en el alelo A_2, asumiendo bajas frecuencias de este alelo, podemos aproximarlo a

$$\Delta q = -\frac{\frac{1}{2}sq(1 - q)}{1 - sq^2} \sim \frac{1}{2}sq(1 - q) \tag{11.72}$$

En este caso, por otra parte, $d = 0$ y $\alpha = a + d(q - p) = a$, por lo que sustituyendo ese valor de α en la ecuación (11.60) e igualando al resultado de la ecuación (11.72), tenemos

$$ipq\frac{a}{\sigma_P} \sim \frac{1}{2}sq(1-q) \tag{11.73}$$

Por otra parte, si la dominancia es completa, entonces $d = a$ y entonces $\alpha = a+a(q-p) = a(p+q) + a(q-p) = 2qa$, por lo que

$$ipq\frac{2qa}{\sigma_P} \sim sq^2(1-q) \tag{11.74}$$

En general, ambas ecuaciones, (11.74) y (11.73), se reducen a

$$s = i\frac{2a}{\sigma_P} \tag{11.75}$$

Por ejemplo, la ecuación (11.73), asumiendo que la aproximación es exacta, se reduce mediante

$$\begin{aligned} ipq\frac{a}{\sigma_P} = \frac{1}{2}sq(1-q) &\Leftrightarrow 2ipq\frac{a}{\sigma_P} = spq \\ &\Leftrightarrow s = i\frac{2a}{\sigma_P} \end{aligned} \tag{11.76}$$

mientras que la ecuación (11.74) se reduce mediante

$$ipq\frac{2qa}{\sigma_P} = sq^2(1-q) \Leftrightarrow ip\frac{2a}{\sigma_P} = s(1-q) \Leftrightarrow \tag{11.77}$$

$$i(1-q)\frac{2a}{\sigma_P} = s(1-q) \Leftrightarrow i\frac{2a}{\sigma_P} = s \tag{11.78}$$

Es decir, la ecuación (11.73) nos dice que hemos encontrado una relación directa entre las dos formas de expresar la selección que venimos trabajando. El coeficiente de selección que hemos usado para encontrar las frecuencias de los genotipos y el cambio de la frecuencia del alelo A_2 es igual a la intensidad de selección multiplicada por el factor $\frac{2a}{\sigma_P}$, que no es otra cosa que la diferencia entre los genotipos (fenotipos) de los homocigotas, estandarizados entre el desvío estándar fenotípico de la característica.

Una otra forma interesante de ver este resultado es haciendo uso de la relación $i = \frac{S}{\sigma_P}$ y de la ecuación del criador. A partir de la ecuación (11.73), utilizando la relación entre intensidad de selección y diferencial de selección podemos hacer

$$s = i\frac{2a}{\sigma_P} = \frac{S}{\sigma_P}\frac{2a}{\sigma_P} = S\frac{2a}{\sigma_P^2} \Leftrightarrow S = s\frac{\sigma_P^2}{2a} \tag{11.79}$$

Sustituyendo este resultado en la ecuación del criador, tenemos

$$R = h^2S = h^2s\frac{\sigma_P^2}{2a} = \frac{\sigma_A^2}{\sigma_P^2}s\frac{\sigma_P^2}{2a} \;\therefore \tag{11.80}$$

$$R = \sigma_A^2 \frac{s}{2a} \tag{11.81}$$

El resultado de la ecuación (11.81) tiene dos formas interesantes de verse. La primera tiene una relación directa con el teorema fundamental de la selección natural que vimos en el capítulo de «Selección natural». Si recuerdas, la idea del mismo era que existía una relación directa entre la tasa de cambio en la aptitud (**fitness**) debida a la selección natural en una característica y la varianza genético-aditiva de la misma. La ecuación (11.81) establece que el progreso genético es directamente proporcional a la varianza genético-aditiva de la característica. Si no hay varianza genética, sin importar la magnitud del coeficiente de selección, no existirá progreso genético. Esto nos lleva directamente a reflexionar sobre la posibilidad de utilizar la clonación para propagar el mejoramiento genético, una idea que pareció extenderse a partir de la clonación de la oveja Dolly. Si clonásemos **al mejor animal existente** (un concepto algo absurdo) y lo utilizáramos para reemplazar todo el stock, todos los animales serían idénticos, por lo que no existiría varianza genético-aditiva y por lo tanto nos anularíamos la posibilidad de mejorar la característica (al menos hasta que aparezcan nuevas mutaciones que introduzcan varianza y que sean en el sentido correcto).

La segunda forma interesante de ver este resultado es si sustituimos en la ecuación (11.81) la varianza genético-aditiva por su valor esperado en el modelo de un locus con dos alelos, ecuación (8.120), $V_A = \sigma_A^2 = 2pq\alpha^2$

$$R = \sigma_A^2 \frac{s}{2a} = \frac{2pq\alpha^2}{2a} s \;\therefore\; R = \frac{pq\alpha^2}{a} s \tag{11.82}$$

De acuerdo con esta última ecuación, a valores determinados del coeficiente de selección y del efecto de sustitución (que depende del modo de acción génica y de las frecuencias alélicas), la respuesta a la selección será máxima cuando el producto pq sea máximo, cosa que como ya vimos ocurre cuando $p = q = \frac{1}{2}$. Es decir, el progreso genético será máximo a frecuencias intermedias y luego irá disminuyendo a medida que la frecuencia del alelo que seleccionamos en contra decrece.

PARA RECORDAR

- Si tomamos como ejemplo que la proporción de individuos seleccionados A_1A_1/A_1A_2 de una distribución es mayor que la de los individuos A_2A_2. Este es el caso particular en que $q^2 = 1/2$, por lo que hay la misma proporción en las dos clases fenotípicas, pero lo mismo se aplicaría a cualquier frecuencia si consideramos la relación entre las frecuencias de las dos clases antes de la selección y en los individuos seleccionados. Eso es así debido a que la distribución de la clase A_2A_2 está más a la izquierda respecto al umbral y por lo tanto solo los valores más extremos respecto a la media de la clase alcanzan el umbral.

- La nueva frecuencia del alelo A_2 será $q_1 = q + b_{qP}S$. Es decir, la nueva frecuencia será igual a la frecuencia anterior q, más la regresión de la frecuencia en el diferencial de selección.
- El cambio de frecuencia Δq esperable en una generación producto de la selección será igual a$\Delta q = -ipq\frac{\alpha}{\sigma_P}$. Dónde el cambio en la frecuencia del alelo A_2 será negativo ya que estamos seleccionando en contra del mismo. La **intensidad del cambio será directamente proporcional a la intensidad de selección i, al producto de las frecuencias de los alelos (pq) y al efecto de sustitución alélica estandarizado entre el desvío fenotípico de la característica.**
- Cuando no hay dominancia, el cambio de frecuencia en el alelo A_2, asumiendo bajas frecuencias de este alelo, podemos aproximarlo a $\Delta q = -\frac{\frac{1}{2}sq(1-q)}{1-sq^2} \sim \frac{1}{2}sq(1-q)$
- Por otra parte, si la dominancia es completa, entonces $ipq\frac{2q\alpha}{\sigma_P} \sim sq^2(1-q)$
- Podemos reducir ambas ecuaciones a $s = i\frac{2\alpha}{\sigma_P}$, que no es otra cosa que la intensidad multiplicada por un factor que representa la diferencia entre los genotipos (fenotipos) de los homocigotas, estandarizados entre el desvío estándar fenotípico de la característica.
- A valores determinados del coeficiente de selección y del efecto de sustitución (que depende del modo de acción génica y de las frecuencias alélicas), la respuesta a la selección será máxima cuando el producto pq sea máximo ($p = q = 1/2$). Es decir, el progreso genético será máximo a frecuencias intermedias y luego irá disminuyendo a medida que la frecuencia del alelo que seleccionamos en contra decrece: $R = \sigma_A^2\frac{s}{2a} = \frac{2pq\alpha^2}{2a}s \therefore R = \frac{pq\alpha^2}{a}s$

11.9 El efecto de la selección en la variabilidad genética: el efecto Bulmer

Como el proceso de selección implica elegir determinados genotipos sobre otros, en particular en características determinadas por varios genes, resulta intuitivo que la variabilidad genética debería disminuir. Sin embargo, en su libro de 1970, Crow y Kimura demuestran que a medida que se incluyen más *loci* en el modelo, la pérdida de varianza se va haciendo cada vez menor, hasta que desaparece cuando hay un número infinito de *loci* involucrados. Como esto resulta en principio contra-intuitivo, en un artículo muy interesante, Bulmer (1971) demuestra que si $G = \sum_{i=1}^{N} g_i$, con g_i el genotipo en cada uno de los *loci*, es posible descomponer la varianza de acuerdo con

$$V_G = V(G) = Var(\sum_{i=1}^{N} g_i) = \sum_{i=1}^{N} Var(g_i) + \sum_{i \neq j} Cov(g_i,g_j) \qquad (11.83)$$

y claramente, cuando existe apareamiento aleatorio y en ausencia de selección, las covarianzas no existen y lo anterior queda reducido a $V_G = \sum_{i=1}^{N} Var(g_i)$.

Veamos que ocurre con las covarianzas bajo selección en el caso de dos *loci* y para ello recurriremos a todo los trabajado en el capítulo en el que analizamos la «Dinámica de dos *loci*». Supongamos que, usando la misma notación que en ese capítulo, en el *locus A* tenemos

dos alelos, A_1 y A_2, mientras que en el B los alelos son B_1 y B_2. A la combinación A_1B_1 le habíamos asignado frecuencia x_1, a la A_1B_2 frecuencia x_2, a la A_2B_1 frecuencia x_3 y a la A_2B_2 frecuencia x_4. A las frecuencias de los alelos los habíamos llamado p_1 y q_1 a los del *locus* A, mientras que p_2 y q_2 a los del locus B. Supongamos que, solo para entender la evolución de la covarianza, asumimos que tanto A_1 como B_1 poseen un valor g, mientras que los otros dos alelos no aportan nada, es decir 0. Si llamamos m_1 y m_2 a las medias de cada *loci*, tenemos que

$$\begin{aligned} m_1 &= p_1 g + q_1(0) = p_1 g \\ m_2 &= p_2 g + q_2(0) = p_2 g \end{aligned} \tag{11.84}$$

Si definimos dos variables estadísticas y_1 e y_2 con los valores de los alelos en los *loci* 1 y 2, sabemos de estadística que $Cov(y_1,y_2) = \sum_i (y_{1,i}y_{2,i}) freq(y_{1,i},y_{2,i}) - m_1 m_2$. En el caso de que los dos *loci* se comporten como estadísticamente independientes, considerando que tenemos 4 combinaciones de alelos presentes en los dos *loci* (A_1B_1, A_1B_2, A_2B_1 y A_2B_2)tenemos que

$$\begin{aligned} Cov(y_1,y_2) = (p_1 g)(p_2 g) + (p_1 g)(q_2 0) + (q_1 0)(p_2 g) + (q_1 0)(q_2 0) - (p_1 g)(p_2 g) = \\ (p_1 g)(p_2 g) + 0 + 0 + 0 - (p_1 g)(p_2 g) = 0 \therefore Cov(y_1,y_2) = 0 \end{aligned} \tag{11.85}$$

como era de esperar. Sin embargo, en el caso de no-independencia, ahora las frecuencias $(y_{1,i},y_{2,i})$ ya no serán el producto de las frecuencias de los alelos que forman el gameto, sino que serán las frecuencias reales x_1, x_2, x_3 y x_4.

$$\begin{aligned} Cov(y_1,y_2) = x_1(g.g) + x_2(g.0) + x_3(0.g) + x_4(0.0) - (p_1 g)(p_2 g) = \\ x_1 g^2 + 0 + 0 + 0 - (p_1 p_2) g^2 = x_1 g^2 - p_1 p_2 g^2 = g^2(x_1 - p_1 p_2) \therefore \\ Cov(y_1,y_2) = g^2(x_1 - p_1 p_2) \end{aligned} \tag{11.86}$$

Pero si recordamos del capítulo 5 que una de las definiciones del desequilibrio de ligamiento era precisamente $D = x_1 - p_1 p_2$, es decir la diferencia entre la frecuencia real del gameto y la esperada si los alelos de los *loci* que lo componen fuesen independientes, tenemos entonces que

$$Cov(y_1,y_2) = g^2 D \tag{11.87}$$

En general, esta covarianza es negativa, ya que D suele serlo. La razón de esto se encuentra en que si bien el gameto A_1B_1 se ve favorecido, también lo son A_1B_2 y A_2B_1, todo a expensas del A_2B_2. Como $D = x_1x_4 - x_2x_3$, en general, $x_2x_3 > x_1x_4$ y por lo tanto $D < 0$. Por ejemplo, en el caso de que la selección actúe en contra del gameto A_2B_2, eliminándolo (frecuencia igual a 0), si llamamos f_{11} a la frecuencia del gameto con los dos alelos favorables (el suscrito $_{11}$ se debe a que es el alelo 1 en ambos *loci*), es decir, A_1B_1, los otros dos gametos tendrán frecuencias intermedias iguales. Si definimos un coeficiente

$b < 1$ que establezca la relación entre gametos, es decir bf_{11} para los gametos A_1B_2 y A_2B_1, como la suma de todas las frecuencias debe ser igual a 1, tenemos

$$f_{11} + bf_{11} + bf_{11} + 0 = 1 \therefore f_{11}(1 + 2b) = 1 \therefore \\ f_{11} = \frac{1}{(1 + 2b)} \tag{11.88}$$

Pero $p_1 = f_{11} + bf_{11}$ y $p_2 = f_{11} + bf_{11}$ ya que son las frecuencias obtenidas de los gametos $A_1B_1 + A_1B_2$ (A_1) y $A_1B_1 + A_2B_1$ (B_1), por lo que su producto será $p_1p_2 = (f_{11} + bf_{11})^2 = f_{11}^2(1 + b)^2$. Si analizamos ahora la relación entre la frecuencia de gameto A_1B_1 esperada bajo apareamiento aleatorio sin ligamiento en relación a la frecuencia real del gameto (f_{11}, tenemos

$$\frac{p_1p_2}{f_{11}} = \frac{f_{11}^2(1 + b)^2}{f_{11}} = f_{11}(1 + b)^2 = \frac{(1 + b)^2}{(1 + 2b)} = \frac{(1 + 2b) + b^2}{(1 + 2b)} \geq 1 \iff \\ p_1p_2 > f_{11} \tag{11.89}$$

Es decir, demostramos en este caso que particular que el producto de las frecuencias de los alelos que forman el gameto A_1B_1 debe ser mayor (solo es igual cuando $b = 0$) que la frecuencia del gameto. Más aún, este resultado al que arribamos para dos *loci* se puede extender al modelo de alelos infinitos.

Por lo tanto, llegamos a que en presencia desequilibrio de ligamiento suelen existir covarianzas negativas entre alelos. Si ahora introducimos este resultado en la ecuación (11.83), tenemos que bajo selección tenemos una reducción de la varianza genética, producto de las combinaciones específicas que fueron seleccionadas.

La reducción de la varianza ocurre como producto de seleccionar individuos extremos y es particularmente importante en las primeras generaciones de selección, para luego ir disminuyendo. Esta reducción es producto de que los individuos seleccionados se parecen más entre que sí que con el resto de la población, induciendo una correlación negativa entre los efectos alélicos. Bulmer (1971) demuestra que la reducción en la varianza fenotípica en los padres seleccionados es $V_P^\star = V_P(1 - k)$, con $k = i(i - x)$, donde i es la intensidad de selección y x el punto de truncamiento (umbral de selección). Uniendo esto con la definición ya vista de heredabilidad, $h^2 = V_A/V_P \iff V_A = h^2V_P$, tenemos que la reducción en varianza aditiva es igual a

$$V_A^\star = V_A(1 - kh^2) \tag{11.90}$$

En el caso de que los progenitores seleccionados generen familias enteras, con correlación genética igual a $1/2$, la mitad de la varianza en la descendencia corresponderá a varianza dentro de familia V_{AW} y la otra mitad a varianza entre familia V_{AB}.

Como en ausencia de ligamiento la varianza dentro de familias no ve afectada por el fenotipo de los padres, tenemos ahora

$$\begin{aligned} V_{AW} &= \frac{1}{2}V_A \\ V_{AB} = \frac{1}{2}V_A^\star &= \frac{1}{2}V_A(1-kh^2) \\ V_{AW} + V_{AB} &= V_A(1-\frac{1}{2}kh^2) \end{aligned} \tag{11.91}$$

por lo que el efecto debido al desequilibrio se redujo a la mitad. Esto continuará de la misma forma en las próximas generaciones, por lo que la ecuación de recurrencia que describe la reducción de la varianza en la generación t toma la forma

$$V_{A,t} = \frac{1}{2}(1-kh_{t-1}^2)V_{A,t-1} + \frac{1}{2}V_{A,0} \tag{11.92}$$

En resumen, tenemos que

- La selección no afecta la varianza aditiva en cada uno de los genes.
- La selección reduce la varianza aditiva de los genes considerados en conjunto.
- La explicación de esto surge de aparición de covarianzas negativas entre los alelos seleccionados.

Para ampliar en el tema, recomendamos los libros de Caballero (2020) y Blasco (2025), además del artículo original de Bulmer (1971).

PARA RECORDAR

- El efecto Bulmer implica la reducción de la varianza aditiva en la población como resultado de seleccionar individuos extremos.
- Su efecto es importante en la primeras generaciones de selección y luego disminuye hasta ser casi insignificante.
- Mientras que la selección no afecta la varianza aditiva en cada uno de los genes, la misma reduce la varianza aditiva de los genes considerados en conjunto, debido a la aparición de covarianzas negativas entre los alelos seleccionados.
- La ecuación de recurrencia que describe la reducción de la varianza en la generación t, a partir de la varianza aditiva inicial $V_{A,0}$, es

$$V_{A,t} = \frac{1}{2}(1-kh_{t-1}^2)V_{A,t-1} + \frac{1}{2}V_{A,0} \tag{11.93}$$

11.10 El diferencial de selección direccional y la identidad de Robertson-Price

Una perspectiva integradora del diferencial de selección **direccional** viene de la llamada **identidad de Robertson-Price** (Robertson 1966; Price 1970, 1972). En esta sección usaremos la notación de Lynch y Walsh (1998), así como el razonamiento empleado, con z representando el fenotipo de una característica de interés para el **fitness.** Hacemos énfasis en **direccional**, ya que es el tipo de selección que hemos manejado esencialmente en este capítulo (seleccionamos los individuos que se encontraban hacia un extremo de la distribución). Como vimos previamente, el diferencial de selección es igual a la diferencia de los individuos seleccionados (pero antes de la reproducción) y la media de la población en general, es decir

$$S = \mu_s - \mu \tag{11.94}$$

Por simplicidad asumamos que las diferencias entre los individuos solo conciernen a la probabilidad de llegar a la madurez, por lo que la función de **fitness** $W(z)$ es la probabilidad de que los individuos con fenotipo z alcancen la edad reproductiva. En particular, no asumimos acá ninguna forma particular de $W(z)$, por lo que esta podría ser una función discontinua, como en **selección por truncamiento**, o una continua, como en la **selección gaussiana** (Houchmandzadeh (2014)). Si llamamos $p(z)$ a la función de densidad correspondiente al fenotipo z **antes de la selección**, entonces la densidad después de la selección será

$$p_s(z) = \frac{W(z)\, p(z)}{\int W(z)\, p(z)\, dz} \tag{11.95}$$

La lógica del numerador de la ecuación (11.95) es sencilla, si pensamos que $W(z)$ es la ponderación del éxito de cada fenotipo z, multiplicado en el numerador por la densidad correspondiente a ese fenotipo. Notemos que el denominador ya no integra necesariamente a 1 y por lo tanto es el factor de escalado de la nueva densidad. Por otra parte, $\int W(z)\, p(z)\, dz = \bar{W}$, el **fitness medio individual.** Si ahora llamamos a $w(z) = \frac{W(z)}{\bar{W}}$ el **fitness relativo** del fenotipo z, entonces la ecuación (11.95) se transforma en

$$p_s(z) = \frac{W(z)\, p(z)}{\int W(z)\, p(z)\, dz} = \frac{W(z)\, p(z)}{\bar{W}} = \frac{W(z)}{\bar{W}} p(z) \;\; \therefore \tag{11.96}$$

$$p_s(z) = w(z)\, p(z) \tag{11.97}$$

El fenotipo medio después de la selección estará dado por

$$\mu_s = \int z\, p_s(z)\, dz = \int z\, w(z)\, p(z)\, dz = E[z\, w(z)] \tag{11.98}$$

que es la esperanza del producto del fenotipo por su **fitness relativo.** Por otra parte, el **fitness medio relativo** en la población estará dado por

$$\bar{w} = \int w(z)\, p(z)\, dz = \frac{1}{\bar{W}} \int W(z)\, p(z)\, dz = \frac{\bar{W}}{\bar{W}} = 1 \tag{11.99}$$

El resultado de la ecuación (11.99) nos muestra que el **fitness medio relativo** de una población siempre será igual a 1. Al mismo tiempo, la esperanza de la media de la población antes de la selección será igual a

$$\mu = E(z) \times E(w) = E(z) \times 1 \tag{11.100}$$

Uniendo este último resultado con la ecuación (11.98) y sustituyendo ambos en la ecuación (11.94), tenemos

$$S = \mu_s - \mu = E[z\ w(z)] - E(z) \times E(w) = \sigma[z,w(z)] \tag{11.101}$$

con $\sigma[z,w(z)]$ representando la covarianza entre el fenotipo y el **fitness relativo.**

Bajo el supuesto de linealidad en la respuesta del promedio de los padres hacia su descendencia, como asumimos en otras partes, el cambio generacional en función del diferencial de selección direccional será

$$R = \Delta\mu = \mu_o - \mu = \beta(\mu_s - \mu) = \beta S = h^2\ S \tag{11.102}$$

la última igualdad (asumiendo que el coeficiente de regresión lineal es $\beta = h^2$) correspondiendo a la **ecuación del criador**, que vimos antes. Uniendo esto último con la ecuación (11.101), tenemos que (utilizando la notación de z para el fenotipo) podemos escribir el progreso genético generacional como

$$R = h^2\ S = \frac{\sigma_A^2}{\sigma_z^2} S = \frac{\sigma_A^2}{\sigma_z^2} \sigma[z,w(z)]\ \therefore \tag{11.103}$$

$$R = \sigma_A^2 \frac{\sigma[z,w(z)]}{\sigma_z^2} = \sigma_A^2 b_{w(z),z} \tag{11.104}$$

donde $b_{w(z),z}$ representa el coeficiente de regresión del **fitness relativo** en el **fenotipo.** La forma de la ecuación (11.104) tiene una estrecha relación con la ecuación (11.81) donde estudiamos el cambio de frecuencias bajo selección y nuevamente resalta la relación existente entre la respuesta a la selección y la varianza aditiva de la característica.

PARA RECORDAR

- Si asumimos que las diferencias entre los individuos solo conciernen a la probabilidad de llegar a la madurez, por lo que la función de fitness $W(z)$ es la probabilidad de que los individuos con fenotipo z alcancen la edad reproductiva. En particular, no asumiendo ninguna forma particular de $W(z)$, por lo que esta podría ser una función discontinua, como en selección por truncamiento, o una continua, como en la selección gaussiana. Si llamamos $p(z)$ a la función de densidad correspondiente al fenotipo z antes de la selección, entonces la densidad después de la selección será $p_s(z) = \frac{W(z)\ p(z)}{\int W(z)\ p(z)\ dz}$

- El **fitness medio relativo** de una población **siempre será igual a 1**. Al mismo tiempo, la esperanza de la media de la población antes de la selección será igual $E(z) \cdot 1$, con el diferencial de selección siendo igual a $\sigma[z,w(z)]$ representando la covarianza entre el fenotipo y el fitness relativo.
- Utilizando la notación de z para el fenotipo podemos escribir el progreso genético generacional como $R = h^2\ S = \sigma_A^2 b_{w(z),z}$ donde $b_{w(z),z}$ representa el coeficiente de regresión del **fitness relativo** en el **fenotipo**.

11.11 Conclusión

Hasta este momento hemos estado aproximándonos al estudio de caracteres fenotípicos influidos por múltiples genes. Hemos dejado de lado, sin embargo, qué acontece cuando fuerzas externas influyen sobre las frecuencias de los alelos referentes a dichos loci. Cuando esta influencia es ejercida por la acción de un ser humano, se habla generalmente de selección artificial. En este capítulo nos volcamos a estudiar cómo responden caracteres fenotípicos de interés a la selección artificial.

De alguna forma retomamos conceptos que desarrollamos en la segunda parte de este libro, pero con un cambio fundamental, dado por nuestro objeto de estudio. Sabemos que el cambio fenotípico se debe al cambio en las frecuencias alélicas de múltiples loci, pero intentar desarrollar un modelo explícito para este fenómeno sería imposible; para hacerlo tendríamos que tener en cuenta la posibilidad de cambio en múltiples frecuencias a la vez, y modelar todas las interacciones y motivos de desvío subyacentes. Como alternativa, decidimos tomar una aproximación desde la estadística, analizando media, varianza y covarianza del carácter fenotípico de interés. Al desarrollar desde esta perspectiva llegamos a algunos resultados fundamentales de la genética cuantitativa, como puede ser **la ecuación del criador**, la cual justamente relaciona la respuesta a la selección con la intensidad de la misma y la heredabilidad del carácter (un concepto central que desarrollamos en el capítulo anterior).

En los dos siguientes capítulos ampliaremos nuestra perspectiva, teniendo como base la siguiente cuestión: ¿cómo afecta la selección en un carácter fenotípico a otros caracteres, y cómo podemos tener esto en cuenta al momento de ejercer la selección artificial?

11.12 Actividades

11.12.1 Control de lectura

1. ¿Qué papel juega la heredabilidad de una característica durante el proceso de selección artificial? Desarrolla conceptualmente y realice un planteo matemático explícito que sintetice los conceptos presentados.
2. ¿Qué tipos de factores de corrección existen? Menciona en qué casos se aplica cada uno.
3. ¿Qué relación existe entre el diferencial de selección y la varianza fenotípica de una característica? ¿Por qué es importante tener esto en cuenta al comparar presiones de selección?
4. ¿Qué relación existe entre la proporción de animales seleccionados y la intensidad de selección? ¿Brindan la misma información estas medidas?
5. ¿Cómo se evalúa a nivel práctico la respuesta de una característica a la selección artificial?

11.12.2 ¿Verdadero o falso?

1. Existen diferencias biológicas subyacentes que llevan a la necesidad de emplear factores de corrección al momento de comparar una característica en diferentes clases de estudio (por ejemplo, sexos). Estos factores son universales para dicha característica.
2. Generalmente, se cuestiona por qué en la selección por truncación la distribución de la característica fenotípica seleccionada en los progenitores no se mantiene igual en la descendencia. Esta discrepancia se debe a la influencia del ambiente en la obtención de los fenotipos.
3. Un aumento en el intervalo generacional conlleva un incremento en la respuesta a la selección anual.
4. En una población animal comercial de tamaño estable los machos contribuyen en mayor medida al progreso genético que las hembras.
5. Si la heredabilidad de una característica es baja, es posible realizar selección compensando esta falta de varianza genético-aditiva con factores ambientales.

Soluciones

1. **Falso**. Los factores de corrección son idiosincráticos, dependiendo de la población considerada. Por ejemplo, podemos considerar que no será lo mismo corregir un factor de altura en poblaciones asiáticas que en poblaciones africanas. A su vez, es de esperar que estos factores de corrección también varíen con el tiempo.
2. **Verdadero**. Es importante resaltar la influencia del ambiente en la distribución de valores que toma la característica fenotípica de nuestro interés. Aun cuando su genotipo no posea gran potencial, los individuos pueden verse beneficiados por un ambiente que fomente a la característica estudiada (lo mismo puede ocurrir en sentido opuesto, perjudicando a individuos de alto potencial genético). No obstante, solo el genotipo es heredado de una generación a la siguiente (es decir, los ambientes no son heredables). Esto implica que, aun cuando se haya establecido un régimen de selección artificial con base en los fenotipos, no seremos capaces de asegurar que la genética

de los individuos seleccionados lleve necesariamente a obtener una progenie que correlacione perfectamente a nivel fenotípico.

3. **Falso.** Notemos que la respuesta de selección anual es una medida de respuesta estandarizada (es decir, en un intervalo definido de tiempo). Recordemos que el intervalo generacional (IG) es el promedio de la edad de sus progenitores cuando nacen los descendientes que los reemplazarán. A la vez, el IG es el denominador del cálculo del progreso genético anual; de esto queda claro que un aumento en el IG conlleva necesariamente una reducción de esta medida. Conceptualmente, esto es coherente con el hecho de que cuanta mayor edad poseen los progenitores al generar la progenie de reemplazo, mayor será el tiempo necesario para observar el resultado del proceso selectivo llevado a cabo.
4. **Verdadero.** La presión de selección que se puede hacer en los machos es mucho mayor, en tanto se requiere un menor número por generación para mantener el tamaño poblacional estable. En el caso de las hembras, por el contrario, los tiempos de gestación suelen llevar a tener que ejercer una presión selectiva comparativamente menor (de lo contrario se elegiría un número absurdamente chico de progenitoras).
5. **Falso.** La selección (ya sea natural o artificial) es un proceso que requiere como condición necesaria la herencia.

11.12.3 Ejercicios

Ejercicio 11.1

Un grupo de criadores se dispone a seleccionar perros de la raza basset hound con el fin de aumentar el largo de sus orejas. Para ello llevan a cabo un proceso de selección por truncación, eligiendo perros de cada camada. Cuentan con 150 perros en total, los cuales presentan una media de largo de oreja que sigue una distribución normal $N(\mu = 32{,}5 \text{ cm}, \sigma = 1$

Asumiendo que por cada perro seleccionado se puede obtener como máximo cuatro perros, ¿cuál es el valor máximo para media de largo de oreja que se puede seleccionar manteniendo constante el número de perros en cada camada?

Solución

Recordemos que la intensidad de selección i se comporta tal que

$$i = \frac{S}{\sigma_p} = \frac{\bar{x}_S - \bar{x}_0}{\sigma_P} = \frac{z}{p} \tag{11.105}$$

donde S es el diferencial de selección (como desvío de la media; $\bar{x}_S - \bar{x}_0$), σ_P es la desviación fenotípica, z es la «altura» de la distribución normal de densidad $N(0,1)$ y p es la proporción de individuos seleccionados como reproductores. La distribución de largos de orejas es una distribución normal con desviación $\sigma_P = 1$, por lo que a efectos prácticos podremos trabajar con esta fórmula, teniendo en cuenta que reemplazamos la media por $\mu = X - \mu$ a efectos de tener en cuenta una distribución normal estándar $N(0,1)$.

Si se puede seleccionar a lo sumo $\frac{1}{4}$ de los individuos de cada camada (a efectos de mantener constante el número de la misma), esto equivale a decir $\frac{z}{p} = 0{,}25$. Por lo tanto, tenemos que en el caso extremo se da

$$i = \frac{\bar{x}_S - \bar{x}_0}{\sigma_P} = 0{,}25 \tag{11.106}$$

$$\bar{x}_S - \bar{x}_0 = 0{,}25 \times \sigma_P = 0{,}25 \times 1 \tag{11.107}$$

$$\bar{x}_S = \bar{x}_0 + 0{,}25 \Rightarrow \tag{11.108}$$

$$\bar{x_S} = 32{,}5 + 0{,}25 = 32{,}75 \text{ cm} \tag{11.109}$$

Es decir, con los datos planteados se deduce que si se seleccionan individuos con una media de selección que supere los 32,75 cm, resultaría inviable mantener el tamaño de camada constante.

Ejercicio 11.2

(Adaptado de J. Felsenstein 2019)

Consideremos un carácter fenotípico cuyo valor numérico está controlado por dos loci, sin varianza ambiental. Si los valores especificados por los genotipos de estos loci bialélicos son

	BB	Bb	bb
AA	6	7	8
Aa	7	8	9
aa	8	9	10

y las proporciones de las frecuencias génicas en ambos casos son 50 : 50 (tal que $p_A = 0{,}5$ y $p_B = 0{,}5$)

a. ¿Cuál será el valor medio para este fenotipo?

b. ¿Cuál será su varianza? ¿Y su desviación estándar?

c. ¿Son estos loci individualmente aditivos para sus efectos alélicos? ¿Muestran algún tipo de interacción (epistasis)?

d. En vista del punto anterior, ¿cuál será la heredabilidad del carácter?

e. Si realizamos selección artificial quedándonos solo con los individuos que poseen un valor fenotípico mayor o igual a 8, ¿cuál será la respuesta de selección en la primera generación obtenida?

Solución

a. El valor medio para el carácter fenotípico será la media ponderada de los valores fenotípicos según su respectiva frecuencia genotípica. A efectos de calcular estas últimas, asumiremos que los loci se encuentran segregando bajo equilibrio de Hardy-Weinberg y en equilibrio de ligamiento. Dado que las frecuencias génicas son $p_A = 0{,}5$ y $p_B = 0{,}5$, las frecuencias genotípicas serán

$$\begin{cases} f(AABB) = (p_A^2) \times (p_B^2) = 0{,}5^4 \\ f(AaBB) = (2 \times p_A p_a) \times (p_B^2) = 2 \times 0{,}5^4 \\ f(aaBB) = (p_a^2) \times (p_B^2) = 0{,}5^4 \\ f(AABb) = (p_A^2) \times (2 \times p_B p_b) = 2 \times 0{,}5^4 \\ f(AaBb) = (2 \times p_A p_a) \times (2 \times p_B p_b) = 4 \times 0{,}5^4 \\ f(aaBb) = (p_a^2) \times (2 \times p_B p_b) = 2 \times 0{,}5^4 \\ f(AAbb) = (p_A^2) \times (p_b^2) = 0{,}5^4 \\ f(Aabb) = (2 \times p_A p_a) \times (p_b^2) = 2 \times 0{,}5^4 \\ f(aabb) = (p_a^2) \times (p_b^2) = 0{,}5^4 \end{cases} \tag{11.110}$$

con lo que el valor medio (que llamaremos $\bar{x}$) estará dado por

$$\begin{aligned} \bar{x} &= (0{,}5^4 \times 6) + (2 \times 0{,}5^4 \times 7) + (0{,}5^4 \times 8) + (2 \times 0{,}5^4 \times 7) \\ &\quad + (4 \times 0{,}5^4 \times 8) + (2 \times 0{,}5^4 \times 9) + (0{,}5^4 \times 8) + (2 \times 0{,}5^4 \times 9) \\ &\quad + (0{,}5^4 \times 10) \\ \bar{x} &= 7{,}5 \end{aligned} \tag{11.111}$$

b. Calcularemos la varianza según

$$\sigma_x^2 = \sum p_i \times (x_i - \bar{x})^2 \tag{11.112}$$

donde x_i es el valor fenotípico para el i-ésimo genotipo y p_i su frecuencia (es decir, debemos ponderar al momento de calcular la varianza). Tenemos así

$$\sigma_x^2 = [0{,}5^4] \times (6 - 7{,}5)^2 + [2 \times 0{,}5^4] \times (7 - 7{,}5)^2 + [0{,}5^4] \times (8 - 7{,}5)^2 + \tag{11.113}$$

$$[2 \times 0{,}5^4] \times (7 - 7{,}5)^2 + [4 \times 0{,}5^4] \times (8 - 7{,}5)^2 + [2 \times 0{,}5^4] \times (9 - 7{,}5)^2 + \tag{11.114}$$

$$[0{,}5^4] \times (8 - 7{,}5)^2 + [2 \times 0{,}5^4] \times (9 - 7{,}5)^2 + [(0{,}5^4] \times (10 - 7{,}5)^2 \tag{11.115}$$

$$\sigma_x^2 \approx 1{,}9375 \Rightarrow \sigma_x = \sqrt{\sigma_x^2} \approx \sqrt{1{,}9375} \approx 1{,}3919 \tag{11.116}$$

c. Observemos que en todos los casos el cambio de un alelo por otro genera un cambio en el valor fenotípico del respectivo genotipo de 1. Esto es una señal clara de aditividad: si existieran efectos epistáticos deberíamos observar desviaciones debido a las interacciones entre loci. A su vez es también señal de que no existen efectos de dominancia (de lo contrario, los fenotipos heterocigotas no serían equidistantes de los homocigotas). En resumen, la variación fenotípica parece estar determinada, en su componente genético, solo por la componente aditiva.

d. Dado que toda la variación genética está dada por el componente aditivo, tenemos que en este caso $h^2 = 1$.

e. Para calcular la respuesta de selección podemos aplicar la ecuación del criador ($R = h^2 S$, donde $S = \bar{x}_s - \bar{x}_0$ es el diferencial de selección). Podemos calcular la media fenotípica para los genotipos seleccionados (es decir, $\bar{x}_S$) de forma análoga a lo realizado en la primera parte del ejercicio. Solo debemos tener en cuenta que las frecuencias genotípicas cambiarán por un factor común en todos los casos, ya que cambió la composición de la población (ya que descartamos individuos). La frecuencia total de los individuos seleccionados en la población original es de $f(aaBB)+f(AaBb)+f(aaBb)+f(AAbb)+f(Aabb)+f(aabb) = 0{,}5^4 + 4 \cdot 0{,}5^4 + 2 \times 0{,}5^4 + 0{,}5^4 + 2 \times 0{,}5^4 + 0{,}5^4 = 0{,}6785$. Podemos realizar un ajuste de las proporciones genotípicas en la nueva población con una regla de tres (o lo que es igual, aplicando un factor de $\frac{1}{0{,}6785}$. Las nuevas frecuencias serán de

$$\begin{cases} f(aaBB) = f(aaBB) \times \frac{1}{0{,}6785} \\ f(AaBb) = f(AaBb) \times \frac{1}{0{,}6785} \\ f(aaBb) = f(aaBb) \times \frac{1}{0{,}6785} \\ f(AAbb) = f(AAbb) \times \frac{1}{0{,}6785} \\ f(Aabb) = f(Aabb) \times \frac{1}{0{,}6785} \\ f(aabb) = f(aabb) \times \frac{1}{0{,}6785} \end{cases} \tag{11.117}$$

De esto sale que la media de los individuos seleccionados es de

$$\bar{x}_s = (0{,}5^4 \times \frac{1}{0{,}6785} \times 8) + (4 \times 0{,}5^4 \frac{1}{0{,}6785} \times 8) + \tag{11.118}$$

$$(2 \times 0{,}5^4 \frac{1}{0{,}6785} \times 9) + (0{,}5^4 \times \frac{1}{0{,}6785} \times 8) + \tag{11.119}$$

$$(2 \times 0{,}5^4 \times \frac{1}{0{,}6785} \times 9) + (0{,}5^4 \times \frac{1}{0{,}6785} \times 10) \tag{11.120}$$

$$\bar{x}_s \approx 8{,}6588 \tag{11.121}$$

Por lo tanto, $R \approx 1 \times (8{,}6588 - 7{,}5) = 1{,}1588$

Ejercicio 11.3

En un rodeo de bovinos, se tiene como objetivo aumentar el peso adulto ($h^2 = 0{,}45$). En este rodeo, el porcentaje de destete logrado es del 75 % y el desvío fenotípico de la característica es de 30 kg. La estructura del mismo se presenta a continuación:

Edad al parto	Hembras	Machos
3	210	10
4	190	8
5	120	6
6	60	
7	20	
TOTAL	600	24

¿Cuál es el progreso genético logrado al cabo de una generación?

Solución

En primer lugar, determinaremos la proporción de animales seleccionados

$600 \times 0{,}75 = 450$ animales nacidos de los cuales 225 serán hembras y 225 machos

Luego determinaremos la proporción de animales seleccionados de cada sexo

$$p_h = \frac{210}{225} \approx 0{,}93 = 93\,\% \tag{11.122}$$

$$p_m = \frac{10}{225} \approx 0{,}04 = 4\,\% \tag{11.123}$$

A partir de la tabla de intensidades podremos determinar que estas proporciones de animales seleccionados implican una intensidad de selección de 0,1444 en las hembras y de 2,1543 en los machos. Por lo tanto, la intensidad de selección poblacional es de $\frac{0{,}1444+2{,}1543}{2} = 1{,}14935$

Si recordamos que $i = \frac{S}{\sigma_P}$ entonces $S = i \times \sigma_P$. Entonces el diferencial de selección será $S = 1{,}14935 \times 30 = 34{,}48$ kg por lo que podremos calcular el progreso genético mediante la ecuación del criador tal que $PG = 0{,}45 \times 34{,}48 \approx 15{,}52$ kg.

Ejercicio 11.4

Un equipo se encuentra trabajando en el mejoramiento de dos características fenotípicas que poseen valores de heredabilidad muy dispares ($h_1^2 = 0{,}121$ y $h_1^2 = 0{,}932$). Si pretenden obtener una misma respuesta de selección para estos, ¿cuántas veces mayor será el diferencial de selección S_2 respecto al S_1?

Solución

Emplearemos la ecuación del criador ($R = h^2 S$). Si $R_1 = R_2$, esto implica que

$$h_1^2 S_1 = h_2^2 S_2 \Rightarrow \tag{11.124}$$

$$\frac{h_1^2}{h_2^2} = \frac{S_2}{S_1} \tag{11.125}$$

$$\frac{0{,}932}{0{,}121} = \frac{S_2}{S_1} \tag{11.126}$$

$$\frac{S_2}{S_1} = 7{,}7 \tag{11.127}$$

Vemos que el diferencial de selección debe ser casi 8 veces mayor. En efecto, debe compensar exactamente en la misma proporción la diferencia entre las heredabilidades (ya que ambos términos se encuentran multiplicando en la ecuación del criador).

Ejercicio 11.5

El objetivo de un criador de ovinos es la disminución del diámetro de la fibra, y para ello se propone realizar selección en los carneros a usar. En este sentido, anualmente se logran señalar 150 machos de los cuales el productor planea seleccionar con los 8 mejores para utilizar como carneros. La media de la población en cuanto a diámetro de fibra es de $28{,}4\mu$ y el registro de cada uno de los machos seleccionados fue de $18, 22, 17, 21, 20, 19, 20, 21\mu$. A su vez, se ha estimado que la heredabilidad de la característica es de 0,48.

a. ¿Cuánto progreso genético se logra al cabo de una generación? **b.** Si en lugar de seleccionar 8 machos se seleccionarán 5, ¿qué progreso genético se lograría?

Solución

a. En primer lugar deberemos calcular la media de los machos seleccionados

$$\bar{x} = \frac{18 + 22 + 17 + 21 + 20 + 19 + 20 + 21}{8} = 19{,}75 \tag{11.128}$$

A partir de la misma podremos calcular el diferencial de selección de los machos y el poblacional para el que deberemos tener en cuenta que no se está haciendo selección en las hembras

$$S_m = 19{,}75 - 28{,}45 = -8{,}7 \tag{11.129}$$

$$S_p = \frac{-8{,}7 + 0}{2} = -4{,}35 \tag{11.130}$$

Por último, podremos calcular el progreso genético

$$PG = h^2 \times S = 0{,}48 \times -4{,}35 = -2{,}088 \tag{11.131}$$

Es decir que cada generación es posible disminuir en algo más de 2μ el grosor de la fibra.

b. En el caso en que se seleccionaran 5 machos en lugar de 8, deberemos repetir el mismo procedimiento

$$\bar{x} = \frac{18 + 17 + 20 + 19 + 20}{5} = 18{,}8 \tag{11.132}$$

$$S_m = 18{,}8 - 28{,}45 = -9{,}65 \tag{11.133}$$

$$S_p = \frac{-9{,}65 + 0}{2} = -4{,}825 \tag{11.134}$$

$$PG = h^2 \times S = 0{,}48 \times -4{,}825 = -2{,}316 \tag{11.135}$$

Capítulo 12

Correlaciones y respuesta correlacionada

La probabilità che una certa persona sia stupida è indipendente da qualsiasi altra caratteristica della stessa persona.[1]

Carlo Cipolla

Hasta ahora, en la mayor parte de nuestro recorrido nos hemos limitado a examinar un solo fenotipo cada vez. Hemos construido modelos sobre el genotipo, descomponiendo su expresión en la suma de los aportes genéticos y ambientales y en algunos casos su interacción. Más aún, en el capítulo «Selección artificial I» estudiamos cómo sería la respuesta a nuestro esfuerzo de selección direccional en función de la presión de selección que ejercemos y de los parámetros genéticos asociados a la característica. Sin embargo, por lo general estamos interesados en mejorar más de un fenotipo en los individuos. En ovinos deseamos aumentar el peso del vellón, pero al mismo tiempo deseamos reducir el diámetro de las fibras. En producción lechera (en nuestro país) deseamos aumentar la producción de grasa y de proteína (en kg), pero al mismo tiempo deseamos reducir los litros de leche (ya que la industria castiga el volumen), reducir el recuento de células somáticas (al premiado por la industria) y al menos mantener el nivel de fertilidad del rodeo. En el capítulo «Selección artificial II» veremos diferentes estrategias para alcanzar estos objetivos múltiples, pero para llegar a eso primero debemos entender la base causal de las relaciones entre diferentes características fenotípicas.

En el presente capítulo definiremos los tipos de correlaciones fenotípica (observable), genética (aditiva) y ambiental, en un modelo genético simplificado, donde el ambiente extendido contendrá además de los efectos ambientales propiamente dichos, todos los efectos genéticos no-aditivos. Esto nos permitirá simplificar los análisis mediante la comprensión de las correlaciones genéticas y ambientales.

[1]Segunda Ley Fundamental de la estupidez humana, *Allegro ma non troppo*, Carlo Cipolla (1988). La traducción podría ser algo así como: «La probabilidad de que una determinada persona sea estúpida es independiente de cualquier otra característica de esa persona». En este caso, la referencia podemos asociarla a la correlación entre características con cierta determinación genética.

OBJETIVOS DEL CAPÍTULO

☐ Definir tipos de correlaciones y utilizarlas para analizar las correlaciones genéticas y ambientales.
☐ Explorar las causas subyacentes de estas correlaciones, incluyendo pleiotropía y ligamiento.
☐ Analizar la respuesta correlacionada en la selección de características.
☐ Introducir el análisis de coeficientes de paso (*path analysis*).
☐ Describir la relación entre correlaciones fenotípicas, genéticas y ambientales.
☐ Discutir métodos para estimar correlaciones genéticas y ambientales.
☐ Organizar la información genética, ambiental y fenotípica mediante matrices de varianza-covarianza y calcular parámetros genéticos a partir de estas.
☐ Generalizar la ecuación del criador a varias características (ecuación de Lande).
☐ Analizar la información proporcionada por la matriz de varianza-covarianza genética en la selección y las restricciones que la misma impone a la evolución de características.

12.1 Causas genéticas y ambientales de las correlaciones

El fenotipo, de acuerdo con nuestro modelo genético básico, lo hemos descompuesto en los aportes genéticos y los aportes ambientales, más (en algunos casos) los aportes de la interacción entre el genotipo y el ambiente. Por lo tanto, cuando hablamos de correlaciones fenotípicas entre características que observamos en los individuos, no es de extrañar que la base de estas correlaciones se pueda descomponer también en los aportes de la genética, es decir las correlaciones genéticas y los aportes ambientales y sus correspondientes correlaciones. En la presente sección vamos a discutir las bases biológicas y causales de estas correlaciones, mientras que en las siguientes secciones de este capítulo desarrollaremos la teoría para interpretarlas y cuantificarlas.

Correlación fenotípica entre dos características, medidas ambas en los mismos individuos **es la correlación entre los valores fenotípicos** (observables) de cada individuo en ambas características.

Causas de las correlaciones genéticas

Claramente, podemos calcular correlaciones entre cualquier par de mediciones que tengamos en los mismos individuos y por lo tanto podemos luego buscar las causas, genética y ambientales de los resultados obtenidos. Sin embargo, en algunos casos las características que pretendemos medir como dos características diferentes son apenas distinguibles entre sí, por lo que no es sorprendente que en esos casos la correlación genética se explique por sí sola. Por ejemplo, si pesamos a los individuos en una balanza que usa el Sistema Internacional

de Unidades (SI, kg) y luego las pesamos en otra balanza que utiliza las unidades del sistema imperial británico (libras), la correlación entre las dos características debería ser muy alta, particularmente si las balanzas funcionan bien y el intervalo entre las dos medidas realizadas es corto. Técnicamente son dos características distintas, peso en la primera balanza y peso en la segunda balanza, pero más allá de factores que podemos asociar al error (calibración y variaciones temporales mínimas, por ejemplo, una vaca que orinó entre medidas), la variable biológica de interés es exactamente la misma. Por lo tanto, la genética que controla el peso de los individuos es la misma y no hace falta buscar causas misteriosas. De hecho, si las estimaciones de la correlación genética entre ambas características nos dieran un valor bajo, deberíamos sospechar fuertemente de problemas serios en el estudio. Sin duda se trata de un caso extremo, pero en general resulta útil reflexionar sobre el origen posible de cualquier correlación entre características para las que obtengamos estimaciones de correlaciones fenotípicas, genéticas y ambientales.

Correlación genética (aditiva) entre dos características, medidas ambas en los mismos individuos **es la correlación entre los valores de cría** de cada individuo en ambas características.

Desde el **punto de vista genético** hay dos causas que claramente pueden estar por detrás de la correlación entre dos mediciones realizadas en individuos (muestreados al azar) de la población: **pleiotropía** y **ligamiento**. En el capítulo «Dinámica de dos loci» discutimos lo que acontecía cuando la proximidad física de los loci disminuía la probabilidad de recombinación entre los mismos. En ese caso, determinadas combinaciones de alelos (**haplotipos**) tienden a aparecer con mayor frecuencia de lo que sería esperable por azar **si no estuviesen «ligados»** (que sería el producto de las frecuencias de cada alelo). Si, por ejemplo, dos loci se encuentran fuertemente ligados y cada uno de ellos controla (genes de efecto mayor) una característica diferente, entonces valores específicos en una característica estarán asociados a valores específicos en otra característica, lo que genera correlación. Supongamos un caso sencillo, dos loci A y B, cada uno gobernando una característica diferente, con dos alelos cada uno: $A = 5$, $a = 1$ para la primera característica, $B = 4$, $b = 2$ para la segunda característica. Si los dos loci no se encuentran ligados, la segregación distribuirá los gametos en forma independiente para los dos loci y por lo tanto no existirá correlación de valores entre las dos características. Sin embargo, en el caso de que se encuentren totalmente ligados, supongamos que A va siempre con B y que a va siempre con b. En ese caso, solo tendremos dos haplotipos posibles: AB y ab. Sin pérdida de generalidad, supongamos que las frecuencias de A y B son iguales a $\frac{1}{2}$, por lo que los haplotipos AB y ab tendrán también frecuencias de $\frac{1}{2}$. Ahora, los posibles genotipos y sus valores para las dos características serán

Genotipo	Frecuencia	Caract. 1	Caract. 2	Dif. 1	Dif. 2	Prod. 1,2
AB/AB	$\frac{1}{4}$	$5+5=10$	$4+4=8$	4	2	8
AB/ab	$\frac{1}{2}$	$5+1=6$	$4+2=6$	0	0	0
ab/ab	$\frac{1}{4}$	$1+1=2$	$2+2=4$	-4	-2	8
		$\mu_1 = 6$	$\mu_2 = 6$	$\sigma_1^2 = 8$	$\sigma_2^2 = 2$	$\sigma_{12} = 4$

En el cuadro, **Dif. 1** y **Dif. 2** representan las diferencias de los valores de cada genotipo respecto a las medias correspondientes de cada característica (que están en la última línea

de la tabla, en la columna respectiva), mientras que **Prod. 1,2** se refiere al producto cruzado de esas diferencias.

Las varianzas de las características 1 y 2 y la covarianza entre ellas se calculan como de costumbre, es decir

$$\sigma_1^2 = \frac{1}{4}(4)^2 + \frac{1}{2}(0)^2 + \frac{1}{4}(-4)^2 = 4 + 0 + 4 = 8 \tag{12.1}$$

$$\sigma_2^2 = \frac{1}{4}(2)^2 + \frac{1}{2}(0)^2 + \frac{1}{4}(-2)^2 = 1 + 0 + 1 = 2 \tag{12.2}$$

$$\sigma_{12} = \frac{1}{4}8 + \frac{1}{2}0 + \frac{1}{4}8 = 2 + 0 + 2 = 4 \tag{12.3}$$

De acuerdo con la definición del coeficiente de correlación, podemos calcularla a partir de los datos de la tabla de arriba como

$$r_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2\,\sigma_2^2}} = \frac{4}{\sqrt{8 \times 2}} = \frac{4}{\sqrt{16}} = \frac{4}{4} = 1 \tag{12.4}$$

Es decir, en este caso, la correlación genética (aditiva) entre las dos características es igual a 1. Como no hemos invocado ningún otro fenómeno, solo el **ligamiento completo** entre estos dos loci, pese a que hemos tomado valores arbitrarios para los alelos, la posibilidad de que el ligamiento sea causa de la correlación genética queda asentado.

En el caso de la correlación inducida por el desequilibrio de ligamiento, se trata de un fenómeno de carácter temporal; excepto en las regiones cromosómica que no poseen recombinación, en la medida en que discurra el tiempo, dependiendo de la tasa de recombinación, en algún momento se producirán dichos eventos de recombinación entre los loci y por lo tanto el fenómeno de ligamiento tenderá a desaparecer, como vimos en la sección La evolución en el tiempo del desequilibrio de ligamiento del capítulo «Dinámica de dos loci». De hecho, como vimos en dicha sección, la tasa de decaimiento del ligamiento es aproximadamente exponencial, ya que para una tasa de recombinación r, el desequilibrio en el tiempo t (medido en generaciones) es igual a

$$D_t = (1 - r)^t D_0 \approx e^{-rt} D_0 \tag{12.5}$$

Este comportamiento exponencial del decaimiento en el desequilibrio de ligamiento (ver figura 5.4), lleva a que en unas pocas generaciones el efecto del mismo en la correlación genética prácticamente desaparezca.

El segundo fenómeno que puede explicar fácilmente la correlación genética entre dos características es la **pleiotropía.** Existen muchas definiciones de pleiotropía ya que el concepto ha ido cambiando con el tiempo, pero básicamente es el efecto que un locus produce en más de una característica fenotípica, sea esta entendida en el sentido tradicional o en el extendido que incluye, por ejemplo, la expresión génica (el transcriptoma). En la figura 12.1 se ilustra el mecanismo básico de la pleiotropía. La línea horizontal a trazos separa lo que sería lo determinado a nivel genómico de lo que observamos a nivel fenotípico (en el sentido tradicional, pesos, alturas, colores, diámetros de características, cantidad de productos, porcentaje de materias, etcétera), aunque esta distinción no se aplica a los **fenotipos de expresión génica.**

La característica fenotípica 1 se encuentra controlada desde el punto de vista genético por un solo gen, que es a su vez regulado por un regulador exclusivo para ese gen. Por lo tanto, **desde**

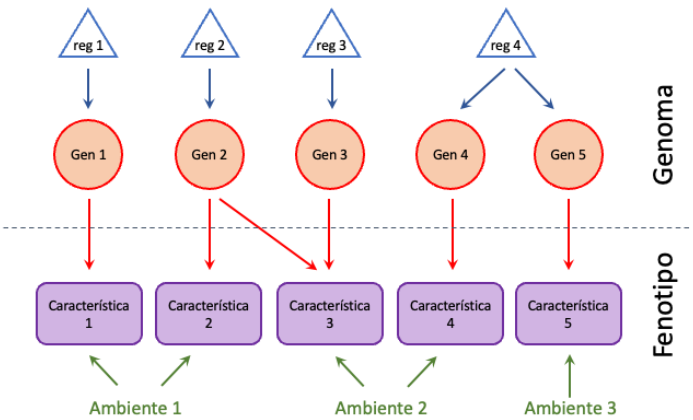


Figura 12.1. Ilustración de algunos mecanismos moleculares que generan pleiotropía. La característica 1 (rectángulos violetas) se encuentra controlada a nivel genético por un solo gen (círculos rojos), que es controlado a su vez por un regulador (triángulos azules) que no afecta ningún otro gen, por lo que su correlación genética con el resto de las características en el diagrama es cero. Las características fenotípicas 2 y 3 se encuentran al menos parcialmente controladas por el gen 2 (que tiene un solo regulador), por lo que este gen afecta simultáneamente la expresión de ambas características y es posible que aparezca una correlación genética. Finalmente, las características 4 y 5 se encuentran controladas a nivel genético por diferentes genes, pero ambos genes están regulados (posiblemente en forma parcial) por un solo regulador (el regulador 4), por lo que ambos genes tendrán expresiones correlacionadas y posiblemente se observe correlación entre las características fenotípicas. Los diferentes «ambientes» que influyen en las características pueden incidir en las correlaciones fenotípicas si son compartidos entre características, como, por ejemplo, el ambiente 1 y el ambiente 2.

el punto de vista genético no existe razón para que la pleiotropía induzca una correlación de la característica 1 con cualquiera de las otras. En cambio, mientras que la **característica 2** está controlada desde el punto de vista genético por un solo gen, el **gen 2** con su propio regulador, **la característica 3 está gobernada por dos genes**, el **gen 2** y el **gen 3**. Esto hace que la expresión del **gen 2** controle, al menos parcialmente, las dos características. Sin duda, es una buena razón para esperar que esta causa en común genere una correlación (de origen genético) en los datos fenotípicos observados. Por último, las características **4** y **5** están cada una de ellas gobernadas por un solo gen, distinto, porque podríamos pensar que no hay razón para que exista pleiotropía. Sin embargo, ambos genes, el **gen 4** y el **gen 5** comparten un mismo regulador (además de la posibilidad de que tengan otros reguladores), por lo que la expresión de ambos genes será una función del mismo regulador y por lo tanto tendrán expresiones correlacionadas, lo que, a su vez, posiblemente induzca una correlación a nivel fenotípico.

Como hemos discutido previamente en otros capítulos, son raras las características de interés cuantitativas gobernadas por un solo gen y no demasiadas las que parecen gobernadas por unos pocos genes de efecto mayor. Esto lleva a una dificultad importante a la hora de identificar el mecanismo preciso a nivel molecular y estimar sus efectos: existen pocas características cuantitativas gobernadas por muy pocos genes, donde se podrían identificar en forma relativamente sencilla los mecanismos de las interacciones. La mayoría están gobernadas por muchos genes, donde por lo tanto resulta casi imposible determinar todas las posibles interacciones entre los mismos (reguladores, inhibidores, etcétera) y mucho menos estimar sus efectos. Sin embargo, entre las correlaciones genéticas estudiadas hay muchas que son fuertes o moderadamente fuertes y que persisten a lo largo del tiempo, por lo que debemos suponer que existe un nivel importante de pleiotropía para estas.

Existen numerosos ejemplos de pleiotropía en características más cualitativas que cuantitativas (en muchos casos porque las mismas resultan difíciles de cuantificar, como, por ejemplo, con el grado de padecimiento de una enfermedad). Un ejemplo clásico de pleiotropía a nivel de enfermedades que se ven en humanos es la **fenilcetonuria** (también conocida como **PKU**), una alteración congénita del metabolismo causada por reducción o ausencia de la actividad en la enzima **fenilalanina hidroxilasa**. Esta última enzima es codificada por el gen PAH, localizado en el cromosoma 12 entre las bandas 12q22-q24.2 y tiene como cometido metabolizar el aminoácido **fenilalanina** (Phe), convirtiéndolo en otros metabolitos útiles para el cuerpo, es especial el aminoácido **tirosina** (Tyr). Se conocen más de 400 mutaciones en ese gen que producen una alteración en la capacidad metabólica del mismo, lo que lleva bajo una dieta normal a un incremento patológico en los niveles de fenilalanina. Estos niveles elevados de fenilalanina dañan el cerebro (usualmente en recién nacidos), provocando anormalidades en su funcionamiento, microencefalia, discapacidad intelectual, desórdenes de conducta, pero también otras manifestaciones físicas como decoloración de cabellos y piel, un olor muy particular y eccema. Claramente, el efecto genético de un gen (PAH) se manifiesta en diferentes características fenotípica, lo que es nuestra definición de pleiotropía. Afortunadamente, en nuestro país y en gran parte del mundo se les realiza a los recién nacidos un test que se conoce como **despistaje de enfermedades metabólicas** (y que en nuestro país incluye hipotiroidismo congénito, fenilcetonuria, hiperplasia suprarrenal congénita y fibrosis quística) por lo que detectada en forma temprana alcanza con un control estricto de la dieta para eliminar sus consecuencias.

Entre los animales domésticos se debate el rol que pudo (y puede haber tenido la pleiotropía) en los procesos de domesticación. De hecho, pese a que hemos domesticado a diferentes

especies, de grupos tan distantes como las aves y los mamíferos, en general determinadas características se repiten cuando comparamos la especie domesticada con la correspondiente salvaje (mansedumbre, tolerancia al humano, menor curiosidad, menor tamaño de las glándulas suprarrenales, etcétera). Sin duda que podemos apelar en algunos casos al ligamiento entre características y en algunos casos se han encontrado determinantes genéticos en el mismo entorno genómico, pero parece difícil de generalizar a todas las características. Sin embargo, por las dificultades que hemos mencionado previamente para identificar las bases moleculares de la pleiotropía en características cuantitativas, ha resultado una teoría con poco sustento (D. Wright 2015).

Un ejemplo de pleiotropía en animales es el de la sordera en gatos blancos de ojos azules, donde el 40 % de estos padecen la enfermedad. Una pista acerca de la base genética de esta enfermedad proviene de que los gatos con un ojo amarillo y el otro azul solo son sordos del lado del ojo azul. Se trata de una patología no exclusivamente observable en gatos, ya que, por ejemplo, el síndrome de Waardenburg (SW) en humanos es un trastorno que se caracteriza por sordera (en grados variables) y en anomalías en la pigmentación de los ojos, cabello y piel (todas ellas estructuras derivadas de la cresta neural). Otro ejemplo de pleiotropía en animales domésticos es la que produce el gen «frizzle» en aves, por ejemplo, la raza kirin de gallinas. En efecto, mutaciones en este gen provocan un curvado hacia afuera de las plumas, en lugar de su disposición paralela al cuerpo original. Pero además de ese llamativo fenotipo, dicho gen también incrementa la temperatura corporal, aumenta la tasa metabólica y el flujo de sangre, además de incrementar la capacidad digestiva y disminuir la producción de huevos (Lobo 2008).

Causas de las correlaciones ambientales

Muchas de las características fenotípicas de interés poseen, por varias razones, una base causal muy similar y que a veces subyace a varias características. Por ejemplo, todas aquellas características que aumentan su expresión con el tamaño de los animales tendrán una correlación derivada de esa base común. Resulta lógico que, dentro de la misma población y de la misma raza, ovejas más grandes (más pesadas) tiendan a producir una mayor superficie de cuero y un vellón mayor (más pesado) ya que tienen más superficie. Por lo tanto, para las características de superficie de cuero ovino y peso de vellón limpio esperamos una correlación fenotípica positiva. Pero ¿cuál es la base causal de esta correlación? Como lo expresamos antes, la causa está en que, a una densidad de fibras determinada, la superficie se vincula directamente con el número de fibras totales. Si nos ponemos más finos en la búsqueda de las causas, podríamos pensar que el tamaño (volumen) de los animales influencia directamente la superficie, ya que las formas de las ovejas son bastante similares (hay pocas ovejas esféricas, mucho menos cúbicas y no conocemos reportes de ovejas cónicas) y, por lo tanto, dentro de una misma raza y población las relación superficie/volumen es una función relativamente estable. De hecho, esta relación tiene que ver con el tamaño de los animales y con el clima donde viven, ya que el calor se genera proporcionalmente al volumen, mientras que se disipa en proporción al área. Como dato curioso, **algunos osos** que habitan en climas muy fríos son, desde este punto de vista **casi «esféricos»**, ya que la esfera es el cuerpo con menor relación superficie volumen.

Correlación ambiental entre dos características, medidas ambas en los mismos individuos, estrictamente **es la correlación entre los desvíos ambientales** que forman parte del valor fenotípico de cada individuo en ambas características. Usualmente, en la práctica se considera un ambiente extendido, que incluye **además del ambiente propiamente dichos los desvíos genéticos no-aditivos.**

Volvamos a nuestras ovejas. Por lo anterior, el tamaño de la oveja está determinando varias características fenotípicas que podemos medir. Lo mismo ocurre en vacas lecheras, donde el peso del animal se correlaciona con la producción de leche, grasa y proteína. Claramente, el tamaño de la vaca (de la misma raza y población) es un componente causal importante de todas estas características. A primera vista podríamos pensar que esta correlación es genética, ya que vacas más grandes y productivas producen crías más grandes y productivas. Sin embargo, un poquito de reflexión nos permite dudar de este argumento: el crecimiento de los animales depende de su alimentación también, es decir, animales mal alimentados no crecerán pese a tener el potencial. La alimentación la consideramos ambiente, por lo tanto, el ambiente influencia también el tamaño, que de alguna manera determina las características productivas que nos interesa medir (producción de leche, grasa y proteína, por ejemplo). Desvíos positivos ambientales (producto de mejor alimentación que el promedio) en una característica, también (en este caso) lo serán en la otra, por lo que la correlación ambiental será positiva.

En ovinos, mejor alimentación suele producir vellones con mayor peso, pero al mismo tiempo incrementar el diámetro de la lana. En plantas, mejores suelos mejorarán varias características de rendimiento al mismo tiempo, por lo que si realizamos mediciones de pares de características en plantas obtenidas de un número de sustratos diferentes vamos a tener una correlación de origen ambiental importante. De hecho, esto sugiere un método para estimar dicha correlación en plantas clonales, ya que al ser idéntica la genética, toda la correlación fenotípica observada será producto de las diferencias en el ambiente.

PARA RECORDAR

- Llamaremos **correlación fenotípica** entre dos características, medidas ambas en los mismos individuos a la correlación entre los valores fenotípicos (observables) de cada individuo en ambas características.
- Llamaremos **correlación genética (aditiva)** entre dos características, medidas ambas en los mismos individuos a la correlación entre los valores de cría de cada individuo en ambas características.
- Desde el punto de vista genético existen dos causas aparentes para que exista correlación entre características, ellas son la **pleitropía** y el **ligamiento**.
- En el **caso del ligamiento**, determinadas combinaciones de alelos (haplotipos) tienden a aparecer con mayor frecuencia de lo que sería esperable por azar si no estuviesen «ligados». En este caso, si por ejemplo dos loci se encuentran fuertemente ligados y cada uno de ellos controla (genes de efecto mayor) una característica diferente, entonces valores específicos en una característica estarán asociados a valores específicos en otra característica, lo que genera correlación.

- En el caso de la correlación inducida por el **desequilibrio de ligamiento**, se trata de un fenómeno de carácter temporal; excepto en las regiones cromosómica que no poseen recombinación, en la medida en que discurra el tiempo, dependiendo de la tasa de recombinación, en algún momento se producirán dichos eventos de recombinación entre los loci y por lo tanto el fenómeno de ligamiento tenderá a desaparecer.
- En el caso de la **pleitropía**, la misma se entiende como el efecto que un locus produce en más de una característica fenotípica, sea esta entendida en el sentido tradicional o en el extendido que incluye, por ejemplo, la expresión génica.
- Llamaremos **correlación ambiental** entre dos características, medidas ambas en los mismos individuos, a estrictamente la correlacción entre los desvíos ambientales que forman parte del valor fenotípico de cada individuo en ambas características. Usualmente, en la práctica se considera un ambiente extendido, que incluye además del ambiente propiamente dichos los desvíos genéticos no-aditivos.

12.2 Introducción al *path analysis*

En muchos casos podemos observar correlaciones entre variables (que por lo tanto son observables), pero sabemos que dichas correlaciones son producto de relaciones entre otras variables que las explican, pero que no son observables. De hecho, en muchos casos podemos establecer modelos claros con las relaciones entre las distintas variables, tanto las observables como las no-observables. El objetivo pasa a ser entonces entender y estimar los coeficientes del modelo. Con eso en mente, Sewall Wright (1918, 1934b) desarrolló el método conocido como de los **coeficientes de paso** (*path coefficients* en inglés).

La idea es sencilla, como se ilustra en la figura 12.2. Supongamos que tenemos una serie de variables, representadas con letras mayúsculas en la figura. Tenemos dos tipos básicos de relaciones entre variables: *a)* las **causales**, que vamos a representar con flechas rectas y con una sola punta, que apuntan desde la variable que explica a la explicada y *b)* las **no-causales**, que representan asociación de algún tipo entre variables (posiblemente debido a una causa común, no modelada), que vamos a representar con flechas curvas de dos puntas. A las variables que **no reciben ninguna flecha recta** (es decir, que no son explicadas por otras en el modelo) se las conoce como **exógenas**, mientras que **las que reciben alguna flecha recta** se las conoce como **endógenas**. Las flechas rectas representan, en un modelo lineal, variables que explican la variable a la que apuntan y por lo tanto, cuando trabajamos con correlaciones entre variables (es decir, variables estandarizadas), los coeficientes correspondientes serán los **coeficientes de regresión estandarizados**. Los coeficientes correspondientes a las flechas curvas de dos puntas serán simplemente los coeficientes de correlación entre las dos variables, marcando una situación de simetría que no tienen las flechas rectas.

De hecho, al observar con atención la figura 12.2 vemos que hay un «tercer tipo» de flecha, representadas en color verde, que son flechas rectas pero que no salen desde ninguna variable específica. Estas flechas se usan para representar la variación residual, es decir, la no explicada por el modelo. En ciertos casos estas flechas son omitidas de los diagramas,

pero deben ser tenidas en cuenta a la hora de entender la varianza explicada por las distintas variables del modelo. En algunos casos, como veremos más adelante, el residuo se modela específicamente (por ejemplo, dentro del ambiente), por lo que este tipo de flechas no aparece en dichos diagramas.

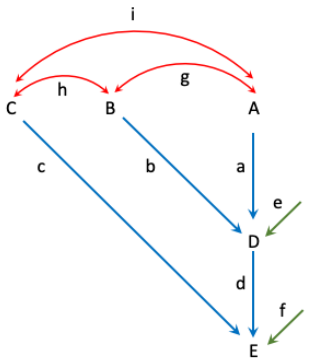


Figura 12.2. Ejemplo de un diagrama de caminos (o de paso), ilustrando las relaciones entre variables. Las líneas rectas con una sola flecha (marcadas en azul aquí para mayor claridad) representan relaciones «causales» entre variables, apuntando desde la que explica a la explicada, mientras que las flechas curvas con dos puntas (en rojo) explican asociaciones de correlación entre variables.

De acuerdo con Wright, si un diagrama de caminos está correctamente representado, entonces **la correlación entre dos variables** puede representarse como **la suma de caminos compuestos**, en los que **para cada camino se multiplican los coeficientes que lo integran**, que siguen tres reglas:

1. **No hay bucles** (es decir, no se puede pasar dos veces por la misma variable en un camino).
2. **No se puede avanzar y luego retroceder** (para una correlación solo se utilizan las causas comunes, no los efectos comunes).
3. Solo se admite un **máximo de una flecha curva por trayectoria.**

Veamos algún ejemplo, utilizando la figura 12.2 antes de pasar a la aplicación en genética que es de nuestro interés. La correlación entre A y B está determinada por el coeficiente correspondiente al único camino que cumple las tres reglas, es decir g. Lo mismo, para cualquiera de los pares de variables exógenas (A, B y C), con los coeficientes h e i. Si queremos ahora calcular la correlación entre A y D, que denotaremos como r_{AD}, tenemos un camino directo entre A y D, con coeficiente a y otro camino indirecto, desde A hacia B y luego hacia D, es decir, con coeficientes g y b (que deben multiplicarse). Sumando los dos caminos, tenemos entonces que

$$r_{AD} = a + b \times g \tag{12.6}$$

Pongámosles números a los coeficientes y veamos si podemos calcular alguna correlación más compleja. Supongamos que los coeficientes estimados (de alguna forma) son los siguientes: $a = 0{,}3$; $b = 0{,}1$; $c = 0{,}4$; $d = 0{,}2$; $e = 0{,}15$; $f = 0{,}25$; $g = 0{,}05$; $h = 0{,}05$; $i = 0{,}15$. Vamos a calcular la correlación esperada entre A y E. De acuerdo con las tres reglas que vimos antes, los caminos posibles son $a \times d$, $g \times b \times d$ y $i \times c$. Por lo tanto, la correlación buscada será

$$r_{AE} = a \times d + g \times b \times d + i \times c = 0{,}3 \times 0{,}2 + 0{,}05 \times 0{,}1 \times 0{,}2 + 0{,}15 \times 0{,}40 = 0{,}121 \tag{12.7}$$

La idea es bastante sencilla, aunque la estimación estadística de los coeficientes suele ser relativamente compleja y la especificación de los modelos muy importante. Al mismo tiempo que no deben faltar variables que sean relevantes, ya que sesgarían los estimados de los coeficientes, el agregar flechas innecesarias vuelve inestable la estimación de los coeficientes y por lo tanto representa modelos poco fiables a la hora de extraer conclusiones de los mismos. Existen muchos tratados sobre el tema, pero un excelente tratamiento del método y sus aplicaciones se encuentra en Loehlin y Beaujean (2017).

La relación entre las correlaciones fenotípica, genética y ambiental

La pregunta que se nos viene en mente es cómo se relaciona esto con la genética, y en particular con las correlaciones fenotípicas, genéticas y ambientales que trataremos en este capítulo. Un excelente y extenso tratamiento del tema se puede ver en el libro Li (1955), que se recomienda como lectura suplementaria. Como vimos previamente en el capítulo «El modelo genético básico», por lo general solo observamos los fenotipos de los individuos, no los diferentes efectos genéticos y ambientales. Pero como vimos antes también, la correlación que observamos entre dos características medidas en cada uno de los individuos de una población posee, en general, una base genética y una base ambiental. La tentación inicial sería entonces de suponer (de forma errónea) que la correlación fenotípica es la suma de las correlaciones genética y ambiental. Sin embargo, tenemos una situación un poco más compleja y para modelarla vamos a utilizar el método de los coeficientes de paso desarrollado por Sewall Wright.

Para simplificar nuestro modelo, supongamos que solamente vamos a modelar el fenotipo de cada característica como la suma de los efectos aditivos y el resto de los componentes, que llamaremos «ambiente», pero que tiene también los componentes genéticos no-aditivos y los errores de medición. Es decir, si llamamos E^* al ambiente propiamente dicho (más los errores de medición, entonces los fenotipos de las dos características que consideramos los modelamos como

$$\begin{aligned} P_1 &= A_1 + E_1;\ E_1 = D_1 + I_1 + E_1^* \\ P_2 &= A_2 + E_2;\ E_2 = D_2 + I_2 + E_2^* \end{aligned} \tag{12.8}$$

En la figura 12.3 podemos observar una representación del diagrama de paso que representa la relación entre las variables correspondientes a las dos características (por ejemplo, entre

Peso de Vellón Limpio y Diámetro de Fibra, entre Peso al Destete y Peso a los 18 meses, etcétera). De acuerdo con nuestro modelo especificado en las ecuaciones (12.8), no existe interacción entre genotipo y ambiente, por lo que estas variables (y los caminos que las conectan) serán independientes. La línea horizontal a trazos representa la separación entre lo observable y lo no observable. **La flecha curva a trazos *no* se especifica** normalmente en los diagramas ya que precisamente dicha correlación es la que el diagrama intenta explicar (con las otras variables); por dicha razón aparece representada a trazos y algo separada.

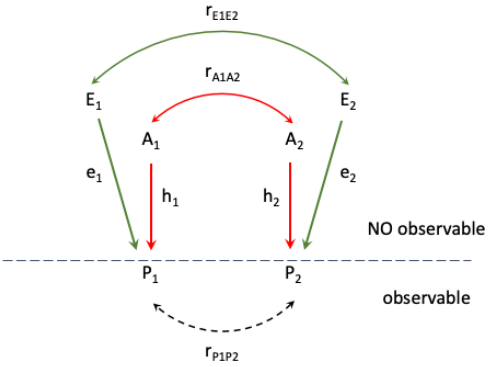


Figura 12.3. Método de los «coeficientes de paso» para determinar las relaciones entre los componentes causales de la correlación fenotípica entre dos características en el mismo individuo. La línea a trazos horizontal representa la separación entre lo observable y lo no-observable. Formalmente, el diagrama de pasos no incluye la curva a trazos representando $r_{P_{12}}$ (la correlación fenotípica entre las variables). En este modelo el «ambiente» E incluye todos los efectos excepto los genéticos aditivos, es decir $E = I + D + E^*$ (con E^* el «ambiente» propiamente dicho, incluyendo los errores de medición).

Existen entonces dos caminos independientes que debemos sumar, el genético (aditivo) y el ambiental. Para ello necesitamos obtener los coeficientes asociados a cada uno de los pasos. Los coeficientes de las flechas curvas son precisamente las correlaciones genética y ambiental, que dependerán de cada característica y que aún no sabemos cómo estimar (lo veremos más adelante), pero que a los fines de esta sección (relación entre correlaciones) nos alcanza con dejar así. Por otra parte, si recordamos de antes que los coeficientes de las líneas rectas eran los coeficientes de regresión estandarizados, debemos entonces determinar su valor.

Primero procederemos con los coeficientes de paso que unen a las variables $A_1 \rightarrow P_1$ y $A_2 \rightarrow P_2$. De acuerdo con la definición de coeficiente de regresión estandarizado, $\beta^*_{yx} = b_{yx}\frac{\sigma_x}{\sigma_y}$, si **el fenotipo es la variable dependiente** y **el valor de cría la independiente** (es decir, la que explica parcialmente el fenotipo), el correspondiente coeficiente de regresión estandarizado sería

$$\beta^*_{PA} = b_{PA}\frac{\sigma_A}{\sigma_P} = \frac{Cov(A,P)}{\sigma^2_A}\frac{\sigma_A}{\sigma_P} = \frac{\sigma^2_A}{\sigma^2_A}\frac{\sigma_A}{\sigma_P} = \frac{\sigma_A}{\sigma_P} = h \qquad (12.9)$$

Por lo tanto, el coeficiente de paso para $A_1 \rightarrow P_1$ será h_1 y el correspondiente a $A_2 \rightarrow P_2$ será h_2.

Es conocido (de los cursos previos de estadística) que la proporción de la varianza explicada por una variable en la regresión lineal es el coeficiente de regresión estandarizado elevado al cuadrado (conocido como R^2, no confundir esta R con la respuesta a la selección para la que también usamos R, o con la repetibilidad, que en algunos textos también utilizan la misma letra), por lo que lo explicado por los efectos genéticos aditivos sería, para la característica 1, $(h_1)^2 = h_1^2$, es decir la heredabilidad. Como hay solo dos variables independientes para cada fenotipo (lo genético-aditivo y el resto, ambiente más genética no-aditiva, incluyendo el error), entonces el otro componente debe explicar el resto de la varianza, es decir

$$(h_1)^2 + (e_1)^2 = 1 \therefore e_1^2 = 1 - h_1^2 \therefore e_1 = \sqrt{1 - h_1^2} \tag{12.10}$$

De la misma manera, para la segunda característica

$$(h_2)^2 + (e_2)^2 = 1 \therefore e_2^2 = 1 - h_2^2 \therefore e_2 = \sqrt{1 - h_2^2} \tag{12.11}$$

Aplicando la reglas de Wright, solo hay dos caminos posibles para explicar la correlación entre P_1 y P_2, el camino formado por $h_1\ r_{A_{12}}\ h_2$ y el camino formado por $e_1\ r_{E_{12}}\ e_2$ y por lo tanto debemos sumarlos. Por lo tanto (reordenando los términos en cada camino), la correlación fenotípica observada es igual a

$$r_{P_{12}} = h_1 \times h_2 \times r_{A_{12}} + e_1 \times e_2 \times r_{E_{12}} \tag{12.12}$$

Ejemplo 12.1

En las **ovejas verdes** se observa una correlación (fenotípica) entre el Peso al Año de edad (PA) y el Peso de Vellón Sucio (PVS) de $r_P = 0{,}70$ (no vamos a usar los subíndices $_{12}$ en las correlaciones para simplificar la notación). Las heredabilidades de ambas características fueron estimadas en $h^2_{PVS} = 0{,}49$ y $h^2_{PA} = 0{,}36$. Por otra parte, un estudio reciente determinó que el mejor estimado de la correlación genética (aditiva) entre ambas características es de $r_A = 0{,}55$. Estimar la correlación ambiental (en sentido amplio) y discutir los resultados.

De acuerdo con la ecuación (12.12), la relación entre las correlaciones fenotípica, genética (aditiva) y ambiental es

$$r_P = h_1 \times h_2 \times r_A + e_1 \times e_2 \times r_E \tag{12.13}$$

Por lo tanto, si despejamos la correlación ambiental, tenemos que

$$r_{E_{12}} = \frac{r_{P_{12}} - h_1 \times h_2 \times r_{A_{12}}}{e_1 \times e_2} \tag{12.14}$$

Los coeficientes de paso genéticos se calculan directamente haciendo la raíz cuadrada de las correspondientes heredabilidades, es decir

$$\begin{aligned} h_{PA} &= \sqrt{h^2_{PA}} = \sqrt{0{,}36} = 0{,}6 \\ h_{PVS} &= \sqrt{h^2_{PVS}} = \sqrt{0{,}49} = 0{,}7 \end{aligned} \tag{12.15}$$

Por otra parte, los coeficientes de paso ambientales se calculan utilizando los resultados de las ecuaciones (12.10) y (12.11)

$$\begin{aligned} e_{PA} &= \sqrt{1 - h^2_{PA}} = \sqrt{1 - 0{,}36} = \sqrt{0{,}64} = 0{,}80 \\ e_{PVS} &= \sqrt{1 - h^2_{PVS}} = \sqrt{1 - 0{,}49} = \sqrt{0{,}51} \approx 0{,}71 \end{aligned} \tag{12.16}$$

Ahora, poniendo estos valores en la ecuación para $r_{E_{12}}$, tenemos

$$r_E = \frac{r_P - h_{PA} \times h_{PVS} \times r_A}{e_{PA} \times e_{PVS}} = \frac{0{,}70 - 0{,}6 \times 0{,}7 \times 0{,}55}{0{,}80 \times 0{,}71} \approx 0{,}83 \tag{12.17}$$

Por lo tanto, observamos que tanto la correlación fenotípica, como la genética (aditiva) y la ambiental son todas positivas y de carácter importante. La razón de la correlación genética entre las dos características consideradas es bastante esperable (dentro de la misma raza y población) ya que animales más pesados suelen ser más grandes y por lo tanto tener mayor superficie para generar lana. Se trata de una relación que estaría fundamentada, de alguna manera, en la pleiotropía, ya que los genes que regulan el crecimiento afectan a ambas características. Al mismo tiempo, la correlación ambiental también parece esperable, por similares razones a las de la correlación genética: animales que crecen más (por factores ambientales), van a producir medidas más altas en las dos características.

PARA RECORDAR

- Haciendo una introducción al *path analysis*, de acuerdo con Wright, si un diagrama de caminos está correctamente representado, entonces la **correlación entre dos variables puede representarse como la suma de caminos compuestos**, en los que para cada camino se multiplican los coeficientes que lo integran, que siguen tres reglas: 1) No hay bucles (es decir, no se puede pasar dos veces por la misma variable en un camino); 2) No se puede avanzar y luego retroceder (para una correlación solo se utilizan las causas comunes, no los efectos comunes); 3)Solo se admite un máximo de una flecha curva por trayectoria.
- En el proceso de entender la relación entre las correlaciones fenotípica, genética y ambiental, **supongamos que solamente vamos a modelar el fenotipo de cada característica como la suma de los efectos aditivos y el resto de los componentes (que llamaremos «ambiente»)**, pero que tiene también los componentes genéticos no-aditivos y los errores de medición. Es decir, si llamamos E^* al ambiente propiamente dicho (más los errores de medición, entonces los fenotipos de las dos características que consideramos los modelamos como
  $P_1 = A_1 + E_1;\ E_1 = D_1 + I_1 + E^*_1$
  $P_2 = A_2 + E_2;\ E_2 = D_2 + I_2 + E^*_2$
- Relacionando todas las variables, podemos decir que la correlación fenotípica observada es $r_{P_{12}} = h_1 \times h_2 \times r_{A_{12}} + e_1 \times e_2 \times r_{E_{12}}$

12.3 Métodos para determinar la correlación genética

Hasta acá hemos discutido acerca de los distintos tipos de correlaciones, fenotípica, genética (aditiva) y ambiental, sin preocuparnos demasiado en su determinación. De hecho, la correlación fenotípica se puede calcular directamente de las observaciones. Por otra parte, también vimos la relación entre las tres correlaciones, por lo que si contamos con dos de ellas y las heredabilidades también podemos estimar la faltante. Más aún, veremos más adelante cómo calcular la correlación genética a partir de la matriz de varianza-covarianza genética **G**, pero no discutimos cómo llegamos a la determinación de los elementos de la misma (las varianzas y covarianzas genéticas).

Históricamente, los métodos experimentales que se usaron para estimar la correlación genética fueron los mismos que se usaron para determinar las heredabilidades de las características. De hecho, con un mínimo de planificación, el mismo experimento permite obtener todas las estimaciones de interés. En la actualidad, la mayor parte de las estimaciones se realizan a través de modelos mixtos lineales en su forma multivariada, usando métodos iterativos sobre la versión multivariada del BLUP, propuesta inicialmente por C. R. Henderson y Quaas (1976) y que veremos en el capítulo «Selección artificial II».

De cualquier manera, simplemente como forma de comprender las bases de la estimación vamos a ver uno de los diseños experimentales posibles. La idea es obtener estimaciones a través de un diseño del tipo **progenitor-progenie**. Para eso vamos a medir las dos características en todas las parejas de progenitor-progenie. Vamos a llamar X a los progenitores e Y a las progenies, mientras que vamos a utilizar los subíndices 1 y 2 para las dos características.

De acuerdo con la notación anterior, vamos a obtener las siguientes 4 covarianzas:

- $Cov_{X_1Y_2}$ es la covarianza de la característica 1 en progenitor con la característica 2 en la progenie
- $Cov_{X_2Y_1}$ es la covarianza de la característica 1 en progenitor con la característica 2 en la progenie
- $Cov_{X_1Y_1}$ es la covarianza de la característica 1 en progenitor con la misma característica en la progenie
- $Cov_{X_1Y_1}$ es la covarianza de la característica 2 en progenitor con la misma característica en la progenie

La fórmula para las covarianzas observadas, para N parejas de progenitor progenie es

$$Cov_{XY} \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N-1} \tag{12.18}$$

Desde el punto de vista causal, las **covarianzas cruzadas** (una característica en el progenitor y la otra, diferente, en la progenie) se pueden describir como

$$Cov_{X_1Y_2} = Cov_{X_2Y_1} = \frac{1}{2}Cov_{A_{12}} + \frac{1}{4}Cov_{AA_{12}} + \ldots \tag{12.19}$$

es decir, un medio de la covarianza aditiva entre las características 1 y 2, un cuarto de la covarianza epistática del tipo $A \times A$, etcétera.

Por otro lado, como vimos en la sección «Semejanza entre parientes», las covarianzas fenotípicas progenitor-progenie estiman un medio de varianza aditiva, por lo que

$$\begin{aligned} Cov_{X_1Y_1} &= \frac{1}{2}V_{A_1} \\ Cov_{X_2Y_2} &= \frac{1}{2}V_{A_2} \end{aligned} \tag{12.20}$$

Por lo tanto, si despreciamos en la ecuación (12.19) los términos epistáticos, un estimador razonable de la correlación genética entre las dos características sería

$$\hat{r}_{A_{12}} = \frac{Cov_{X_1Y_2} + Cov_{X_2Y_1}}{2\sqrt{Cov_{X_1Y_1} \times Cov_{X_2Y_2}}} \tag{12.21}$$

Siempre es importante recordar que se trata de un estimador del «verdadero» valor de un parámetro genético y, por lo tanto, la estimación es función de una o más variables aleatorias, es decir, con incertidumbre asociada a la estimación. Aunque la derivación escapa el alcance de la presente sección (Robertson 1959; Tallis 1959), un estimador aproximado del error estándar ($EE(r_A)$) del estimador de la correlación genética es

$$EE(r_A) \approx (1 - r_A^2)\sqrt{\frac{EE(h_1^2)\ EE(h_2^2)}{2\ h_1^2\ h_2^2}} \tag{12.22}$$

Un diseño alternativo es el de medios hermanos, donde un padre (de I disponibles) es apareado con varias (K) madres. Con el cambio de notación de que ahora X e Y representan las dos características de interés, de acuerdo con el análisis de varianza-covarianza para este diseño, tenemos la siguiente tabla de las esperanzas de cuadrados medios y del producto medio:

Factor	g.l.	$E[CM_X]$	$E[CM_Y]$	$E[PM_{XY}]$
Intergrupos	$I1$	$\sigma^2_{WX} + K\sigma^2_{BX}$	$\sigma^2_{WY} + K\sigma^2_{BY}$	$cov_W + Kcov_B$
Error	$I(K-1)$	σ^2_{WX}	σ^2_{WY}	cov_W

A su vez, desde el punto de vista de los componentes causales, tenemos que

$$\begin{aligned} \sigma^2_{BX} &= \frac{V_{A_X}}{4} \\ \sigma^2_{BY} &= \frac{V_{A_Y}}{4} \\ cov_B &= \frac{Cov_{A_{XY}}}{4} \end{aligned} \tag{12.23}$$

Por lo tanto, de acuerdo con la definición del coeficiente de correlación genético-aditiva, si dividimos la covarianza entre grupos de la ecuación (12.23) entre la raíz cuadrada del producto de las varianzas entre grupos, tenemos

$$r_{A_{XY}} = \frac{cov_B}{\sigma_{BX}\ \sigma_{BY}} = \frac{\frac{Cov_{A_{XY}}}{4}}{\sqrt{\frac{V_{A_X}}{4}\frac{V_{A_Y}}{4}}} = \frac{Cov_{A_{XY}}}{\sigma_{A_X}\sigma_{A_Y}} \tag{12.24}$$

que es el estimador de la correlación genético-aditiva para este **diseño de medios hermanos.**

Un tratamiento mucho más detallado, así como diferentes estimados de correlaciones genéticas para diferentes características de interés productivo se encuentra en Cardellino y Rovira (1987). También, en la sección «Respuesta correlacionada» veremos un procedimiento sencillo para estimar la correlación genética entre dos características, basado en la medición simultánea de la respuesta directa y la respuesta correlacionada.

Ejemplo 12.2

La morfometría craneal en ovinos es un área de investigación en expansión. En un experimento para determinar la correlación genética entre el largo de la sutura frontal (**Lsf**) y el largo del hueso parietal (**Lhp**) en **ovejas verdes**, se midieron ambas variables en un conjunto de 100 cráneos, obtenidos de 50 parejas madre-hija.

Los resultados del experimento se pueden resumir de la siguiente manera

Covarianza	Valor
Lsf en la madre con **Lhp** en hija	0,33 cm^2
Lhp en la madre con **Lsf** en hija	0,29 cm^2
Lsf en la madre con **Lsf** en hija	0,36 cm^2
Lhp en la madre con **Lhp** en hija	0,39 cm^2

Estimar la covarianza genética entre el largo de la sutura frontal (**Lsf**) y el largo del hueso parietal (**Lhp**).

Utilizando la ecuación (12.21), tenemos que un estimador de la correlación genética a partir de estos datos está dado por

$$\hat{r}_{A_{12}} = \frac{Cov_{X_1Y_2} + Cov_{X_2Y_1}}{2\sqrt{Cov_{X_1Y_1} \times Cov_{X_2Y_2}}} \tag{12.25}$$

Sustituyendo los valores, tenemos que

$$\hat{r}_{A_{12}} = \frac{0{,}33\ cm^2 + 0{,}29\ cm^2}{2\sqrt{0{,}36\ cm^2 \times 0{,}39\ cm^2}} = 0{,}8273 \tag{12.26}$$

Es decir, una correlación positiva e importante entre ambas medidas craneales.

PARA RECORDAR

- Mediante un diseño de **progenitor-progenie** donde llamaremos X a los progenitores e Y a las progenies y utilizando los subíndices 1 y 2 para las dos características, un estimador razonable de la correlación genética entre ambas sería $\hat{r}_{A_{12}} = \frac{Cov_{X_1Y_2} + Cov_{X_2Y_1}}{2\sqrt{Cov_{X_1Y_1} \times Cov_{X_2Y_2}}}$
- Al mismo tiempo, un estimador aproximado del error estándar del estimador de la correlación genética sería $EE(r_A) \approx (1 - r_A^2)\sqrt{\frac{EE(h_1^2)\ EE(h_2^2)}{2\ h_1^2\ h_2^2}}$
- Si por otra parte utilizáramos un diseño de **medios hermanos**, el estimador de la correlación genético-aditiva para ese diseño sería $r_{A_{XY}} = \frac{Cov_{A_{XY}}}{\sigma_{A_X}\sigma_{A_Y}}$

12.4 La correlación fenotípica y su relación con otras correlaciones

Antes, en la sección «Introducción al *path analysis*», dedujimos la relación entre las correlaciones fenotípica, genética y ambiental usando esa herramienta poderosa. Ahora vamos a arribar al mismo resultado partiendo desde otra perspectiva y discutiremos algunas restricciones asociadas a la relación entre las tres correlaciones y las heredabilidades de las dos características.

Nuevamente, como hicimos antes, usaremos una versión simplificada de nuestro modelo genético básico, $P = A + E$, con $E = D + I + E^*$ y E^* representando el ambiente propiamente dicho, más los errores de medición. La simplificación anterior significa que **nuestro fenotipo ahora solo lo dividimos** en el componente **genético aditivo** y **el resto**, que notamos como E en este contexto, pero que comprende tanto lo ambiental como lo genético no-aditivo. De acuerdo con esto, a partir de las reglas que ya conocemos para las covarianzas y asumiendo que no existe covariación entre el componente aditivo de un individuo y el ambiente (extendido) en el otro,

$$\begin{aligned} P &= A + E \\ Cov(P_1,P_2) = Cov(A_1 + E_1, A_2 + E_2) &= Cov(A_1,A_2) + Cov(E_1,E_2) \end{aligned} \tag{12.27}$$

A su vez, a partir de la definición de coeficiente de correlación,

$$r_{P_{12}} = \frac{Cov(P_1,P_2)}{\hat{\sigma}_{P_1}\ \hat{\sigma}_{P_2}} \iff Cov(P_1,P_2) = r_{P_{12}}\ \hat{\sigma}_{P_1}\ \hat{\sigma}_{P_2} \tag{12.28}$$

$$r_{A_{12}} = \frac{Cov(A_1,A_2)}{\hat{\sigma}_{A_1}\ \hat{\sigma}_{A_2}} \iff Cov(A_1,A_2) = r_{A_{12}}\ \hat{\sigma}_{A_1}\ \hat{\sigma}_{A_2} \tag{12.29}$$

$$r_{E_{12}} = \frac{Cov(E_1,E_2)}{\hat{\sigma}_{E_1}\ \hat{\sigma}_{E_2}} \iff Cov(E_1,E_2) = r_{E_{12}}\ \hat{\sigma}_{E_1}\ \hat{\sigma}_{E_2} \tag{12.30}$$

Si asumimos que la heredabilidad representa la partición de la información que corresponde a los efectos aditivos, podemos definir un coeficiente de «ambientalidad» (que claramente no representa el ambiente, sino todo lo que no es heredable), que sumado a la heredabilidad debe dar 1 (el todo)

$$h_1^2 + e_1^2 = 1 \iff e_1^2 = 1 - h_1^2 \iff e_1 = \sqrt{1 - h_1^2} \tag{12.31}$$

$$h_2^2 + e_2^2 = 1 \iff e_2^2 = 1 - h_2^2 \iff e_2 = \sqrt{1 - h_2^2} \tag{12.32}$$

$$h_1 = \sqrt{h_1^2} = \sqrt{\frac{\sigma_{A_1}^2}{\sigma_{P_1}^2}} = \frac{\sigma_{A_1}}{\sigma_{P_1}} \tag{12.33}$$

$$h_2 = \sqrt{h_2^2} = \sqrt{\frac{\sigma_{A_2}^2}{\sigma_{P_2}^2}} = \frac{\sigma_{A_2}}{\sigma_{P_2}} \tag{12.34}$$

$$r_{P_{12}} = r_{A_{12}}\, h_1\, h_2 \; + r_{E_{12}}\, e_1\, e_2 \tag{12.35}$$

Ahora, sustituyendo las covarianzas en la parte de antes por su correspondiente producto del coeficiente de correlación multiplicado por los desvíos estándar, tenemos

$$Cov(P_1,P_2) = Cov(A_1,A_2) + Cov(E_1,E_2) \iff \tag{12.36}$$

$$r_{P_{12}}\, \hat{\sigma}_{P_1}\, \hat{\sigma}_{P_2} = r_{A_{12}}\, \hat{\sigma}_{A_1}\, \hat{\sigma}_{A_2} + r_{E_{12}}\, \hat{\sigma}_{E_1}\, \hat{\sigma}_{E_2} \tag{12.37}$$

$$r_{P_{12}} = \frac{r_{A_{12}}\, \hat{\sigma}_{A_1}\, \hat{\sigma}_{A_2} + r_{E_{12}}\, \hat{\sigma}_{E_1}\, \hat{\sigma}_{E_2}}{\hat{\sigma}_{P_1}\, \hat{\sigma}_{P_2}} = \frac{r_{A_{12}}\, \hat{\sigma}_{A_1}\, \hat{\sigma}_{A_2}}{\hat{\sigma}_{P_1}\, \hat{\sigma}_{P_2}} + \frac{r_{E_{12}}\, \hat{\sigma}_{E_1}\, \hat{\sigma}_{E_2}}{\hat{\sigma}_{P_1}\, \hat{\sigma}_{P_2}} \tag{12.38}$$

$$r_{P_{12}} = r_{A_{12}}\, \frac{\hat{\sigma}_{A_1}}{\hat{\sigma}_{P_1}}\, \frac{\hat{\sigma}_{A_2}}{\hat{\sigma}_{P_2}} \; + r_{E_{12}}\, \frac{\hat{\sigma}_{E_1}}{\hat{\sigma}_{P_1}}\, \frac{\hat{\sigma}_{E_2}}{\hat{\sigma}_{P_2}} \tag{12.39}$$

Lo que nos lleva, finalmente a la relación

$$r_{P_{12}} = r_{A_{12}}\, h_1\, h_2 + r_{E_{12}}\, e_1\, e_2 \tag{12.40}$$

Claramente, la correlación fenotípica no solo depende de las correlaciones genética (aditiva) y ambiental (extendida), sino que también depende de las exactitudes de las dos características (o de sus heredabilidades). Cuanto mayores sean las heredabilidades, mayor será la influencia relativa de la correlación genética y viceversa para la ambiental.

Una perspectiva interesante de la correlación fenotípica es analizar el comportamiento de la respuesta en la ecuación (12.40) al variar alguno de los parámetros. Un problema mayor es que el número de parámetros a variar no es pequeño; tenemos dos correlaciones, más dos heredabilidades (o sus raíces cuadradas). Notemos que tanto e_1 como e_2 no son parámetros independientes (se pueden calcular directamente a partir de h_1 y h_2. Sin embargo, a los efectos de analizar el impacto en la respuesta correlacionada, basta observar que más que el efecto individual de h_1 y h_2, lo que es importante observar es el producto de estas para el aporte de la correlación genética aditiva, mientras que para la correlación fenotípica, a partir de las heredabilidades, tenemos

$$\begin{aligned} e_1 = \sqrt{1 - h_1^2} \text{ y } e_2 = \sqrt{1 - h_2^2} \;\therefore \\ e_1 \times e_2 = \sqrt{1 - h_1^2} \times \sqrt{1 - h_2^2} = \sqrt{(1 - h_1^2) \times (1 - h_2^2)} \\ e_1 \times e_2 = \sqrt{1 - h_1^2 - h_2^2 + h_1^2 \times h_2^2} \end{aligned} \tag{12.41}$$

Ahora, claramente no alcanza con una función que tome el producto de las heredabilidades para caracterizar el aporte de la correlación fenotípica. De hecho, para el mismo resultado del producto $h_1^2 \times h_2^2$, $e_1 \times e_2$ será menor cuanto mayor la diferencia entre h_1^2 y h_2^2. Sin pérdida de generalidad, solo para ilustrar el comportamiento de la correlación fenotípica

respecto a la genética y la ambiental, asumiendo para simplificar que $h_1^2 = h_2^2 = h^2$, tenemos

$$\sqrt{(1 - h_1^2) \times (1 - h_2^2)} = \sqrt{1 - 2 \times h^2 + (h^2)^2} \tag{12.42}$$

Es decir, bajo el supuesto anterior, a partir del producto de las heredabilidades es posible analizar la correlación fenotípica como función de las correlaciones genética y ambiental. Los resultados de este análisis, para distintos valores del producto de heredabilidades se pueden observar en la figura 12.4. Si llamamos

$$a = \sqrt{(h^2)^2} = h^2 \text{ y } b = \sqrt{1 - 2 \times h^2 + (h^2)^2} \tag{12.43}$$

entonces

$$r_{P_{12}} = a \times r_{A_{12}} + b \times r_{E_{12}} \tag{12.44}$$

O sea, dado el producto de las heredabilidades (bajo el supuesto de que provengan de iguales heredabilidades), la correlación fenotípica es una combinación lineal de las correlaciones genética y ambiental. Más aún, se trata de una combinación lineal aunque no se cumpla la igualdad de heredabilidades, solo que ahora

$$b = \sqrt{1 - h_1^2 - h_2^2 + h_1^2 \times h_2^2} \tag{12.45}$$

y por lo tanto será en general menor que cuando son idénticas las heredabilidades. Para demostrar esto último, basta que se cumpla

$$\sqrt{1 - h_1^2 - h_2^2 + h_1^2 \times h_2^2} \leq \sqrt{1 - 2 \times h^2 + (h^2)^2} \tag{12.46}$$

para lo que alcanza con que

$$h_1^2 + h_2^2 \geq 2 \times h^2 \text{ ya que } (h^2)^2 = h_1^2 \times h_2^2 \ \therefore \tag{12.47}$$

$$h_1^2 + h_2^2 \geq 2 \times \sqrt{h_1^2 \times h_2^2} \iff (h_1^2 + h_2^2)^2 \geq 4 \times h_1^2 \times h_2^2 \tag{12.48}$$

$$(h_1^2)^2 + 2 \times h_1^2 \times h_2^2 + (h_2^2)^2 \geq 4 \times h_1^2 \times h_2^2 \iff (h_1^2)^2 - 2 \times h_1^2 \times h_2^2 + (h_2^2)^2 \geq 0 \tag{12.49}$$

$$\iff (h_1^2 - h_2^2)^2 \geq 0 \tag{12.50}$$

El único caso donde el lado izquierdo de la última expresión se anula es cuando $h_1^2 = h_2^2$, mientras que en el resto de los casos $h_1^2 + h_2^2 > 2 \times h^2$. Por lo tanto, en el caso de que las heredabilidades no sean idénticas el aporte de la correlación ambiental será menor respecto a la genética (a un mismo valor del producto de las heredabilidades).

El comportamiento de la relación entre las correlaciones genética aditiva y ambiental con la correlación fenotípica se puede ver en la figura 12.4, para distintos valores del producto de las heredabilidades (considerándolas iguales). En la misma, se aprecia claramente cómo al aumentar el producto de las heredabilidades (la varianza aditiva explica una mayor parte

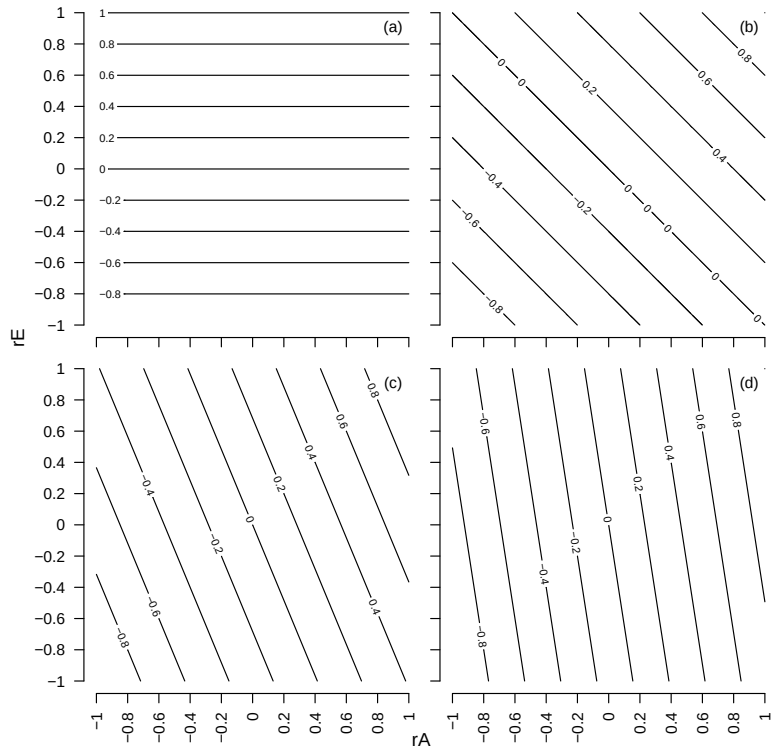


Figura 12.4. Correlaciones fenotípicas para distintas combinaciones de correlaciones genéticas (aditivas) y ambientales (a) $h_1^2 \times h_2^2 = 0$, (b) $h_1^2 \times h_2^2 = 0{,}25$, (c) $h_1^2 \times h_2^2 = 0{,}50$, (d) $h_1^2 \times h_2^2 = 0{,}75$

de la fenotípica), las curvas de nivel de correlación fenotípica pasan de la horizontalidad ($r_{P_{12}} = r_{E_{12}}$) a gradualmente irse hacia la vertical (donde $r_{P_{12}} = r_{A_{12}}$). Además, en la figura se observa claramente la relación lineal entre las tres correlaciones ya que cualquier curva de nivel (correlación fenotípica) se puede construir como combinación lineal de las otras dos correlaciones. Una discusión interesante de este tema se puede encontrar en Searle (1961).

Ejemplo 12.3

Dos características fenotípicas se encuentran correlacionadas a través de una correlación genética r_A y una correlación ambiental r_E. Si las heredabilidades de las dos características son $h_1^2 = 0{,}25$ y $h_2^2 = 0{,}75$, ¿cuál es la relación entre el aporte de la correlación genética y la ambiental a la correlación fenotípica? ¿Cuáles son las ponderaciones de la correlación genética y de la correlación ambiental?

Si las heredabilidades fuesen $h_1^2 = 0{,}40$ y $h_2^2 = 0{,}60$, ¿cuál sería la relación? ¿Cuáles son ahora las ponderaciones de la correlación genética y de la correlación ambiental?

De acuerdo con la ecuación (12.40), el aporte de la correlación genética estará ponderado por $h_1 h_2 = \sqrt{h_1^2 h_2^2}$, mientras que el aporte de la correlación ambiental estará ponderado por $\sqrt{1-h_1^2}\sqrt{1-h_2^2}$. La relación se aportes será entonces

$$\text{rel}_{\text{A/E}} = \frac{\sqrt{h_1^2 h_2^2}}{\sqrt{1-h_1^2}\sqrt{1-h_2^2}} = \sqrt{\frac{h_1^2 h_2^2}{(1-h_1^2)(1-h_2^2)}} \tag{12.51}$$

En el caso de que $h_1^2 = 0{,}25$ y $h_2^2 = 0{,}75$, entonces lo anterior se transforma en

$$\text{rel}_{\text{A/E}} = \sqrt{\frac{0{,}25 \times 0{,}75}{(1-0{,}25)(1-0{,}75)}} = \sqrt{\frac{0{,}25 \times 0{,}75}{0{,}75 \times 0{,}25}} = \sqrt{1} = 1 \tag{12.52}$$

En este caso el peso de la correlación genético-aditiva y de la correlación ambiental será igual en su contribución a la correlación fenotípica. La ponderación para la correlación genética será $\sqrt{h_1^2 h_2^2} = \sqrt{0{,}25 \times 0{,}75} = 0{,}433$, mientras que la ponderación para la correlación ambiental será de $\sqrt{1-h_1^2}\sqrt{1-h_2^2} = \sqrt{0{,}75}\sqrt{0{,}25} = \sqrt{0{,}25 \times 0{,}75} = 0{,}433$ (obviamente, ya que $\text{rel}_{\text{A/E}} = 1$).

En el segundo caso, sustituyendo $h_1^2 = 0{,}40$ y $h_2^2 = 0{,}60$, tenemos

$$\text{rel}_{\text{A/E}} = \sqrt{\frac{0{,}40 \times 0{,}60}{(1-0{,}40)(1-0{,}60)}} = \sqrt{\frac{0{,}40 \times 0{,}60}{0{,}60 \times 0{,}40}} = \sqrt{1} = 1 \tag{12.53}$$

mientras que las ponderaciones de la contribución genética y ambiental serán $\sqrt{h_1^2 h_2^2} = \sqrt{0{,}40 \times 0{,}60} = 0{,}49$ (ambas son iguales porque $\text{rel}_{\text{A/E}} = 1$).

En general, para que las ponderaciones de lo genético y lo ambiental sean iguales alcanza con que

$$\text{rel}_{\text{A/E}} = \sqrt{\frac{h_1^2 h_2^2}{(1-h_1^2)(1-h_2^2)}} = 1 \Leftrightarrow h_1^2 h_2^2 = (1-h_1^2)(1-h_2^2) \Leftrightarrow \tag{12.54}$$

$$\Leftrightarrow h_1^2 h_2^2 = 1 - h_1^2 - h_2^2 + h_1^2 h_2^2 \Leftrightarrow 1 - h_1^2 - h_2^2 = 0 \Leftrightarrow h_1^2 + h_2^2 = 1 \Leftrightarrow \tag{12.55}$$

$$\Leftrightarrow h_1^2 = 1 - h_2^2 \tag{12.56}$$

es decir, siempre que la suma de heredabilidades sea igual a 1.

PARA RECORDAR

- Retomando la relación entre las correlaciones fenotípica, genético-aditiva y ambiental a través de la siguiente expresión: $r_{P_{12}} = r_{A_{12}}\, h_1\, h_2 + r_{E_{12}}\, e_1\, e_2$, podemos notar que la **correlación fenotípica no solo depende de las correlaciones genética (aditiva) y ambiental (extendida)**, sino que **también depende de las exactitudes de las dos características (o de sus heredabilidades)**. Es decir, **cuanto mayores sean las heredabilidades, mayor será la influencia relativa de la correlación genética y viceversa para la ambiental.**
- A su vez, a un mismo valor del producto $h_1^2 \times h_2^2$, $e_1 \times e_2$ será menor cuanto mayor la diferencia entre h_1^2 y h_2^2.

12.5 Respuesta correlacionada

Hasta ahora hemos venido trabajando las correlaciones entre características desde el punto de vista de la observación de sus efectos y de las relaciones entre ellas (fenotípica, genética y ambiental). Sin embargo, desde el punto de vista del mejoramiento genético es fundamental entender lo que sucede con las características que están correlacionadas desde el punto de vista genético con las características por las que realizamos la selección. Es decir, si dos características se encuentran correlacionadas genéticamente, al realizar selección en una de ellas, también estaremos desplazando la media de la otra característica, a pesar de no haber realizado ninguna selección en forma explícita en ella.

Para analizar cómo la selección en una característica afecta a las características correlacionadas genéticamente vamos a suponer que tenemos dos características, una que llamaremos X y sobre la que realizaremos la selección y una segunda característica Y, que se encuentra correlacionada con la primera (notemos el cambio de notación que hicimos, ya que en lugar de llamarlas 1 y 2 vamos a usar las tradicionales X e Y).

Como vimos antes, la respuesta en la característica bajo selección, X en nuestro caso, es equivalente al valor reproductivo medio de los individuos seleccionados (porque es lo que se transmite completamente). Al mismo tiempo, el cambio que ocurrirá en Y, es decir la respuesta, lo podemos plantear entonces como la regresión de los valores reproductivos de Y en X. El coeficiente de esta regresión sería por lo tanto

$$b_{A_{YX}} = \frac{Cov_A(X,Y)}{\sigma^2_{A_X}} \tag{12.57}$$

Pero de la ecuación (12.30) sabemos que la relación entre la covarianza genética (aditiva) y su correspondiente coeficiente de correlación nos permite escribirla como

$$r_{A_{XY}} = \frac{Cov_A(X,Y)}{\sigma_{A_X}\sigma_{A_Y}} \therefore Cov_A(X,Y) = r_{A_{XY}}\sigma_{A_X}\sigma_{A_Y} \tag{12.58}$$

por lo que, sustituyendo este último resultado en la ecuación (12.57), tenemos que el coeficiente de regresión es igual a

$$b_{A_{YX}} = \frac{r_{A_{XY}} \sigma_{A_X} \sigma_{A_Y}}{\sigma^2_{A_X}} = r_{A_{XY}} \frac{\sigma_{A_Y}}{\sigma_{A_X}} \tag{12.59}$$

Como vimos en el capítulo «Selección artificial I», la respuesta a la selección (en una generación realizando selección individual) para la característica X la podemos escribir como

$$R_X = i_X \; h_X \; \sigma_{A_X} \tag{12.60}$$

Por lo tanto, la respuesta correlacionada en la característica Y es

$$\begin{aligned} RC_Y = b_{A_{YX}} \; R_X = b_{A_{YX}} \; i_X \; h_x \; \sigma_{A_X} = \\ = r_{A_{XY}} \frac{\sigma_{A_Y}}{\sigma_{A_X}} \; i_X \; h_X \; \sigma_{A_X} \therefore \\ RC_Y = i_X \; h_X \; r_{A_{XY}} \; \sigma_{A_Y} \end{aligned} \tag{12.61}$$

El resultado de la ecuación anterior es intuitivo y fácil de interpretar. El cambio que vamos a esperar en la característica Y cuando se realiza selección en la característica X es igual a la intensidad de selección (en X) multiplicada por el desvío estándar aditivo de la característica Y (que es la que estamos observando y que, además, debe tener las unidades de la respuesta) y por un «coeficiente de paso» formado por $h_X \; r_{A_{XY}}$. Si observamos nuevamente la figura 12.3, vemos que A_2 (que actúa con efecto h_2 sobre P_2) podría ser sustituida por $h_1 \; r_{A_{12}}$ (la correlación entre A_2 y P_1). Cambiando 1 por X y 2 por Y, se entiende perfectamente el rol que juega el producto $h_X \; r_{A_{XY}}$ en lugar de h_Y.

Finalmente, si tenemos en cuenta que $h_Y = \frac{\sigma_{A_Y}}{\sigma_{P_Y}}$ y que por lo tanto $\sigma_{A_Y} = h_Y \; \sigma_{P_Y}$, una forma alternativa de escribir la ecuación (12.61) es

$$RC_Y = i_X \; h_X \; r_{A_{XY}} \; \sigma_{A_Y} = i_X \; h_X \; r_{A_{XY}} \; h_Y \; \sigma_{P_Y} \tag{12.62}$$

que en lugar de depender del desvío aditivo de la característica, depende ahora del desvío fenotípico de la misma.

De hecho, estas últimas ecuaciones sugieren un método para estimar la correlación genética entre las dos características. La idea es sencilla: separamos la población en dos grupos, uno sobre el que vamos a realizar selección directa en la característica X y el otro sobre el que vamos a realizar selección directa sobre la característica Y. En ambos grupos realizamos la misma presión de selección, por lo que $i_X = i_Y = i$ y en ambos vamos a medir la respuesta a la selección en ambas características, la seleccionada directamente y la correlacionada. El cociente entre selección directa e indirecta para la característica X es igual a

$$\frac{RC_X}{R_X} = \frac{i_Y \; h_Y \; r_{A_{XY}} \; \sigma_{A_X}}{i_X \; h_X \; \sigma_{A_X}} = \frac{h_Y}{h_X} \; r_{A_{XY}} \tag{12.63}$$

mientras que para la característica Y es igual a

$$\frac{RC_Y}{R_Y} = \frac{i_X \; h_X \; r_{A_{XY}} \; \sigma_{A_Y}}{i_Y \; h_Y \; \sigma_{A_Y}} = \frac{h_X}{h_Y} \; r_{A_{XY}} \tag{12.64}$$

Ahora, si multiplicamos los resultados de las ecuaciones (12.63) y (12.64), tenemos que

$$\frac{RC_X}{R_X}\frac{RC_Y}{R_Y} = \frac{h_Y}{h_X}\, r_{A_{XY}} \frac{h_X}{h_Y}\, r_{A_{XY}} = \frac{h_Y}{h_X}\frac{h_X}{h_Y}\,(r_{A_{XY}})^2 \therefore$$
$$r^2_{A_{XY}} = \frac{RC_X}{R_X}\frac{RC_Y}{R_Y} \;\therefore\; r_{A_{XY}} = \sqrt{\frac{RC_X}{R_X}\frac{RC_Y}{R_Y}} \tag{12.65}$$

que es un buen estimador del valor absoluto de la correlación genética (aditiva) entre las dos características (notemos que al estar la correlación $r_{A_{XY}}$ elevada al cuadrado, su signo siempre será positivo), siempre y cuando las respuestas recíprocas coincidan, de acuerdo con lo esperado por la teoría.

Ejemplo 12.4

La oveja galesa de montaña **balwen** (**Balwen Welsh Mountain**), figura 12.5, es una raza que se cría especialmente para carne. A fin de definir estrategias para mejorar su producción de lana, para estimar la correlación genética entre peso de vellón sucio (**PVS**) y peso de vellón limpio (**PVL**) se realizó un experimento de selección directa y selección indirecta para ambas características, donde se obtuvieron las siguientes respuestas en una generación:

Característica Seleccionada	Característica de Interés	Tipo	Valor
PVL	**PVL**	Directa	0,113 kg
PVS	**PVS**	Directa	0,101 kg
PVS	**PVL**	Indirecta	0,084 kg
PVL	**PVS**	Indirecta	0,082 kg

Estimar la correlación genética (aditiva) entre ambas características.

Utilizando la ecuación (15.3), tenemos que un estimador de la correlación genética (aditiva) a partir de experimentos de este tipo está dada por

$$r^2_{A_{XY}} = \frac{RC_X}{R_X}\frac{RC_Y}{R_Y} \tag{12.66}$$

Sustituyendo los valores que tenemos en la tabla, si llamamos X a **PVL** e Y a **PVS**, tenemos entonces que

$$r^2_{A_{XY}} = \frac{0{,}084\text{ kg}}{0{,}113\text{ kg}}\frac{0{,}082\text{ kg}}{0{,}101\text{ kg}} = 0{,}6035$$
$$r_{A_{XY}} = \sqrt{r^2_{A_{XY}}} = \sqrt{0{,}6035} = 0{,}7769 \tag{12.67}$$

que es una correlación genética positiva y mediana.

Figura 12.5. La oveja galesa de montaña **balwen** (**Balwen Welsh Mountain**) es originaria del valle de Tywi, en Gales y es una de las variedades criadas esencialmente para la producción de carne. Tiene un patrón de color característico, con el cuerpo negro y las extremidades blancas. (Foto: Mick Lobb, CC BY-SA 2.0, https://commons.wikimedia.org/w/index.php?curid=14555593).

PARA RECORDAR

- Para analizar cómo la **selección en una característica afecta a las características correlacionadas genéticamente** vamos a suponer que tenemos dos características, una que llamaremos X y sobre la que realizaremos la selección y una segunda característica Y, que se encuentra correlacionada con la primera.
- Podemos plantear el cambio que ocurrirá en Y, es decir la respuesta, como la regresión de los valores reproductivos de Y en X. El coeficiente de esta regresión sería $b_{A_{YX}} = \frac{Cov_A(X,Y)}{\sigma^2_{A_X}} = r_{A_{XY}} \frac{\sigma_{A_Y}}{\sigma_{A_X}}$
- La **respuesta correlacionada** en la característica Y es $RC_Y = i_X\, h_X\, r_{A_{XY}}\, \sigma_{A_Y}$. Según esto, el cambio que vamos a esperar en la característica Y cuando se realiza selección en la característica X es igual a la intensidad de selección (en X) multiplicada por el desvío estándar aditivo de la característica Y (que es la que estamos observando y que, además, debe tener las unidades de la respuesta) y por un «coeficiente de paso» formado por $h_X\, r_{A_{XY}}$. Alternativamente podemos escribir esta expresión como $RC_Y = i_X\, h_X\, r_{A_{XY}}\, h_Y\, \sigma_{P_Y}$ que en lugar de depender del desvío aditivo de la característica, depende ahora de su desvío fenotípico.

Eficiencia en relación con la selección directa

Los resultados obtenidos respecto al cambio de la media (respuesta, progreso genético generacional) en una característica cuando se selecciona por otra sugiere la posibilidad de hacer **selección indirecta**, es decir usar criterios de una característica (su fenotipo, intensidad de selección, etcétera) para mejorar otra, la de nuestro interés.

Supongamos que queremos mejorar la característica Y y estamos analizando la posibilidad de hacerlo directamente, o como alternativa, utilizar la información de la característica X y realizar selección indirecta. Como vimos en la ecuación (12.64), el cociente entre los dos tipos de selección, **indirecta** en el numerador, **directa** en el denominador, es igual a

$$\frac{RC_Y}{R_Y} = \frac{i_X\, h_X\, r_{A_{XY}}\, \sigma_{A_Y}}{i_Y\, h_Y\, \sigma_{A_Y}} = \frac{i_X}{i_Y} \frac{h_X}{h_Y}\, r_{A_{XY}} \qquad (12.68)$$

En general, cuando por razones de manejo de la población (rodeo, majada, piara, etcétera) debemos retener una determinada proporción de animales como reproductores, la misma proporción será retenida más allá del criterio o característica por la que realicemos la selección. Más aún, si las dos características X e Y poseen distribución normal, como vimos previamente en la sección «Intensidad de selección y proporción seleccionada», la intensidad de selección es una función de la proporción p seleccionada ($i = \frac{z}{p}$) y por lo tanto, $i_X = i_Y$. En ese caso, la ecuación (12.68) se transforma en

$$\frac{RC_Y}{R_Y} = \frac{h_X}{h_Y}\, r_{A_{XY}} \qquad (12.69)$$

lo que es una expresión muy sencilla de evaluar. Para que nos convenga el método de **selección indirecta**, el cociente $\frac{RC_Y}{R_Y}$ debe ser mayor a 1. Por lo tanto, aplicando este criterio en la ecuación (12.69) tenemos que la condición puede expresarse como

$$\frac{h_X}{h_Y}\, r_{A_{XY}} > 1 \Leftrightarrow r_{A_{XY}} > \frac{h_Y}{h_X} \tag{12.70}$$

Dicho de otra forma, para que nos convenga utilizar **selección indirecta** (usando información de X para seleccionar en Y), la correlación aditiva entre las características debe ser mayor que la relación entre las precisiones $\frac{h_Y}{h_X}$. Cuanto mayor sea la precisión para selección fenotípica individual en la característica indirecta (h_X), menor será el cociente $\frac{h_Y}{h_X}$ y mayor la probabilidad de que $r_{A_{XY}}$ lo supere. Entendido todo esto, igualmente es necesario comprender que los principales motivos para realizar selección indirecta no se basan en estos conceptos, sino que la facilidad y el bajo costo de medición de la característica correlacionada respecto a la de interés son lo que usualmente motiva la aplicación de este tipo de selección.

La ecuación (12.70) también puede arreglarse de forma tal que la condición quede expresada como $h_X\, r_{A_{XY}} > h_Y$, es decir, que el producto de la precisión en la característica indirecta por la correlación genética sea mayor que la precisión directa. Es decir, en términos cualitativos, se debe cumplir que la precisión de la indirecta sea alta respecto a la directa, así como la correlación genética también sea relativamente alta. En general esto se aplica a condiciones donde la heredabilidad de la característica de interés es muy baja y que al mismo tiempo sea posible encontrar una característica fuertemente correlacionada a nivel genético con la de interés y que posea una heredabilidad razonable. Las características difíciles de medir poseen en general una baja heredabilidad, producto de que el error de medición constituye una parte importante de la varianza fenotípica, por lo que la heredabilidad (la fracción de varianza aditiva) resulta reducida.

Un problema en el que la selección indirecta (a través de la respuesta correlacionada) es de gran utilidad, es el de aquellas características que se expresan solamente en un sexo, pero que tienen alguna característica fuertemente correlacionada desde el punto de vista genético. Eso permitiría utilizar la selección por la característica directa en el sexo donde la característica de interés se expresa (a menos que la indirecta sea también más eficiente) y utilizar la selección indirecta en el otro sexo, mejorando de esta forma la intensidad de selección.

Finalmente, otro uso de las correlaciones genéticas tiene que ver con la interacción entre genotipo y ambiente. Como vimos antes, el valor de cría de los individuos es una función de la heredabilidad de la característica y se refiere a una determinada población en determinadas condiciones. Más aún, la existencia de interacción entre genotipo y ambiente hace que las diferencias esperadas entre individuos sea una función del ambiente. Por ejemplo, no podemos pretender que las diferencias en producción de leche de vacas holando de alta producción sea la misma en sistemas estabulados en climas templados que en la producción pastoril en los trópicos. Por lo tanto, si bien la producción de leche parece ser la misma característica, es dudoso que las mismas combinaciones de expresión génica estén actuando en los dos ambientes, por lo que no suena insensato considerarlas (en términos operativos) como dos características diferentes. En este sentido, las producciones en los dos ambientes serían consideradas como dos características diferentes y que, por lo tanto, se les pueden calcular las correlaciones fenotípica, aditiva y ambiental. A partir de esto, será

posible predecir el comportamiento esperado de un animal con información en un ambiente en el otro ambiente.

Ejemplo 12.5

En una población de ganado de la raza **braunvieh** el peso al destete (**PD**) tiene una heredabilidad de $h^2_{PD} = 0{,}36$, mientras el peso a los 18 meses (**P18**) tiene una heredabilidad de $h^2_P 18 = 0{,}25$. Suponiendo que los animales se aparean por primera vez a los 2 años y que la característica que nos interesa mejorar es **P18**, teniendo en cuenta que la correlación genética (aditiva) entre **PD** y **P18** es de $r_A = 0{,}5$, ¿seleccionaríamos en forma **directa** por **peso a los 18 meses** o **indirectamente** por **peso al destete** para mejorar P18?

Si la intensidad de selección es la misma en las dos características (como sería el caso), de acuerdo con la ecuación (12.69), utilizando RC para la respuesta correlacionada en Y al seleccionar por X y RD para la respuesta a la selección directa, tenemos que

$$\frac{RC}{RD} = \frac{h_X}{h_Y}\, r_{A_{XY}} \tag{12.71}$$

En nuestro caso Y es el peso a los 18 meses, mientras que X es **PD**, por lo que poniéndole números

$$RC/RD = \frac{0{,}5 \times \sqrt{0{,}36}}{\sqrt{0{,}25}} = \frac{0{,}5 \times 0{,}6}{0{,}5} = 0{,}6 \tag{12.72}$$

Por lo tanto, es preferible **seleccionar directamente** para mejorar **P18**. La razón de esto es que las heredabilidades no son muy distintas (por lo que la raíz de su relación no es mucho mayor a uno), sumado a que la correlación genético-aditiva es intermedia.

PARA RECORDAR

- Cuando por razones de manejo de la población debemos retener una determinada proporción de animales como reproductores, la misma proporción será retenida más allá del criterio o característica por la que realicemos la selección. Más aún, si las dos características X e Y poseen distribución normal, la intensidad de selección será $i_X = i_Y$. En ese caso, podemos determinar que $\frac{RC_Y}{R_Y} = \frac{h_X}{h_Y}\, r_{A_{XY}}$
- Dicho esto, para evaluar qué método de selección nos conviene aplicar, basta una simple observación de los resultados: si el cociente $\frac{RC_Y}{R_Y}$ es mayor a 1, nos será más conveniente realizar selección indirecta. Dicho de otra forma, la correlación aditiva entre las características debe ser mayor que la relación entre las precisiones $\frac{h_X}{h_Y}$
- Como norma general, las características difíciles de medir poseen en general una baja heredabilidad, producto de que el error de medición constituye una parte importante de la varianza fenotípica, por lo que la heredabilidad (la fracción de varianza aditiva) resulta reducida.
- Más allá de estas relaciones matemáticas, la selección indirecta es fundamental en los casos en los que las características se expresan solo en un sexo, son costosas o muy difíciles de medir.

El signo de la correlación y el beneficio-demérito de esta

Un detalle a considerar en nuestras disquisiciones anteriores sobre el mérito relativo de la selección indirecta es el del signo de la correlación genética. Si la correlación es fuertemente negativa, pongamos, por ejemplo $r_{A_{XY}} = -1$, sustituyendo ese valor en la ecuación (12.70) tendríamos que $-1 > \frac{h_Y}{h_X} \Leftrightarrow \frac{h_Y}{h_X} < -1$, lo que claramente es imposible para cualquier valor de h_Y y h_X ya que los mismos deben ser positivos (son las raíces cuadradas de heredabilidades).

Dicho de otra forma, pese a que existe una asociación perfecta entre variables, como el signo de la misma es negativo no podríamos usarla para realizar **selección indirecta**. Claramente, debe existir una inconsistencia en nuestro razonamiento que lleva a este absurdo. La inconsistencia proviene del supuesto que $i_X = i_y = i$ que nos permitía simplificar la relación de la ecuación (12.68) a la de la ecuación (12.69), de donde derivamos la ecuación (12.70). De hecho, en términos generales $i = \frac{S}{\sigma_P}$. Como el diferencial de selección tiene signo y el desvío estándar fenotípico es siempre positivo (raíz cuadrada de una varianza), entonces la intensidad de selección tendrá signo también. Asumiendo que seleccionamos la misma proporción de individuos para cualquiera de los dos métodos (directo e indirecto), como los sentidos son opuestos (signo negativo de la correlación aditiva), entonces su cociente será igual a -1 y por lo tanto la ecuación equivalente a la ecuación (12.70) $-r_{A_{XY}} > \frac{h_Y}{h_X}$

Por otra parte, es importante entender que el signo de una correlación, por ejemplo, la genética, no habla de por sí de si la misma es favorable o desfavorable. Pongamos un ejemplo sencillo de entender. En la producción lanera, en las condiciones actuales, el ingreso por venta de lana que tienen los productores es función directa de la cantidad (kg) de lana producida e inversa al diámetro de las fibras. Esto se debe, entre otras cosas, a que la lana fina produce una mejor sensación al tacto y por lo tanto se usa preferentemente para vestimenta de alta calidad. En ausencia de otros grandes defectos en la lana (color, fibras pigmentadas, largo de mecha), los objetivos de mejora del productor lanero deberían ser aumentar el peso de vellón (limpio, por ejemplo) y disminuir el diámetro de las fibras. Desafortunadamente para el productor lanero, la naturaleza tenía otra cosa en mente al desarrollar el «pelaje» de los mamíferos. En una oveja de un tamaño determinado, por ejemplo, hay tres formas en las que la naturaleza podría incrementar el peso del vellón limpio anual, a saber: *a)* aumentar el número de folículos por unidad de superficie, *b)* aumentar la velocidad de crecimiento (largo) de las fibras, *c)* aumentar su diámetro (es decir, el área de la sección transversal). Sin descartar la importancia de las dos primeras, la tercera sin duda juega un papel importante: fibras de mayor diámetro tienen mayor superficie en la sección transversal (aun siendo relativamente huecas), que se transforman en mayor volumen (al mismo largo) y si no disminuye la densidad del material, mayor peso. Esto explica la **correlación positiva** entre peso de vellón y diámetro, que desafortunadamente va en contra de los criterios de pago del mercado actual. Por lo tanto, **esta correlación genética es desfavorable**, positiva y desfavorable. Si por alguna extraña razón, por ejemplo, si no existiese más interés en elaborar prendas finas de lana y el la gente deseara sentir un fuerte escozor al usar prendas de lana, posiblemente el mercado pagaría más por las lanas gruesas, por lo que ahora **la misma correlación genética positiva pasaría a ser favorable**.

En general, es difícil que en corto tiempo cambie el signo de una correlación genética, aunque como veremos en la sección La forma generalizada de la ecuación del criador, las correlaciones genéticas también evolucionan en el tiempo (usualmente, el tiempo en términos evolutivos). Sin embargo, las tendencias del mercado, que es de última lo que refleja

las preferencias de los clientes, usualmente cambian en el tiempo en forma mucho más acelerada. Por ejemplo, la importancia en la ecuación de ingresos del productor lechero de la grasa ha cambiado sustancialmente en los últimos cincuenta años y en los diferentes mercados: desde ignorar completamente su valor, asignarle un valor importante, reducir su valor relativo (favoreciendo las proteínas), hasta volver a incrementar su valor relativo. Más aún, actualmente existen mercados que pagan muy bien por ella, otros que pagan poco y algunos nada. Sin embargo, su correlación genética con la producción de leche, con la producción de proteínas, etcétera sigue siendo aproximadamente la misma en cada población de vacas holando. Por ejemplo, mientras que la correlación genética (aditiva) entre producción de leche (litros o kilogramos) y la producción de grasa o proteína es positiva, mientras que en los **mercados de leche fluida** (aquellos en los que el destino principal es el consumo de leche) se trata de una relación **neutra o aún favorable**, en los mercados de leche destinada a la producción mayoritaria de sólidos (manteca, queso, leche en polvo), la misma es totalmente **desfavorable**.

PARA RECORDAR

- Es muy importante **no confundir** correlaciones positivas o negativas con favorables y desfavorables. Es decir, el **signo de la correlación no representa si esta nos es favorable** productivamente.
- Como ejemplo podemos pensar en el micraje de la lana y el peso del vellón, esta correlación es **positiva**, debido a que en general un mayor peso de vellón implica un mayor grosor de la fibra, sin embargo nos es **desfavorable** debido a que lo deseado es aumentar el peso de vellón pero manteniendo o disminuyendo el grosor de la fibra.

12.6 Matrices de varianza-covarianza

Como vimos antes, para cada par de características de interés es posible, en principio, estimar las covarianzas correspondientes a nuestro modelo simplificado. En particular, las covarianzas genético-aditivas, las covarianzas ambientales y las covarianzas fenotípicas. Si tenemos dos características, existirá una sola covarianza de cada tipo (genético-aditiva, ambiental, fenotípica), pero cuando consideramos más características debemos considerar todos los pares de variables posibles para cada tipo de covarianza. Por ejemplo, si tenemos tres características de interés, las covarianzas genético-aditivas serán $Cov(A_1,A_2)$, $Cov(A_1,A_3)$ y $Cov(A_2,A_3)$, o usando una notación más sintética $\sigma_{A_{1,2}}$, $\sigma_{A_{1,3}}$ y $\sigma_{A_{2,3}}$. Claramente, cuando tenemos k características en consideración, habrá $k(k-1)/2$ pares posibles de características (asumiendo que el orden no es relevante, es decir $\sigma_{A_{i,j}} = \sigma_{A_{j,i}}$).

Por otro lado, la covarianza de una variable consigo misma es igual a la varianza de dicha variable, ya que

$$Cov(x,x) = \frac{1}{n-1}\sum(x-\bar{x})(x-\bar{x}) = \frac{1}{n-1}\sum(x-\bar{x})^2 = Var(x) \qquad (12.73)$$

Puesto en la notación anterior, $\sigma_{A_{i,i}} = \sigma^2_{A_i}$. De estas varianzas, habrá k para k características, es decir, una por característica. Esto nos lleva a que el número de **parámetros** para cada uno de los tipos de covarianza (genético-aditiva, ambiental, fenotípica) será igual a

$$\frac{k(k-1)}{2} + k = \frac{k(k-1)+2k}{2} = \frac{k[(k-1)+2]}{2} = \frac{k(k+1)}{2} \qquad (12.74)$$

Una forma sencilla de acomodar todos los **parámetros** es ubicarlos en una matriz, como se ilustra en la figura 12.6 para las varianzas-covarianzas genético-aditivas (matriz que llamaremos **G**) de tres características. El arreglo de los elementos en la misma es muy sencillo: la matriz es cuadrada (es decir, tiene el mismo número de filas que de columnas) de orden $k = 3$ en este caso. Los índices de filas y columnas se corresponden con los números de las características; por ejemplo, la fila $i = 2$ y la columna $j = 3$ se corresponde al parámetro de interés (covarianza, genético-aditiva en este caso) entre las características 2 y 3. Como la diagonal está formada por la restricción en los índices tales que $i = j$, en la misma aparecen las covarianzas entre la característica i consigo misma, que como ya vimos se trata de las varianzas (genético-aditivas) de las características. Fuera de la diagonal, los índices son distintos necesariamente, es decir $i \neq j$ y por lo tanto de trata de covarianzas (genético-aditivas) entre dos características distintas. Además, como $\sigma_{A_{i,j}} = \sigma_{A_{j,i}}$, **la matriz es simétrica**, es decir que a cada elemento de la triangular superior le corresponde un elemento en la triangular inferior, producto de intercambiar los índices i y j. Las matrices de varianza-covarianza son positivas-semidefinidas, es decir que para cualquier vector **z** perteneciente a los reales, el escalar $\mathbf{z'Gz}$ debe ser mayor o igual a cero, con **G** la matriz de varianza-covarianza (en este caso genético-aditiva).

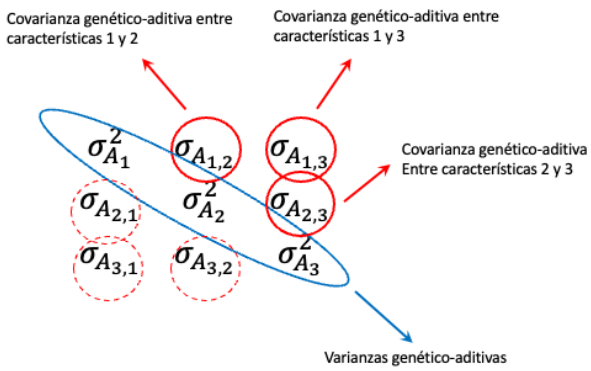


Figura 12.6. La matriz de varianza-covarianza genético-aditiva para 3 características. La diagonal (encerrada en un óvalo azul) representa las covarianzas de cada variable consigo misma, es decir las varianzas genético-aditivas de cada característica. Fuera de la diagonal aparecen las covarianzas genético-aditivas entre las características definidas por la fila y columna correspondiente (encerradas en círculos rojos). Como $\sigma_{A_{i,j}} = \sigma_{A_{j,i}}$, **la matriz es simétrica** y a cada círculo rojo en la triangular superior le corresponde un círculo rojo a trazos en la triangular inferior, resultado de intercambiar los índices de filas y columnas.

De la misma forma que definimos la estructura de la matriz **G** para las varianzas-covarianzas genético-aditivas entre característica, podemos hacer lo mismo para el resto de las covarianzas de nuestro interés, por ejemplo, fenotípicas o ambientales. En la figura 12.7 se representa la estructura de la matriz de varianza-covarianza fenotípica, que llamaremos **P**, para tres categorías de interés. Se aprecia claramente que la estructura es completamente análoga a la que definimos previamente para **G**.

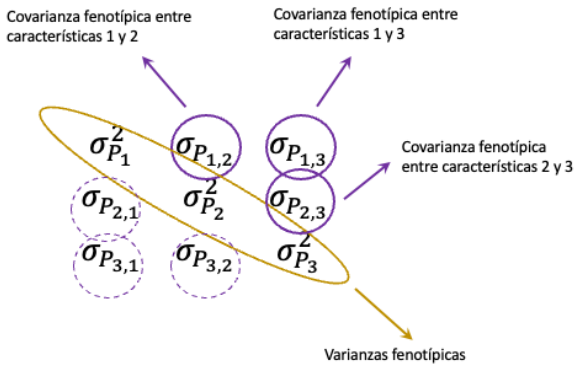


Figura 12.7. La matriz de varianza-covarianza fenotípica para 3 características. La diagonal representa las covarianzas de cada variable consigo misma, es decir las varianzas fenotípicas de cada característica. Fuera de la diagonal aparecen las covarianzas fenotípicas entre las características definidas por la fila y columna correspondiente. Como $\sigma_{P_{i,j}} = \sigma_{P_{j,i}}$, **la matriz es simétrica.**

Como veremos más adelante en este capítulo, organizar todos los parámetros en forma matricial nos permitirá extender la **ecuación del criador** a varias características en forma simultánea manteniendo la notación muy sencilla, y como veremos en el próximo capítulo, será esencial en el desarrollo de los métodos de predicción de valores de cría. Sin embargo, antes de entrar en esos temas, veamos en forma sencilla cómo podemos obtener correlaciones genéticas (aditivas) y fenotípicas, así como las heredabilidades de las características, a partir de las matrices **G** y **P**. Los diferentes programas que se utilizan en mejoramiento genético o en genética cuantitativa para estimar los parámetros genéticos suelen reportarnos ambas matrices y a partir de ellas debemos ser capaces de calcular los parámetros genéticos más relevantes para las decisiones de mejoramiento, usualmente las heredabilidades y las correlaciones genéticas (aditivas) entre características. Supongamos que nos encontramos estudiando k características diferentes y que obtuvimos las correspondientes matrices **G** y **P**, ambas de orden k (k filas y k columnas).

De acuerdo con la definición, la correlación genética entre la característica i y la característica j estará dada por

$$r_{A_{i,j}} = \frac{Cov(A_i,A_j)}{\sqrt{Var(A_i)Var(A_j)}} = \frac{\sigma_{A_{i,j}}}{\sqrt{\sigma^2_{A_i}\sigma^2_{A_j}}} = \frac{\sigma_{A_{i,j}}}{\sigma_{A_i}\sigma_{A_j}} \tag{12.75}$$

Por lo tanto, para calcular la correlación genética (aditiva) entre la característica i y la característica j a partir de la matriz $\mathbf{G}$ el procedimiento es muy sencillo:

1. Ubicar la covarianza genético-aditiva entre las dos características como la entrada $G_{i,j}$ en la matriz $\mathbf{G}$ (o sea, el valor que aparece en la fila i, columna j).
2. Calcular los desvíos estándar aditivos de ambas características, haciendo la raíz cuadrada a los elementos i y j de la diagonal (o sea, a las entradas $G_{i,i}$ y $G_{j,j}$).
3. Dividir la covarianza obtenida en el punto 1 entre el producto de los desvíos estándar obtenidos en el punto 2. Ese será el valor de la correlación genética (aditiva) entre las características i y j.

De la misma manera, las correlaciones fenotípicas quedan definidas por

$$r_{P_{i,j}} = \frac{Cov(P_i,P_j)}{\sqrt{Var(P_i)Var(P_j)}} = \frac{\sigma_{P_{i,j}}}{\sqrt{\sigma^2_{P_i}\sigma^2_{P_j}}} = \frac{\sigma_{P_{i,j}}}{\sigma_{P_i}\sigma_{P_j}} \tag{12.76}$$

por lo que el procedimiento para obtener las correlaciones fenotípicas será igual al descrito antes para las correlaciones genéticas, pero ahora usando la matriz $\mathbf{P}$ en lugar de la matriz $\mathbf{G}$.

Para calcular las heredabilidades de las diferentes características a partir de las matrices $\mathbf{G}$ y $\mathbf{P}$, el procedimiento es aún más sencillo, ya que solo haremos uso de las diagonales de ambas matrices. De acuerdo con la definición de Heredabilidad que vimos en el capítulo «Parámetros genéticos: heredabilidad y repetibilidad», para la característica i la misma sería

$$h^2_i = \frac{Var(A_i)}{Var(P_i)} = \frac{\sigma^2_{A_i}}{\sigma^2_{P_i}} \tag{12.77}$$

Por lo tanto, el procedimiento para calcular nuestra estimación de la heredabilidad de cada característica es trivial: dividir cada elemento de la diagonal de $\mathbf{G}$ entre el correspondiente elemento de la diagonal en la matriz $\mathbf{P}$, es decir

$$h^2_i = \frac{G_{i,i}}{P_{i,i}} \tag{12.78}$$

Como vimos previamente, la relación entre las correlaciones fenotípicas, genéticas (aditivas) y ambientales viene dada por

$$r_{P_{i,j}} = h_i \times h_j \times r_{A_{i,j}} + e_i \times e_j \times r_{E_{i,j}} \tag{12.79}$$

Por lo que, recordando que $e_i = \sqrt{1 - h_i}$, $e_j = \sqrt{1 - h_j}$, podemos obtener también la correlación ambiental (más allá de que a veces tenemos la matriz **E** disponible) a partir de

$$r_{E_{i,j}} = \frac{r_{P_{i,j}} - h_i \times h_j \times r_{A_{i,j}}}{e_i \times e_j} = \frac{r_{P_{i,j}} - h_i \times h_j \times r_{A_{i,j}}}{\sqrt{1 - h_i^2}\sqrt{1 - h_j^2}} = \frac{r_{P_{i,j}} - h_i \times h_j \times r_{A_{i,j}}}{\sqrt{(1 - h_i^2)(1 - h_j^2)}} \tag{12.80}$$

Ejemplo 12.6

A partir de un experimento donde se midieron en todos los animales el largo de la pata trasera izquierda (en cm) y el peso (en kg) a los dos años en **ovejas verdes**, mediante un modelo **BLUP** se pudieron determinar las siguientes matrices de varianza-covarianza

$$\mathbf{G} = \begin{pmatrix} 100 & 180 \\ 180 & 800 \end{pmatrix} \tag{12.81}$$

para la matriz de varianza-covarianza genética y

$$\mathbf{P} = \begin{pmatrix} 250 & 380 \\ 380 & 2400 \end{pmatrix} \tag{12.82}$$

para la matriz de varianza-covarianza fenotípica. Las varianzas se expresan en cm^2 para el largo, en kg^2 para el peso, mientras que las covarianzas son en cm.kg.

A partir de estas matrices, estimar las **heredabilidades** de las dos características, así como sus **correlaciones genéticas** y **fenotípicas**. A partir de los resultados previos, estimar la **correlación ambiental**.

En cada matriz, el elemento $_{1,1}$ corresponde a la **varianza** de la característica 1 (largo de la pata), mientras que el elemento $_{2,2}$ corresponde a la **varianza** de la característica 2 (peso). En general, los elementos de la diagonal son las varianzas de las características (en el caso de **G** las aditivas, mientras que en el caso de **P** las fenotípicas).

Por otra parte, los elementos **fuera de la diagonal** representan **covarianzas**, obviamente entre las características de los subíndices de la matriz. Por ejemplo, $\mathbf{G}_{ij}$ es la covarianza genética (aditiva) entre la característica i y la característica j. Las matrices de varianza-covarianza son simétricas, por lo que $\mathbf{G}_{ij} = \mathbf{G}_{ji}$. En nuestro caso las matrices son de 2×2, por lo que solo existe $\mathbf{G}_{1,2} = \mathbf{G}_{2,1}$ y $\mathbf{P}_{1,2} = \mathbf{P}_{2,1}$.

La definición de heredabilidad como cociente de varianzas ($h^2 = \frac{V_A}{V_P}$) nos lleva a que para la característica 1 (largo de pata trasera izquierda) esta sería

$$h_1^2 = \frac{\mathbf{G}_{1,1}}{\mathbf{P}_{1,1}} = \frac{100 \text{ cm}^2}{250 \text{ cm}^2} = 0{,}400 \tag{12.83}$$

Para la característica 2 (peso), la heredabilidad se estimaría como

$$h_2^2 = \frac{\mathbf{G}_{2,2}}{\mathbf{P}_{2,2}} = \frac{800 \text{ kg}^2}{2400 \text{ kg}^2} = 0{,}333 \tag{12.84}$$

La correlación genética, por definición es igual a $r_A = \frac{\sigma_{A_{1,2}}}{\sqrt{\sigma^2_{A_1}\sigma^2_{A_2}}}$, que utilizando los datos de la matriz $\mathbf{G}$ nos deja con

$$r_A = \frac{180 \text{ cm.kg}}{\sqrt{100 \text{ cm.kg} \times 800 \text{ cm.kg}}} = 0{,}6364 \tag{12.85}$$

Finalmente, la correlación fenotípica es por definición $r_P = \frac{\sigma_{P_{1,2}}}{\sqrt{\sigma^2_{P_1}\sigma^2_{P_2}}}$ y utilizando los valores de la matriz $\mathbf{P}$, sería entonces igual a

$$r_P = \frac{380 \text{ cm.kg}}{\sqrt{250 \text{ cm.kg} \times 2400 \text{ cm.kg}}} = 0{,}4906 \tag{12.86}$$

Utilizando la relación entre las correlaciones fenotípica, genética y ambiental, tenemos que $r_{P_{12}} = h_1 h_2 r_{A_{12}} + e_1 e_2 r_{E_{12}}$. Por definición, tenemos que

$$\begin{aligned} e_1 &= \sqrt{1 - h_1^2} = \sqrt{1 - 0{,}4} = \sqrt{0{,}6} = 0{,}7746 \\ e_2 &= \sqrt{1 - h_2^2} = \sqrt{1 - 0{,}333} = \sqrt{0{,}667} = 0{,}8165 \end{aligned} \tag{12.87}$$

por lo que despejando y sustituyendo estos valores llegamos a que la correlación ambiental en este caso es igual a

$$r_{E_{12}} = \frac{r_{P_{12}} - h_1 h_2 r_{A_{12}}}{e_1 e_2} = \frac{0{,}4906 - (0{,}6364 \times \sqrt{0{,}4 \times 0{,}333})}{0{,}7746 \times 0{,}8165} = \frac{0{,}2583}{0{,}6325} = 0{,}4084 \tag{12.88}$$

PARA RECORDAR

- Como vimos previamente, si tenemos dos características, existirá una sola covarianza de cada tipo (genético-aditiva, ambiental, fenotípica), pero cuando consideramos más características debemos considerar todos los pares de variables posibles para cada tipo de covarianza.
- Si tenemos k características, el número de parámetros para cada uno de los tipos de covarianza será $\frac{k(k+1)}{2}$. Para mayor facilidad, podemos ubicar los mismos en una matriz de varianzas-covarianzas genético- aditivas (**matriz que llamaremos** G).
- La matriz G será cuadrada y simétrica de orden k dónde los índices de filas (i) y columnas (j) se corresponderán con los números de las características. Como la diagonal está formada por la restricción en los índices tales que $i = j$, en la misma aparecen las covarianzas entre la característica i consigo misma, que se trata de las varianzas (genético-aditivas) de las características. Fuera de la diagonal, los índices son distintos necesariamente, es decir $i \neq j$ y por lo tanto de trata de covarianzas (genético-aditivas) entre dos características distintas.

- Llamaremos P a la matriz de varianza-covarianza fenotípica y la misma poseerá una estructura análoga a la descrita para G.
- Si nos encontramos estudiando k características diferentes y que obtuvimos las correspondientes matrices **G** y **P**, ambas de orden k (k filas y k columnas), la correlación genética entre la característica i y la característica j estará dada por $r_{A_{i,j}} = \frac{\sigma_{A_{i,j}}}{\sigma_{A_i}\sigma_{A_j}}$
- Por lo tanto, para **calcular la correlación genética (aditiva)** entre la característica i y la característica j a partir de la matriz **G** el procedimiento: 1) **Ubicar la covarianza genético-aditiva entre las dos características** como la entrada $G_{i,j}$ en la matriz **G** (o sea, el valor que aparece en la fila i, columna j); 2) **Calcular los desvíos estándar aditivos de ambas características**, haciendo la raíz cuadrada a los elementos i y j de la diagonal (o sea, a las entradas $G_{i,i}$ y $G_{j,j}$); 3) **Dividir la covarianza obtenida en el punto 1 entre el producto de los desvíos estándar obtenidos en el punto 2.** Ese será el valor de la correlación genética (aditiva) entre las características i y j.
- El procedimiento para **obtener las correlaciones fenotípicas** será igual al descrito antes para las correlaciones genéticas, pero ahora usando la matriz **P** en lugar de la matriz **G**.
- Para calcular las **heredabilidades** solo haremos uso de las diagonales de ambas matrices, tal que para la característica i la misma sería $h_i^2 = \frac{Var(A_i)}{Var(P_i)} = \frac{\sigma^2_{A_i}}{\sigma^2_{P_i}}$.
  Por lo tanto, el procedimiento para calcular nuestra estimación de la heredabilidad de cada característica será dividir cada elemento de la diagonal de **G** entre el correspondiente elemento de la diagonal en la matriz **P**, es decir $h_i^2 = \frac{G_{i,i}}{P_{i,i}}$
- Recordando la relación entre las correlaciones fenotípicas que vimos previamente, podemos obtener también la correlación ambiental (más allá de que a veces tenemos la matriz **E** disponible) a partir de: $r_{E_{i,j}} = \frac{r_{P_{i,j}} - h_i \times h_j \times r_{A_{i,j}}}{\sqrt{(1-h_i^2)(1-h_j^2)}}$

12.7 La forma generalizada de la ecuación del criador

En el capítulo «Parámetros genéticos: heredabilidad y repetibilidad» estudiamos diferentes definiciones de la heredabilidad (como relación de varianzas o como regresión), pero siempre nuestro contexto fue **univariado**, es decir, estudiamos cada característica de nuestro interés como si no tuviera ningún tipo de relación, dependencia o restricción, con otras características. Más aún, vimos el concepto de heredabilidad lograda, lo que nos permitió una primera conexión entre los conceptos de **heredabilidad** y **progreso genético**. Más adelante, en el capítulo «Selección artificial II» formalizamos la **ecuación del criador**, $R = S\,h^2$, que nos permitía predecir la respuesta (o progreso genético) a la selección direccional en una generación.

A partir de la ecuación del criador, teniendo en cuenta las definiciones de heredabilidad y de intensidad de selección (para características con distribución normal) derivamos la expresión

de la ecuación (11.51), $R = i\, h\, \sigma_A$ que nos muestra que la respuesta a la selección está directamente relacionada con la varianza aditiva de la característica, algo que reafirmamos en la ecuación (11.81) $R = \sigma_A^2 \frac{s}{2a}$ para el modelo de un locus con dos alelos, o en términos más generales en la **identidad de Robertson-Price**, ecuación (11.104) $R = \sigma_A^2 b_{w(z),z}$. En función de esto, no pasó desapercibido el hecho que características que poseían importante varianza aditiva y que se encontraban sujetas a una importante selección direccional no mostrasen casi respuesta a la selección. Dicho de otra manera, todo el marco teórico que venimos desarrollando hasta el momento deja de funcionar en ciertas circunstancias, por lo que es necesario entender las razones y atenderlas en un nuevo marco teórico. Una de las razones posibles a considerar es el hecho de que por lo general nunca la selección actúa en una sola dirección o es posible realizar selección exclusivamente por una característica. Por ejemplo, más allá de nuestro esfuerzo por seleccionar solo una característica de interés, la selección natural actúa de formas muchas veces desconocidas para nosotros y que por lo tanto no podemos evitar. Como vimos antes, existen bases genéticas para la correlación entre muchas características de interés y el éxito reproductivo de los individuos, por lo que muchas veces al seleccionar por determinada característica (es decir, incrementando el **fitness** de dicho fenotipo) estamos eligiendo individuos con menor **fitness** y que por lo tanto dejarán menos descendencia. Es por lo tanto fundamental el mayor conjunto de características posibles en nuestro estudio de la respuesta a la selección.

El primero en desarrollar una teoría formal de la selección multivariada fue Lande (1979), quien propuso que la ecuación del criador se podía extender fácilmente a una forma **multivariada** de acuerdo con la siguiente ecuación matricial

$$\mathbf{R} = \Delta \bar{z} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S} \tag{12.89}$$

donde ahora $\mathbf{R}$ es un vector de respuestas a la selección en cada una de las características (si tenemos k características entonces $\mathbf{R}$ es un vector de k elementos), $\mathbf{S}$ es un vector con los diferenciales de selección direccional en cada una de las características y $\mathbf{G}$ y $\mathbf{P}$ son las matrices de varianza-covarianza genética (aditiva) y fenotípica, respectivamente.

La relación de la ecuación (12.89) con la ecuación del criador es evidente. Si recordamos la definición de heredabilidad como coeficiente de varianzas, la misma era $h^2 = \frac{V_A}{V_P}$. En matrices no existe el concepto de división, pero sí el de multiplicación por la inversa, de tal manera que $\mathbf{P}\mathbf{P}^{-1} = \mathbf{I}$ (la matriz identidad). Por lo tanto, el producto de $\mathbf{G}$ por la inversa de $\mathbf{P}$ (es decir, $\mathbf{P}^{-1}$) es una especie de análogo al concepto de «dividir» la matriz de varianza-covarianza genética entre la matriz de varianza-covarianza fenotípica. Una forma de ver la equivalencia con el caso de considerar las k variables como no correlacionadas a ningún nivel (es decir, como k análisis univariados) es ver que en ese caso, tanto $\mathbf{G}$ como $\mathbf{P}$ serán **matrices diagonales** (las únicas entradas con valores distintos de cero están en la diagonal). En ese caso, la inversa de $\mathbf{P}$ será una matriz diagonal, con valores en la diagonal iguales a $\frac{1}{P_{i,i}} = \frac{1}{\sigma^2_{P_i}}$. Al multiplicar $\mathbf{G}\mathbf{P}^{-1}$, el resultado será una matriz diagonal con valores

$$\frac{G_{i,i}}{P_{i,i}} = \frac{\sigma^2_{G_i}}{\sigma^2_{P_i}} = h_i^2 \tag{12.90}$$

que no son otra cosa que las heredabilidades de cada una de las características. Como lo mencionamos antes para el caso univariado, la validez de la versión multivariada de la **ecuación del criador**, la ecuación (12.89), es dependiente de la distribución conjunta mutivariada normal entre fenotipos y valores de cría.

En el caso univariado, la **identidad de Roberston-Price** nos indicaba que el diferencial de selección era igual a $S = \sigma[z,w(z)]$, ecuación (11.101), es decir la covarianza entre el fenotipo y el **fitness** asociado al mismo. La extensión a multivariado nos indica que ahora **S** es un vector de covarianzas, por lo que, de acuerdo con la teoría de regresión múltiple, el producto

$$\mathbf{P}^{-1}\mathbf{S} = \beta \tag{12.91}$$

es el vector de coeficientes de regresión del **fitness** en los distintos fenotipos de las características bajo consideración. Es decir, el **fitness** global basado en los distintos fenotipos lo podemos describir como

$$w = a + \sum_{i=1}^{k} \beta_i z_i + e_i \tag{12.92}$$

donde los β_i son los coeficientes del gradiente de selección direccional. La idea es que los mismos representan el esfuerzo de selección en cada característica luego de haber removido las correlaciones fenotípicas con el resto de las características bajo selección.

Sustituyendo el resultado de la ecuación (12.91) en la ecuación (12.89), tenemos entonces que

$$\mathbf{R} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S} = \mathbf{G}\beta \tag{12.93}$$

que se conoce como la **ecuación de Lande**. Si los fenotipos son multivariados normales, Lande (1979) muestra que β es el gradiente del logaritmo del **fitness medio**, es decir

$$\beta = \mathbf{P^{-1}S} = \nabla ln\bar{\mathbf{W}} \tag{12.94}$$

donde $\bar{\mathbf{W}}$ es el vector de **fitness medios** (para las distintas características) y ∇ es el operador gradiente, $\nabla = (\frac{\partial}{\partial \bar{z}_1},...,\frac{\partial}{\partial \bar{z}_k})'$. Por lo tanto, al ser β el gradiente del logaritmo del fitness medio, **la dirección que maximizará el cambio en el fitness medio de la población es** precisamente **la dirección del vector** β.

Sin duda lo anterior es muy interesante, pero ¿cuál es la importancia práctica? Si miramos con más cuidado la **ecuación de Lande** vemos que más allá del gradiente de selección direccional que nosotros o la naturaleza les imponga a las características relevantes para desplazar las medias en el sentido deseado, la respuesta es una función de la matriz **G** y que por lo tanto tendrá sus propias restricciones. De hecho, de acuerdo con la **ecuación de Lande**, $\mathbf{R} = \mathbf{G}\beta$, la multiplicación del vector de gradiente β (es decir el «óptimo») por la matriz **G** producirá un vector de respuesta **R**, que en general no coincidirá en dirección con el vector β. En general, la multiplicación de un vector por una matriz produce tanto una rotación, como un escalado del vector resultante.

Una forma sencilla de evaluar la concordancia (o diferencia) en la dirección de dos vectores es examinar el ángulo entre ellos. Si el ángulo es de cero grados, entonces las direcciones de los dos vectores coinciden y la respuesta a la selección sigue exactamente la dirección del gradiente de selección (que si solo incluimos las características de interés para nosotros, representaría la dirección deseada en el progreso genético); por el contrario, si el ángulo es de 90 grados, la respuesta a la selección será totalmente ortogonal a nuestros deseos. En nuestro caso, los dos vectores en consideración son $\mathbf{R}$ y β, por lo que aplicando la definición de producto escalar tenemos que

$$\mathbf{R} \times \beta = ||\mathbf{R}|| \; ||\beta|| cos(\theta) \;\; \therefore \tag{12.95}$$

$$cos(\theta) = \frac{\mathbf{R} \times \beta}{||\mathbf{R}|| \; ||\beta||} \tag{12.96}$$

Por lo tanto, para saber el ángulo entre el vector del gradiente de selección y la respuesta a la selección en una generación alcanza aplicar a esta ecuación la función **arcocoseno**, lo que nos deja con

$$\theta = cos^{-1}(\frac{\mathbf{R} \times \beta}{||\mathbf{R}|| \; ||\beta||}) \tag{12.97}$$

Claramente, la matriz $\mathbf{G}$ es la que impone las restricciones a las direcciones en las que la selección tendrá respuesta, por lo que el análisis de la misma se vuelve esencial. De los cursos de álgebra lineal podemos recordar que llamábamos **valores propios** (c) y **vectores propios** ($\mathbf{v}$) (también conocidos como **eigenvalues** y **eigenvectors**, respectivamente) de una matriz $\mathbf{M}$ a aquellos que cumplían

$$\mathbf{Mv} = c\mathbf{v} \tag{12.98}$$

es decir, aquellos vectores $\mathbf{v}$ que al ser multiplicados por la matriz no cambiaban de dirección. El **valor propio** es el factor de escalado en la multiplicación. Utilizando esto, las direcciones de vectores que no cambiarán al ser multiplicados por $\mathbf{G}$ serán las de los **vectores propios** de la misma. Más aún, entre ellos, los que más responderán serán, en orden, los de mayor valor, por lo que el primer vector propio (el de mayor valor) será la dirección preferida para la evolución de acuerdo con las restricciones de la arquitectura genética de las características. Más aún, en el caso de que la matriz tenga valores propios iguales a cero, aplicando la definición de ecuación (12.98) a la matriz $\mathbf{G}$ tenemos

$$\mathbf{Gv} = c\mathbf{v} = \mathbf{0} \; \mathbf{v} = \mathbf{0} \tag{12.99}$$

lo que significa que no habrá respuesta ninguna en dicha dirección. Usualmente, en mejoramiento genético, se realizan estimaciones regulares de las varianzas y covarianzas genéticas, pero no existe una fuerte tradición en la interpretación de la evolución de las mismas a lo largo del tiempo, asumiendo de alguna forma que los cambios se deben a problemas de estimación. En realidad, a medida que avanzamos en la selección direccional las varianzas aditivas de las características de interés suelen disminuir, pero también se modifican las covarianzas entre las características, por lo que aun existiendo varianza aditiva disponible en ciertos casos no hay respuesta a la selección. Por lo tanto, es fundamental entender que al mismo tiempo que la respuesta está condicionada por la matriz $\mathbf{G}$, las restricciones también pueden evolucionar, cambiando la matriz $\mathbf{G}$ en un juego complejo y dinámico. Un tratamiento más profundo y al mismo tiempo ameno se encuentra en Blows y Walsh (2009).

Ejemplo 12.7

Las ovejas verdes tienen una tendencia importante a la esfericidad de sus cuerpos, lo que dificulta enormemente el esquilado, por lo que uno de los objetivos de selección planteados para mejorar la majada experimental es hacerlos más cilíndricos. Para ello se desarrolló un índice que es igual al diámetro de la panza en la zona media dividida entre el largo del cuerpo y multiplicado por 100, es decir

$$I_e = 100 \times \frac{D_{panza}}{L_{cuerpo}} \tag{12.100}$$

Al mismo tiempo, se plantea incrementar la producción de lana verde, por lo que debemos mantener los dos objetivos en mente. Un estudio preliminar determinó que la matriz de varianza-covarianza genética estaría dada por

$$\mathbf{G} = \begin{pmatrix} 100 & 0 \\ 0 & 800 \end{pmatrix} \tag{12.101}$$

con sus correspondientes unidades (la varianza del índice en unidades de índice al cuadrado y la producción de lana en g^2). Como se trata de una matriz diagonal, no existen en principio restricciones genéticas. Sin embargo, veamos qué pasa si elegimos un gradiente de selección

$$\beta = \begin{pmatrix} -0{,}09 \\ 0{,}03 \end{pmatrix} \tag{12.102}$$

La respuesta a la selección de acuerdo con la **ecuación de Lande** será

$$R = \mathbf{G}\beta = \begin{pmatrix} 100 & 0 \\ 0 & 800 \end{pmatrix} \begin{pmatrix} -0{,}09 \\ 0{,}03 \end{pmatrix} = \begin{pmatrix} -9 \\ 24 \end{pmatrix} \tag{12.103}$$

Si calculamos ahora el ángulo entre los dos vectores, usando la ecuación (12.97) tenemos que

$$\theta = cos^{-1}(\frac{\mathbf{R} \times \beta}{||\mathbf{R}||\ ||\beta||}) = \tag{12.104}$$

$$cos^{-1}(\frac{1{,}53}{2{,}431666}) = 51° \tag{12.105}$$

lo que pese a la inexistencia de correlación genética entre ambas características nos deja con un desfasaje importante entre la dirección de selección elegida y la favorecida por **G**. Esto se debe a que, pese a que no existe correlación genética entre las características, la respuesta se encuentra rotada respecto a la dirección óptima para el incremento del **fitness medio**. La dirección de máxima variación es en peso del vellón, pero en términos relativos estamos haciendo mucha más presión de selección en una dirección (índice de esfericidad) con poca variación, por lo que no debería sorprendernos esta diferencia en las direcciones de gradiente de selección y de respuesta.

Un segundo estudio, sin embargo, con muchos más animales obtuvo un diferente estimado de la matriz de varianzas-covarianzas genéticas, la que fue ahora

$$\mathbf{G} = \begin{pmatrix} 100 & 282{,}8 \\ 282{,}8 & 800 \end{pmatrix} \tag{12.106}$$

Ahora, si utilizamos el mismo gradiente de selección que antes, la respuesta a la selección predicha por la **ecuación de Lande** será

$$\mathbf{R} = \mathbf{G}\beta = \begin{pmatrix} 100 & 282{,}8 \\ 282{,}8 & 800 \end{pmatrix} \begin{pmatrix} -0{,}09 \\ 0{,}03 \end{pmatrix} = \begin{pmatrix} -0{,}5151 \\ -1{,}4547 \end{pmatrix} \tag{12.107}$$

Claramente, la respuesta a la selección en este caso no parece haber seguido siquiera cerca nuestros objetivos. Si calculamos el ángulo ahora entre los dos vectores, tenemos que el mismo es de

$$\theta = cos^{-1}(\frac{\mathbf{R} \times \beta}{||\mathbf{R}||\ ||\beta||}) = \tag{12.108}$$

$$cos^{-1}\left(\frac{0{,}002718}{0{,}1464012}\right) = 88°93' \tag{12.109}$$

lo que indica una restricción casi total (cercana a $\theta = \mathbf{90}°$). La explicación de esto es muy sencilla, la matriz $\mathbf{G}$ posee uno de los valores propios muy cercanos a cero y cuyo vector propio es

$$\begin{pmatrix} -0{,}9428 \\ 0{,}3333 \end{pmatrix} \tag{12.110}$$

lo que coincide casi exactamente con la dirección elegida para la selección, por lo que esta no tendrá respuesta. Esto no debería sorprendernos, ya que en este caso, la correlación genética entre ambas características, de acuerdo con la ecuación (12.75) es de

$$r_A = \frac{\sigma_{A_{i,j}}}{\sigma_{A_i}\sigma_{A_j}} = \frac{282{,}8}{\sqrt{100 \times 800}} = 0{,}9999 \tag{12.111}$$

es decir, queda muy poco espacio disponible para selección de ambas características en sentido opuesto.

PARA RECORDAR

- Podemos extender la ecuación del criador a una forma **multivariada** de acuerdo con la siguiente ecuación matricial: $\mathbf{R} = \Delta\bar{z} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S}$ donde ahora $\mathbf{R}$ es un vector de respuestas a la selección en cada una de las características (si tenemos k características entonces $\mathbf{R}$ es un vector de k elementos), $\mathbf{S}$ es un vector con los diferenciales de selección direccional en cada una de las características y $\mathbf{G}$ y $\mathbf{P}$ son las matrices de varianza-covarianza genética (aditiva) y fenotípica, respectivamente.
- La **ecuación de Lande** ($\mathbf{R} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S} = \mathbf{G}\beta$) nos muestra que más allá del gradiente de selección direccional que nosotros o la naturaleza le impongan a las características relevantes para desplazar las medias en el sentido deseado, la respuesta es una función de la matriz $\mathbf{G}$ y que por lo tanto tendrá sus propias restricciones. De hecho, de acuerdo con esta, $\mathbf{R} = \mathbf{G}\beta$, la multiplicación del vector de gradiente β (es decir el «óptimo») por la matriz $\mathbf{G}$ producirá un vector de respuesta $\mathbf{R}$, que en general no coincidirá en dirección con el vector β. En general, la multiplicación de un vector por una matriz produce tanto una rotación, como un escalado del vector resultante.

12.8 Conclusión

Como vimos anteriormente, podemos realizar una influencia direccional sobre una característica fenotípica, en un proceso que denominamos selección artificial. La eficiencia de esta selección depende tanto de la intensidad de la misma como de la heredabilidad del carácter fenotípico considerado. No obstante, en nuestro análisis dejamos de lado que la selección de un carácter puede influir en otros, fenómeno que puede ser o no deseado según el caso. ¿A qué se debe esta dependencia entre caracteres? Fenómenos como la *pleiotropía* o la influencia de factores ambientales comunes pueden subyacer a esta respuesta correlacionada entre caracteres fenotípicos. Incluso fenómenos que analizamos en capítulos anteriores, como el desequilibrio de ligamiento, pueden ser la causa subyacente a estas correlaciones, lo cual hace que el fenómeno pueda estar influido por la historia de cada población particular. Entender las causas y el comportamiento de estas respuestas correlacionadas es fundamental para los propósitos del mejoramiento genético. Para entender el por qué, podemos pensar en los casos extremos: dos caracteres que se encuentran totalmente correlacionados, o dos en los que no existe ningún tipo de correlación. Resulta claro que es necesario contemplar estas restricciones genético-evolutivas tanto si se tiene interés en ambos caracteres, o si uno lo es pero el otro no. En estos casos resulta informativo incluso poseer conocimiento simplemente del signo de la correlación entre estos caracteres fenotípicos. Si bien es difícil (y en algunos casos, imposible), el separar los componentes genéticos y ambientales de las respuestas correlacionadas, los análisis que permiten tal separación son un aporte fundamental, tanto en el plano práctico como conceptual, de la genética cuantitativa.

12.9 Actividades

12.9.1 Control de lectura

1. ¿Qué tipos de correlaciones existen?
2. ¿Cómo se tienen en cuenta los diferentes tipos de correlaciones al momento de analizar la respuesta a selección?
3. ¿Qué fenómenos pueden subyacer a la existencia de correlación entre dos variables fenotípicas? Expláyate para alguno de estos fenómenos
4. En algunos casos es preferible optar por la selección indirecta de una característica (en lugar de seleccionarla directamente). ¿Cuál es el criterio determinante para tomar esta decisión?
5. Parte de la correlación entre dos características fenotípicas es explicable por la existencia de un componente genético aditivo compartido. No obstante, en la práctica solo podemos observar correlaciones fenotípicas. ¿Qué formas conoces para estimar la correlación aditiva entre dos características?

12.9.2 ¿Verdadero o falso?

1. La correlación fenotípica no es la simple suma de la correlación genotípica y ambiental.
2. Utilizando el *path analysis* es posible demostrar que el coeficiente de correlación estandarizado entre dos fenotipos equivale a la heredabilidad.
3. Se puede realizar selección para una característica de interés seleccionando sobre otra si existe correlación fenotípica entre estas.
4. En casos donde la correlación genética (aditiva) entre dos características es mayor a la precisión entre ambas, conviene optar por la selección indirecta.
5. A partir de una matriz de varianza-covarianza genética (aditiva) para una característica fenotípica de interés es posible calcular su heredabilidad.

Soluciones

1. **Verdadero.** Se puede demostrar que la correlación fenotípica entre dos características involucra a las correlaciones genotípicas y ambientales de forma más compleja que una suma.
2. **Falso.** Recordemos que este coeficiente equivale a h ($\sqrt{h^2}$), no a la heredabilidad (h^2).
3. **Falso.** La existencia de correlación fenotípica no implica necesariamente que ambos caracteres fenotípicos aumenten en la misma dirección; es necesario tener en cuenta el signo de dicha correlación. Resulta claro que si la misma es negativa será imposible seleccionar la característica de interés a partir de la selección de la característica correlacionada (de hecho, se logrará lo contrario a lo deseado).
4. **Verdadero**.
5. **Falso**. Recordemos que la heredabilidad equivale a $h^2 = \frac{\sigma_A}{\sigma_P}$, por lo que para estimar la misma para una característica se requiere también de la matriz de varianza-covarianza fenotípica.

12.9.3 Ejercicios

Ejercicio 12.1

Un grupo de laboratorio se encuentra estudiando una enfermedad rara (es decir„ de baja incidencia en poblaciones celulares comunes). La selección de células que presentan el fenotipo resulta engorrosa, por lo que se plantea la posibilidad de realizar un proceso de selección indirecta, a efectos de enriquecer los cultivos celulares en dichas células. La idea es seleccionar células según su capacidad de metabolización de diferentes medios de cultivo (algo que es fácil de manipular a nivel experimental).

Un estudio previo estimó la matriz de varianza-covarianza genético-aditiva para tres fenotipos: F_E (fenotipo de la enfermedad), F_{M1} (alta capacidad de metabolización de un compuesto) y F_{M2} (alta capacidad de metabolización de otro compuesto). Esta se muestra a continuación

	F_E	F_{M1}	F_{M2}
F_E	0,23	0,76	0,15
F_{M1}	0,76	0,12	0,21
F_{M2}	0,15	0,21	0,1

A su vez, el mismo estudio estima que las heredabilidades de estos tres fenotipos son $h^2_E = 0{,}69$, $h^2_{\text{M1}} = 0{,}82$ y $h^2_{\text{M2}} = 0{,}55$

¿Conviene al grupo realizar selección indirecta para este fenotipo? ¿Empleando cuál de los metabolitos?

Solución

Recordemos que para que sea preferible utilizar el método de selección indirecta (X) sobre una característica seleccionando directamente a otra (Y) se debe cumplir

$$r_{A_{XY}} > \frac{h_X}{h_Y} \tag{12.112}$$

En este caso puntual, debemos tener en cuenta dos comparaciones (las que contemplan a F_E)

Tenemos que

$$h_E = \sqrt{h^2_E} = \sqrt{0{,}69} \tag{12.113}$$

$$h_{M1} = \sqrt{h^2_{M1}} = \sqrt{0{,}82} \tag{12.114}$$

$$h_{M2} = \sqrt{h^2_{M2}} = \sqrt{0{,}55} \tag{12.115}$$

A su vez los coeficientes de correlación genético aditivas pueden calcularse a partir de la matriz de varianza-covarianza, recordando que

$$r_{A_{XY}} = \frac{\text{Cov}(X,Y)}{\sqrt{\sigma^2_{AX}} \times \sqrt{\sigma^2_{AY}}} \tag{12.116}$$

De esto tenemos

$$\begin{cases} r_{A_{E-M1}} = \frac{0,76}{\sqrt{0,23}\times\sqrt{0,12}} \approx 4,58 \\ r_{A_{E-M2}} = \frac{0,15}{\sqrt{0,23}\times\sqrt{0,1}} \approx 0,989 \end{cases} \tag{12.117}$$

Notemos que para la primera comparación se da $r_{A_{E-M1}} > \frac{h_E}{h_{M1}} = \frac{\sqrt{0,69}}{\sqrt{0,82}} \approx 0,917$, mientras que en el segundo caso tenemos $r_{A_{E-M1}} < \frac{h_E}{h_{M2}} = \frac{\sqrt{0,69}}{\sqrt{0,55}} \approx 1,12$. De esto se deduce que el grupo de investigación sí puede optar por una técnica de selección indirecta, siempre y cuando opte por emplear como régimen selectivo la capacidad de metabolización del compuesto $M1$.

Ejercicio 12.2

(Adaptado de Blows and Walsh 2009)

Un productor se encuentra intentando mejorar un carácter fenotípico A mediante selección artificial. Otro carácter fenotípico, B, también es registrado en las mediciones.

a. Las varianzas aditivas estimadas para los caracteres son $\sigma^2_A = 10$ y $\sigma^2_B = 40$. Además, se asume que el régimen selectivo afecta a estos dos caracteres de forma que el cambio en dos unidades en el carácter A lleva a la disminución en una unidad del carácter B. ¿Cuál es la respuesta esperada para cada carácter, si asumimos que no existe covarianza aditiva entre ambos?

b. Un estudio revela que la asunción de inexistencia en la covarianza genético-aditiva entre los dos caracteres es demasiado ingenua. Estos sí están correlacionados, siendo la matriz de varianza-covarianza genético-aditiva

$$\mathbf{G} = \begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix} \tag{12.118}$$

¿Cuál es la respuesta esperada entonces? ¿Cómo explicas conceptualmente el resultado obtenido?

Solución

a. Se pide calcular la respuesta a la selección en más de un carácter fenotípico. En estos casos se puede aplicar la ecuación de Robertson-Price generalizada

$$\mathbf{R} = \mathbf{G}\beta \tag{12.119}$$

donde $\mathbf{G}$ es la matriz de varianza-covarianza genético-aditiva y β es la respuesta a selección de los caracteres (es decir, $\beta = \mathbf{P}^{-1}\mathbf{S}$)

En nuestro caso se plantea que

$$\mathbf{G} = \begin{pmatrix} 10 & 0 \\ 0 & 40 \end{pmatrix} \tag{12.120}$$

y

$$\beta = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \tag{12.121}$$

Tenemos entonces una respuesta esperada **R** de

$$\mathbf{R} = \begin{pmatrix} 10 & 0 \\ 0 & 40 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} \tag{12.122}$$

$$\mathbf{R} = \begin{pmatrix} 20 \\ -40 \end{pmatrix} \tag{12.123}$$

Notemos que la respuesta a selección no es la más eficiente (ya que debería ser un vector alineado a β, múltiplo de este), pero se esperaría un cambio en los valores fenotípicos de estas características.

b. La existencia de covarianza genético-aditiva cambia la situación: ahora el cambio en un carácter influye necesariamente en la respuesta a selección del otro. Ahora tenemos:

$$\mathbf{R} = \begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} \tag{12.124}$$

$$\mathbf{R} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{12.125}$$

¡Es decir que en este caso no tenemos respuesta a la selección! ¿Cómo es esto posible? Este resultado deja en evidencia un aspecto importante de la respuesta a la selección: los organismos están constituidos por múltiples caracteres fenotípicos, por lo que un régimen selectivo operará afectando varios de los mismos; lo que es más, la existencia de correlación genético-aditiva y fenotípica entre estos caracteres hace que la respuesta de una variable afecte a la respuesta de otra. En un caso extremo, como el presente, dos caracteres pueden estar siendo seleccionados pero influyendo a la respuesta del otro de tal forma que no exista una respuesta final a la selección.

Ejercicio 12.3

En una majada comercial, existe interés en mejorar el peso de los borregos al año de edad (PB; $h^2 = 0{,}32$; $\sigma^2_A = 8\,\text{kg}^2$). Sin embargo, debido a que se quiere realizar una selección más temprana de los animales se ha decidido seleccionar a través del peso al destete (PD; $h^2 = 0{,}47$; $\sigma^2_A = 6\,\text{kg}^2$) de estos animales. A la vez, la correlación genético-aditiva entre las características es de 0,7 y la intensidad de selección posible es de 2,4209 para los machos y 0,6439 para las hembras.

¿Cuánto será el progreso genético por generación en PB si se selecciona por PD?

Solución

Para poder estimar el progreso genético que es posible lograr en PB seleccionando por PD usaremos

$$RC_Y = i_X \; h_X \; r_{A_{XY}} \; \sigma_{A_Y} \tag{12.126}$$

Dónde Y corresponde a la característica de nuestro interés (PB) y X a la característica mediante la cual seleccionamos (PD).

Aplicando entonces la fórmula

$$RC_{PB} = \frac{2{,}4209 + 0{,}6439}{2} \times \sqrt{0{,}47} \times 0{,}7 \times \sqrt{8} \approx 2{,}08 \text{ kg} \tag{12.127}$$

Es decir, por generación, al seleccionar por peso al destete, es posible mejorar algo más de 2 kg el peso del borrego al año.

Ejercicio 12.4

El búfalo de agua (*Bubalus bubalis*) es un gran bóvido de origen asiático pero que actualmente se ha diseminado virtualmente por todo el mundo. Este animal se puede encontrar en estado salvaje o doméstico, dónde la presencia de astas presenta dificultades a la hora del manejo especialmente dado el gran tamaño de estos animales. Se ha observado que la correlación fenotípica entre el tamaño de cuerpo (TC) y el tamaño de astas (TA) es de $r_P = 0{,}68$ mientras que las heredabilidades de las características son $h^2_{TC} = 0{,}48$ y $h^2_{TA} = 0{,}56$. Por otra parte, se ha determinado que el mejor estimado de la correlación genética (aditiva) entre ambas características es de $r_A = 0{,}47$.

¿Cuál es la correlación ambiental entre las características?

Solución

Recordemos que podemos expresar la relación entre las correlaciones fenotípica, genética y ambiental como $r_P = h_1 h_2 r_A + e_1 e_2 r_E$ donde podremos despejar la correlación ambiental tal que $r_E = \frac{r_P - h_1 h_2 r_A}{e_1 e_2}$.

Procederemos entonces a calcular los parámetros e

$$e_1 = \sqrt{1 - h_1^2} = \sqrt{1 - 0{,}48} = \sqrt{0{,}52} = 0{,}7211 \tag{12.128}$$

$$e_1 = \sqrt{1 - h_2^2} = \sqrt{1 - 0{,}56} = \sqrt{0{,}44} = 0{,}6633 \tag{12.129}$$

Entonces, la correlación ambiental entre las características será

$$r_E = \frac{0{,}68 - \sqrt{0{,}48} \times \sqrt{0{,}56} \times 0{,}47}{0{,}7211 \times 0{,}6633} = 0{,}913 \tag{12.130}$$

Ejercicio 12.5

(Modificado de D. S. Falconer and MacKay 1996; trabajo original de Reeve and Robertson 1953)

En un antiguo estudio (Reeve and Robertson 1953) Reeve y Robertson se dedicaron a estudiar la correlación genético-aditiva en dos caracteres fenotípicos de la mosca *D. melanogaster* vinculados al tamaño corporal: largo de ala y largo de tórax. Para ello realizaron experimentos con dos líneas de moscas. Al seleccionar individuos a efectos de generar una línea con mayor largo de ala (con un diferencial de selección de $S_{\text{ala}} = 0{,}2$ cm), los investigadores observan un aumento del largo de tórax. La covarianza entre ambos caracteres fenotípicos fue de $\text{Cov}_{\text{ala}-\text{tórax}} = 0{,}726$, con una varianza para largo de alas de $\sigma^2_{\text{ala}} = 1$ y de $\sigma^2_{\text{tórax}} = 2$ para largo de tórax. Por otro lado, la selección de individuos con el fin de aumentar el largo de tórax (con un diferencial de selección de $S_{\text{tórax}} = 1{,}3$ cm) también resultó en un aumento del largo de alas, esta vez con covarianza $\text{Cov}_{\text{tórax}-\text{ala}} = 0{,}679$, siendo las varianzas fenotípicas $\sigma^2_{\text{ala}} = 1{,}2$ y $\sigma^2_{\text{tórax}} = 1{,}8$

Si un estudio previo arroja que la heredabilidad del largo de ala y tórax es de $h^2_{\text{ala}} = 0{,}74$ y $h^2_{\text{tórax}} = 0{,}69$, respectivamente, ¿cuál es la correlación genético-aditiva entre ambas características?

Nota: A efectos de simplificar los cálculos se asume que no existe varianza genética de dominancia, y que la varianza ambiental es despreciable.

Solución

Recordemos que la correlación aditiva entre dos caracteres fenotípicos se puede calcular según

$$r_a^2 = \frac{\mathrm{CR}_{T2}}{R_{T1}} \times \frac{\mathrm{CR}_{T1}}{R_{T2}} \tag{12.131}$$

donde la respuesta correlacionada del carácter $T2$ al cambiar el carácter $T1$ está dada por $\mathrm{CR}_{T2} = ih_{T1}h_{T2}r_a\sigma_{P,T2}$ (y viceversa si se intercambian los roles), $r_a = \frac{\mathrm{Cov}_a(T1,T2)}{\sigma_{a,T1}\times\sigma_{a,T2}}$ y siendo la respuesta de selección directa dada por $R_{\cdot} = ih_{\cdot}\sigma_{A_{\cdot}}$.

Pasemos por lo tanto a calcular cada una de estas cantidades, a efectos de poder calcular la correlación genético-aditiva entre largo de ala y largo de tórax.

Empecemos por la precisión de cada uno de los caracteres. Tenemos

$$\begin{cases} h_{\text{ala}} = \sqrt{h^2_{\text{ala}}} \Rightarrow h_{\text{ala}} = \sqrt{0{,}74} \\ h_{\text{tórax}} = \sqrt{h^2_{\text{tórax}}} \Rightarrow h_{\text{tórax}} = \sqrt{0{,}69} \end{cases} \tag{12.132}$$

Luego tenemos la respuesta a la selección directa de cada característica, dadas por

$$\begin{cases} R_{\text{ala}} = 0{,}2 \times \sqrt{0{,}74} \times \sqrt{1} \Rightarrow R_{\text{ala}} = 0{,}2 \times \sqrt{0{,}74} \\ R_{\text{tórax}} = (\frac{0{,}15}{\sqrt{1{,}8}}) \times \sqrt{0{,}69} \times \sqrt{1{,}8} \end{cases} \tag{12.133}$$

donde las intensidades de selección fueron calculadas según $i_{\text{ala}} = \frac{S_{\text{ala}}}{\sigma_{P,\text{ala}}} = \frac{0{,}2}{\sqrt{1}} = 0{,}2$ e $i_{\text{tórax}} = \frac{S_{\text{tórax}}}{\sigma_{P,\text{tórax}}} = \frac{0{,}15}{\sqrt{1{,}8}}$

Podemos ahora calcular las respuestas correlacionadas para cada carácter

$$\begin{aligned} \mathrm{CR}_{\text{ala}} &= i\, h_{\text{ala}}\, h_{\text{tórax}}\, r_a\, \sigma_{P,\text{ala}} = \left(\frac{1{,}3}{\sqrt{1{,}8}}\right)\sqrt{0{,}84}\sqrt{0{,}79}\,\frac{0{,}879}{\sqrt{1{,}2}\sqrt{1{,}8}}\,\sqrt{1{,}2} \Rightarrow \mathrm{CR}_{\text{ala}} \approx 0{,}517146 \\ \mathrm{CR}_{\text{tórax}} &= i\, h_{\text{ala}}\, h_{\text{tórax}}\, r_a\, \sigma_{P,\text{tórax}} = (1{,}2)\sqrt{0{,}84}\sqrt{0{,}79}\,\frac{0{,}726}{\sqrt{1}\sqrt{2}}\,\sqrt{2} \Rightarrow \mathrm{CR}_{\text{tórax}} = 0{,}709694 \end{aligned} \tag{12.134}$$

donde tenemos en cuenta el efecto de cambio en una característica al haber seleccionado directamente a la otra.

A partir de los datos podemos calcular la correlación genético-aditiva entre características

$$r_a^2 = \frac{\mathrm{CR}_{\text{ala}}}{R_{\text{tórax}}} \times \frac{\mathrm{CR}_{\text{tórax}}}{R_{\text{ala}}} \tag{12.135}$$

$$r_a^2 \approx (\frac{0{,}517146}{(\frac{1{,}3}{\sqrt{0{,}8}}) \times \sqrt{0{,}79} \times \sqrt{1{,}8}}) \times (\frac{0{,}709694}{1{,}2 \times \sqrt{0{,}84}}) \tag{12.136}$$

$$r_a^2 \approx 0{,}193 \tag{12.137}$$

Capítulo 13

Selección artificial II

Perdida en este tedio social
Busco mi justo medio ancestral
Cósmicamente fuerzas de amor en mí
Entonces nos venden mística de moda
El Bhagavad Gita en joda
¿Ves y no entendés nada?
Aquí las vacas no son sagradas
Van a parar al asador
Aquí creemos en nuestro Señor
Santo Gardel del cielo
Santo Gardel del cielo.[1]

Tabaré Rivero

En capítulos anteriores hemos avanzado en la construcción de las estrategias para mejorar desde el punto de vista genético características de interés para nosotros, primero basados en el fenotipo de los individuos que son candidatos a la selección en la misma característica de interés y luego a través de otras características correlacionadas, pero siempre basados en la información fenotípica del candidato. Una excepción a esto fue nuestra definición de valor de cría de un individuo, cuando la consideramos la mitad del desvío fenotípico promedio de los hijos del candidato, cuando este se apareaba con un número importante de individuos del otro sexo escogidos al azar. En este caso, estamos utilizando la información de la progenie del candidato (el criterio de selección) para mejorar una determinada característica.

En el presente capítulo vamos a extender nuestros conocimientos sobre las estrategias de selección posibles, en una primera parte utilizando información proveniente de parientes del individuo, mientras que en la segunda parte vamos a discutir cómo seleccionar por diferentes características a la vez. Cerraremos el capítulo discutiendo sobre métodos «modernos» para la predicción del valor de cría, los que permiten incorporar de forma «automática» la información aportada por parientes del candidato, así como la evaluación de muchas características a la vez.

Como lo notarás en breve, el eje articulador del capítulo son los índices de selección, en particular los índices lineales. Hay una extensa teoría disponible sobre ellos, son sencillos de comprender, fáciles de usar y su interpretación es intuitiva. Para ello haremos uso extensivo de herramientas básicas de álgebra lineal ya que trabajaremos con las matrices de varianza-covarianza, tanto fenotípica como genética, que vimos en la sección «Matrices de varianza-covarianza» del capítulo «Correlaciones y respuesta correlacionada».

[1] De la canción «El cosmonauta interior», disco *Yoganarquía*, de la Tabaré Riverock Banda, de 1997. Aunque es discutible que aquí las vacas no sean sagradas, lo seguro es que gracias a la selección genética, cada día van mejor a parar al asador.

Muchos de los métodos de selección que discutiremos en el capítulo no se utilizan actualmente en la práctica ya que han sido ampliamente superados por metodologías más modernas de selección, que permiten trabajar en forma unificada los datos. Ejemplos de esto son la selección familiar y la selección dentro de familias. Sin embargo, su presentación aquí sigue jugando un papel muy importante en comprender cómo los distintos grados de parentesco aportan información diferente y si bien como métodos de selección han sido superados, su comprensión facilita la comprensión futura de las técnicas más avanzadas, como los modelos lineales mixtos y los estimados BLUP.

OBJETIVOS DEL CAPÍTULO

☐ Entender los criterios, objetivos fundamentales y el uso de información en el proceso de selección genética.

☐ Integrar la información proveniente de diferentes parientes para mejorar la precisión de la selección.

☐ Estudiar cómo se calculan las matrices de varianza-covarianza fenotípica y genética.

☐ Analizar la utilidad y limitaciones de las pruebas de progenie en la selección genética.

☐ Evaluar la precisión del valor de cría derivado de índices de selección.

☐ Comprender los métodos de selección para múltiples características y cómo se combinan estas en índices.

☐ Explorar técnicas avanzadas para la predicción del valor de cría, como BLUP y modelos mixtos lineales, y su aplicación en la mejora genética.

13.1 Criterios y objetivos de selección

En parte importante de este capítulo vamos a trabajar con la selección por varias características, mientras que en el resto trabajaremos con selección en una característica a partir de información obtenida de parientes de los candidatos. En ambos casos es importante entender la distinción que existe entre **objetivos de selección** y **criterios de selección**.

Objetivo de selección define a la característica que queremos mejorar a través de la selección, aquella característica que presenta relevancia propia para nosotros.

En el caso de la producción agropecuaria, los objetivos de selección se vinculan, en general, directamente a los ingresos o egresos del establecimiento. Por ejemplo, si se trata de

producción lechera, el sistema de pago en nuestro país incluye los kilogramos de proteína y los kilogramos de grasa en la **paramétrica de pago**. Por lo tanto, si queremos incrementar los ingresos del establecimiento una buena opción sería mejorar la cantidad de grasa y proteína que genera cada vaca en una lactancia. Como ambas características, producción de grasa y de proteína, poseen una fuerte base genética, entonces resulta razonable plantearse mejorar genéticamente dichas características. Por lo tanto, en caso de elegirlas, las mismas constituyen o**bjetivos de selección** claros.

En algunos casos será posible hacer selección directamente por la característica, basados en el fenotipo del candidato. En otros casos esto resultará imposible, por ejemplo, en el caso de que una característica se exprese en el otro sexo (respecto a los candidatos a la selección). En este caso, será necesario recurrir a otras características correlacionadas, a información de parientes o ambas combinadas. En ese caso, los objetivos de selección y los criterios de selección no son los mismos.

Criterios de selección son los criterios que se utilizan como fuente de información para mejorar los objetivos de selección. Pueden coincidir o no con los objetivos de selección.

En muchos casos, los criterios de selección no tienen ningún valor *per se* para los ingresos y egresos del establecimiento y simplemente se trata de fuentes de información disponibles para mejorar los objetivos de selección. En otros casos se trata del mismo objetivo de selección, lo que simplifica muchísimos la tarea de seleccionar. En general, en las características fáciles de observar y registrar, que se expresan en ambos sexos y que poseen una heredabilidad razonable el **objetivo de selección coincide con el criterio de selección**.

Para que un criterio de selección sea usable para mejorar otras características es fundamental que el mismo sea relativamente fácil y barato de medir, que posea una heredabilidad razonable, pero fundamentalmente que se encuentre correlacionado desde el punto de vista genético con el objetivo que se desea mejorar.

Ejemplo 13.1

Objetivos y criterios de selección suelen ser diferentes. En general, la razón de esto es que la característica de interés es difícil de medir directamente en los individuos candidatos. Por ejemplo, porque el animal no expresa la característica (leche en machos). En ese caso, una alternativa es realizar la medición en parientes.

En la siguiente tabla, solo a modo de ejemplo, se presentan algunas alternativas para mejorar una característica mediante información de parientes (medio hermanas o hijas) de los candidatos a la selección ($h^2 = 0{,}20$):

Alternativa	i	r_{AC}	IG	$\frac{i \times r_{AC}}{IG}$
Medir 5 medio hermanas	1,03	0,228	4,3	0,055
Medir 10 medio hermanas	0,88	0,294	4,3	0,060
Medir 20 medio hermanas	0,70	0,358	4,3	0,058
Medir 30 medio hermanas	0,58	0,391	4,3	0,053
Medir 5 hijas	1,03	0,456	6,3	0,075
Medir 10 hijas	0,88	0,587	6,3	0,082
Medir 20 hijas	0,70	0,716	6,3	0,080
Medir 30 hijas	0,58	0,782	6,3	0,072

Como la respuesta anual depende de $\frac{i \times r_{AC} \times \sigma_A}{IG}$, pero σ_A es igual para todas las alternativas, entonces nos alcanza con optimizar $\frac{i \times r_{AC}}{IG}$ (última columna de la tabla).

PARA RECORDAR

- El **objetivo de selección** define a la **característica que queremos mejorar** a través de la selección, aquella característica que presenta relevancia propia para nosotros. Generalmente, estos objetivos están **directamente vinculados** a los ingresos o egresos del establecimiento.
- Los **criterios de selección** son los criterios que se utilizan como **fuente de información** para mejorar los objetivos de selección. Pueden coincidir o no con los objetivos de selección. En muchos casos, los criterios de selección no tienen ningún valor en sí para los ingresos y egresos del establecimiento y simplemente se tratan de fuentes de información disponibles para mejorar los objetivos de selección.
- En algunos casos será posible hacer selección directamente por la característica, basados en el fenotipo del candidato, en otros casos esto resultará imposible. Entonces, será necesario recurrir a otras características correlacionadas, a información de parientes o ambas combinadas. En ese caso, **los objetivos de selección y los criterios de selección no son los mismos.**
- En general, en las **características fáciles de observar y registrar, que se expresan en ambos sexos y que poseen una heredabilidad razonable el objetivo de selección coincide con el criterio de selección.**
- Para que un **criterio de selección** sea usable para mejorar otras características es fundamental que el mismo sea **relativamente fácil y barato de medir,**

que posea una heredabilidad razonable, pero fundamentalmente que se encuentre correlacionado desde el punto de vista genético con el objetivo que se desea mejorar.

13.2 Selección basada en un solo tipo de fuente de información

En el método de selección individual la idea es utilizar la información fenotípica aportada por el individuo como criterio para la decisión sobre el valor reproductivo del mismo. Este método es extremadamente sencillo de implementar y es el método sobre el que construimos el capítulo «Selección artificial I», donde trabajamos con la **selección por umbral** y que a veces llamamos **selección masal**. Se trata de un método que funciona muy bien cuando la información fenotípica es fácil de colectar y la heredabilidad de la característica es media o alta. Sin embargo, en situaciones donde el fenotipo individual es difícil o aún imposible de obtener (por ejemplo, cuando la característica se expresa en solo un sexo), o cuando la heredabilidad de la característica es baja, aún podemos recurrir a la selección usando otras fuentes de información. En el capítulo «Correlaciones y respuesta correlacionada» vimos cómo utilizar la información de una característica fuertemente correlacionada medida en el mismo individuo para mejorar la eficiencia de la selección. En lo que sigue veremos cómo utilizar la información brindada por algunas estructuras de parentesco particulares y más adelante en este capítulo veremos cómo combinar la información de diferentes relaciones de parentesco. Esta sección está basada en el tratamiento del tema dado por Falconer y Mackay (1996), que es además una excelente referencia sobre el tema.

La idea por detrás de la presente sección (y de otras) es que el fenotipo de un individuo podemos descomponerlo en la suma de dos fuentes diferentes, una la correspondiente al desvío de la media familiar respecto a la media de la población, que llamaremos P_f y otra que corresponde al desvío fenotípico del individuo respecto a la media familiar, que llamaremos P_w (la $_w$ correspondiendo al inglés *within*, dentro). Por lo tanto, el fenotipo de un individuo será

$$P = P_f + P_w \tag{13.1}$$

Ambos componentes aportan cosas diferentes, dependiendo de las relaciones familiares y de los parámetros genéticos, por lo que la idea a continuación será ver los distintos tipos de selección posibles usando estructuras de parentesco definidas y la respuesta a la selección consiguiente. En el caso de selección individual le hemos dado un peso equivalente a ambos componentes, P_f y P_w. Veremos entonces que existe un tipo de selección que llamaremos **selección familiar** donde solo ponderaremos el aporte de P_f, mientras que en otro tipo de selección, que llamaremos **selección intrafamiliar** solo ponderaremos el aporte de P_w.

Diferentes tipos de selección basados en estructuras de parentesco definidas

Selección familiar La idea en selección familiar es darle todo el peso a P_f, dejando de lado el aporte de P_w y de esta manera eligiendo familias enteras como reproductores. Si bien podría parecer una idea terrible a primera vista, la lógica de este tipo de selección es que **los efectos ambientales tienden a cancelarse dentro de la familia**, lo que llevaría a que la media fenotípica de la misma sea similar a la media genotípica. Esto es particularmente interesante cuando los efectos ambientales son una parte importante de la variabilidad fenotípica, es decir, la heredabilidad es baja, ya que el fenotipo individual no es un buen indicador del mérito genético.

En el caso de que existan **efectos del ambiente común**, estos afectarán los promedios familiares, por lo que los promedios familiares tampoco serían una buena indicación del mérito genético. Claramente, además, el funcionamiento de esta lógica es dependiente del número de individuos evaluados por familia, ya que un número bajo dará promedios familiares con mucho error, reduciendo las ventajas del método. Este tipo de selección es bastante común en peces, donde se crían numerosas familias y se eligen, por ejemplo, aquellas de mayor rendimiento. En el caso de que existan **efectos del ambiente común**, estos afectarán los promedios familiares, por lo que los promedios familiares tampoco serían una buena indicación del mérito genético. Claramente, además, el funcionamiento de esta lógica es dependiente del número de individuos evaluados por familia, ya que un número bajo dará promedios familiares con mucho error, reduciendo las ventajas del método. Este tipo de selección es bastante común en peces, donde se crían numerosas familias y se eligen, por ejemplo, aquellas de mayor rendimiento.

La respuesta esperada a este tipo de selección la podemos escribir como

$$R_f = i\,\sigma_f\,h_f^2 \tag{13.2}$$

donde σ_f representa el desvío estándar de las medias fenotípica familiares y h_f^2 es la heredabilidad de los promedios familiares, algo que calcularemos un poco más adelante.

El mayor problema con este tipo de selección es, obviamente, el incremento en la consanguinidad esperada en la población y la depresión endogámica consiguiente en las características que la muestren, algo que discutiremos en el capítulo «Endocría, exocría, consanguinidad y depresión endogámica».

Selección fraternal: Podría considerarse como un caso especial de **selección familiar** en el que el fenotipo del candidato no es incluido en las mediciones. Este tipo de selección es clásico para características que se expresan en un solo sexo, como podrían ser las de producción lechera que solo se expresan en hembras. En este caso, es frecuente utilizar la información proveniente de grupos de hermanos (enteros o medios hermanos, uno de los dos). A medida que el número de individuos medidos por grupo se incrementa, el incluir o no el fenotipo del individuo deja de tener relevancia y los resultados de ambos tipos de selección convergen. Este tipo de selección a su vez es típica en el mejoramiento genético de características de carcasa donde se sacrifican hermanos del candidato, el cual por obvios motivos no podemos medir.

Selección dentro de familia: La idea es exactamente la opuesta a la idea de **selección familiar**, en lugar de seleccionar por la media familiar, seleccionamos por el desvío respecto a la media familiar. La lógica subyacente de esto es que, si el **ambiente común familiar** juega un papel muy importante en la similaridad de los individuos, entonces las diferencias respecto a la media de la familia pueden representar mejor las diferencias genéticas entre individuos. Un ejemplo claro de características de este tipo son las de crecimiento inicial en especies multíparas de mamíferos, como los cerdos, donde toda la camada se cría junta y su «performance» depende en gran medida de las **habilidades maternas**. En el caso particular de que se seleccione un individuo de cada sexo en cada familia, en lugar de uno solo, esto permite asegurar tamaños de familia exactamente iguales, llevando a cero la varianza en el número de progenies efectivo por familia y de esta manera incrementando al doble el tamaño efectivo poblacional (algo que a fines prácticos es mayormente aplicable en poblaciones de laboratorio). De esta forma, con un número no muy elevado de individuos podemos mantener relativamente controlada la consanguinidad.

De forma análoga a la selección familiar, la respuesta esperada a la **selección dentro de familias** vamos a calcularla como

$$R_w = i\ \sigma_w\ h_w^2 \tag{13.3}$$

donde R_w es la respuesta a la **selección dentro de familias**, σ_w es el desvío estándar fenotípico de los desvíos respecto a la media familiar y h_w^2 es la heredabilidad correspondiente a este tipo de selección.

Pruebas de progenie: En este tipo de selección la idea es evaluar el valor reproductivo a partir de los fenotipos de su progenie. De hecho, la definición de valor de cría como el doble del desvío fenotípico observado en la progenie es exactamente esto. Claramente, en la definición recurríamos al concepto de apareamiento con una muestra aleatoria de tamaño infinito de individuos del otro sexo, cosa que en la práctica es imposible. Sin embargo, evaluando a los futuros reproductores a partir de un gran número de hijos, producto de apareamientos al azar, suele ser una buena aproximación a ese concepto. En principio, podría parecer que se trata de una alternativa óptima, pero presenta serias dificultades que lo hacen apenas una alternativa razonable para manejos a gran escala y esfuerzos centralizados.

Por un lado, existe un incremento importante en el intervalo generacional, ya que debemos esperar a que los hijos del candidato crezcan hasta el momento de poder ser evaluados para la característica de interés y luego esperar un período reproductivo entero del candidato elegido (en el que es usado en forma masiva) para que nazcan sus hijos. Esto enlentece en una medida importante el mejoramiento genético anual, algo que discutiremos en la sección «Límites de la eficiencia en pruebas de progenie». Por otra parte, cuantos más individuos tengamos que mantener para evaluar a cada candidato (aumentando la precisión de la evaluación del candidato), menos candidatos podremos evaluar y por lo tanto también bajaremos la intensidad de selección. Se trata de un balance delicado y no-lineal, por lo que la evaluación de las mejores combinaciones es algo complicada.

Cálculo de las heredabilidades en los distintos tipos de selección

Como $\sigma^2_T = \sigma^2_B + \sigma^2_W$ y $t = \frac{\sigma^2_B}{\sigma^2_T}$ y por lo tanto $\sigma^2_B = t\ \sigma^2_T$, entonces $\sigma^2_W = (1-t)\sigma^2_T$. Esta es la partición de la varianza observada en componentes observables, donde la varianza total σ^2_T es la varianza fenotípica y que, como de costumbre, notaremos con V_P.

De la misma manera, podemos particionar la varianza aditiva en componentes entre familias y dentro de familias, como ya discutimos en el capítulo 9. En él vimos que la covarianza dentro de familias (similaridad) era igual a la varianza entre familias. Si, para mantener la consistencia con Falconer y Mackay (1996) usamos r como notación para el coeficiente de parentesco aditivo ($r = 2f$), en lugar de a que habíamos usado, entonces la varianza aditiva entre familias será $r\ V_A$ y la varianza aditiva dentro de familias, por lo tanto, $(1-r)\ V_A$.

Organizando la información anterior en una tabla de partición de las varianzas aditivas y fenotípica y su relación con los componentes observacionales, tenemos

Componente observacional	Varianza Aditiva	Varianza Fenotípica
Entre familias σ^2_B	$r\ V_A$	$t\ V_P$
Dentro de familias σ^2_W	$(1-r)\ V_A$	$(1-t)\ V_P$

Si σ^2_f y σ^2_w son las varianzas observadas sin error, es decir con muestras de tamaño infinito, el observarlas con muestras de tamaño reducido introducirá errores en las medias de grupos (familias) y por lo tanto las varianzas observadas realmente. De esta forma, la varianza entre medias familiares se verá incrementada en la cantidad $\frac{1}{n}\sigma^2_W$ (que es la varianza en la estimación de la media familiar), mientras que la varianza de los desvíos intrafamiliares se verá reducida de la misma manera (la varianza total es la misma).

Entonces, teniendo en cuenta el tamaño de muestras no-infinito, tenemos que **la relación entre los componentes observacionales y causales**, utilizando los resultados que están en la tabla anterior, para las varianzas aditivas están dadas ahora por

$$\sigma^2_B + \frac{1}{n}\sigma^2_W = rV_A + \frac{1}{n}(1-r)V_A = \left[\frac{nr+(1-r)}{n}\right]V_A = \left[\frac{1+(n-1)r}{n}\right]V_A \quad (13.4)$$

y

$$\sigma^2_W - \frac{1}{n}\sigma^2_W = \frac{(n-1)}{n}\sigma^2_W = \frac{(n-1)}{n}(1-r)V_A = \frac{(n-1)(1-r)}{n}V_A \quad (13.5)$$

Con el mismo procedimiento, para las varianzas fenotípicas, tenemos ahora que

$$\sigma^2_B + \frac{1}{n}\sigma^2_W = tV_P + \frac{1}{n}(1-t)V_P = \left[\frac{nt+(1-t)}{n}\right]V_P = \left[\frac{1+(n-1)t}{n}\right]V_P \quad (13.6)$$

y en forma correspondiente

$$\sigma^2_W - \frac{1}{n}\sigma^2_W = \frac{(n-1)}{n}\sigma^2_W = \frac{(n-1)}{n}(1-t)V_P = \frac{(n-1)(1-t)}{n}V_P \quad (13.7)$$

La información anterior podemos organizarla en una tabla como la siguiente, que nuevamente relaciona los componentes observacionales con los causales:

Varianza observada	Componentes observacionales	Aditiva (causal)	Fenotípica (causal)
de medias familiares σ^2_f	$\sigma^2_B + \frac{1}{n}\sigma^2_W$	$\frac{1+(n-1)r}{n}\ V_A$	$\frac{1+(n-1)t}{n}\ V_P$
desvíos intrafamiliares σ^2_w	$\sigma^2_W - \frac{1}{n}\sigma^2_W$	$\frac{(n-1)(1-r)}{n}\ V_A$	$\frac{(n-1)(1-t)}{n}\ V_P$

A partir de esto, la heredabilidad de la **selección familiar** podemos calcularla directamente como la relación entre las varianzas aditiva y fenotípica correspondientes, es decir

$$h_f^2 = \frac{\left[\frac{1+(n-1)r}{n}\right] V_A}{\left[\frac{1+(n-1)t}{n}\right] V_P} = \left(\frac{1+(n-1)r}{1+(n-1)t}\right) \frac{V_A}{V_P} = \left(\frac{1+(n-1)r}{1+(n-1)t}\right) h^2 \quad (13.8)$$

De la misma manera, la heredabilidad de las desviaciones intrafamiliares será

$$h_w^2 = \frac{\frac{(n-1)(1-r)}{n} V_A}{\frac{(n-1)(1-t)}{n} V_P} = \frac{(1-r)}{(1-t)} \frac{V_A}{V_P} = \frac{(1-r)}{(1-t)} h^2 \quad (13.9)$$

En el caso de **selección fraternal**, donde nos vamos a guiar por la información de los hermanos del candidato, la principal diferencia con la selección familiar es que el individuo candidato no es medido. En este caso, resulta más sencillo obtener la heredabilidad apropiada a este tipo de selección como la regresión de los valores de cría de los individuos candidatos (no medidos) en la media fenotípica de sus hermanos medidos. La covarianza entre hermanos (medio o enteros) es igual $r\ V_A$, lo que no se ve afectado por el número de hermanos medidos en el grupo. Por otro lado, la varianza fenotípica en el grupo de medios hermanos sí se ve afectada y su valor, como vimos antes es $\frac{1+(n-1)t}{n} V_P$, por lo que poniendo todo junto, tenemos

$$h_s^2 = \frac{r\ V_A}{\frac{1+(n-1)t}{n} V_P} = \frac{nr}{1+(n-1)t} h^2 \quad (13.10)$$

Si recordamos del capítulo «Parámetros Genéticos: heredabilidad y repetibilidad» que un estimador obtenido como regresión del valor de cría del individuo a partir de su información fenotípica era el producto de la heredabilidad por el desvío fenotípico (respecto a la media de la población), entonces resulta natural extender este concepto a otros tipos de selección, como por ejemplo con estructuras de parentesco definidas. En el caso de selección individual teníamos

$$\hat{A} = b_{AC}(z_i - \bar{z}) = h^2(z_i - \bar{z}) \quad (13.11)$$

Aplicando la misma idea a selección fraternal, que se puede aplicar a estructuras de parentesco definidas y que no contengan la información del propio candidato, tenemos que el coeficiente de regresión es

$$b_{AC} = \frac{r\, n\, h^2}{1 + (n-1)t} \tag{13.12}$$

y por lo tanto, **el estimador del valor de cría del individuo**, evaluado **a través de sus parientes** será

$$\hat{A} = b_{AC}(z_i - \bar{z}) = h_s^2(z_i - \bar{z}) = \frac{n\, r\, h^2}{1 + (n-1)t}(z_i - \bar{z}) \tag{13.13}$$

Para obtener la precisión del método de selección fraternal (y de los otros que tengan relación fija de parentesco y no incluyan al candidato) alcanza con calcular la correlación entre el valor de cría del candidato y el criterio (la media fenotípica de n hermanos). Usando la misma lógica que para derivar la ecuación (13.10), pero ahora recordando que la varianza aditiva de la población de donde sale el candidato es V_A, tenemos

$$r_{AC} = \frac{r\, V_A}{\sqrt{V_A}\sqrt{\frac{1+(n-1)t}{n}V_P}} = \frac{V_A}{\sqrt{V_A}\sqrt{V_P}}\frac{r}{\sqrt{\frac{1+(n-1)t}{n}}} =$$

$$\frac{\sqrt{V_A}\sqrt{V_A}}{\sqrt{V_A}\sqrt{V_P}}\frac{r}{\sqrt{\frac{1+(n-1)t}{n}}} = \frac{\sqrt{V_A}}{\sqrt{V_P}}\frac{r}{\sqrt{\frac{1+(n-1)t}{n}}} = \sqrt{\frac{V_A}{V_P}}\frac{r}{\sqrt{\frac{1+(n-1)t}{n}}} \quad \therefore$$

$$r_{AC} = r\, h\sqrt{\frac{n}{1 + (n-1)t}} \tag{13.14}$$

Utilizando la misma aproximación de regresión, en el caso de las pruebas de progenie, teniendo en cuenta que $r = \frac{1}{2}$ y $t = r_{A_1A_2}\, h^2 = \frac{1}{4}h^2$

$$b_{AC} = r\,\frac{n\, h^2}{1 + (n-1)t} = \frac{1}{2}\left[\frac{n\, h^2}{1 + (n-1)\frac{1}{4}h^2}\right] = \frac{1}{2}\left[\frac{4\, n\, h^2}{4 + (n-1)h^2}\right] = \frac{2\, n\, h^2}{4 + (n-1)h^2} \tag{13.15}$$

que es el mismo resultado que vamos a obtener más adelante en la ecuación (13.49) de la sección «Pruebas de progenie» usando la aproximación de los índices de selección.

PARA RECORDAR

- En la **selección individual** la idea es utilizar la información fenotípica aportada por el individuo como criterio para la decisión sobre el valor reproductivo del mismo. Sin embargo, en situaciones donde el fenotipo individual es difícil o aún imposible de obtener, o cuando la heredabilidad de la característica es baja, aún podemos recurrir a la selección usando otras fuentes de información.
- Podemos **descomponer el fenotipo de un individuo en la suma de dos fuentes diferentes**, una la **correspondiente al desvío de la media familiar respecto a la media de la población**, que llamaremos P_f y otra que corresponde al **desvío fenotípico del individuo respecto a la media familiar**, que llamaremos P_w.
- Cuando realizamos **selección familiar** la idea es darle todo el peso a P_f, dejando de lado el aporte de P_w y así eligiendo familias enteras como reproductores. La lógica de este tipo de selección es que **los efectos ambientales tienden a cancelarse dentro de la familia**, lo que llevaría a que la media fenotípica de esta sea similar a la media genotípica. La respuesta esperada será $R_f = i\ \sigma_f\ h_f^2$ donde σ_f representa el desvío estándar de las medias fenotípica familiares y h_f^2 es la heredabilidad de los promedios familiares.
- El problema más notable de este tipo de selección es el incremento en la consanguinidad esperada en la población y la depresión endogámica consiguiente en las características que la muestren.
- Podemos definir la **selección fraternal** como un caso particular de **selección familiar** en el que el fenotipo del candidato no es incluido en las mediciones. En este caso, es frecuente utilizar la información proveniente de grupos de hermanos. A medida que el número de individuos medidos por grupo se incrementa, el incluir o no el fenotipo del individuo deja de tener relevancia y los resultados de ambos tipos de selección convergen.
- Cuando realizamos **selección dentro de la familia**, la idea es la opuesta a la de selección familiar, en lugar de seleccionar por la media familiar, seleccionamos por el desvío respecto a la media familiar. Este tipo de selección se aplica cuando el **ambiente común familiar juega un papel muy importante en la similaridad de los individuos**, entonces las diferencias respecto a la media de la familia pueden representar mejor las diferencias genéticas entre individuos. La respuesta esperada será $R_w = i\ \sigma_w\ h_w^2$ donde R_w es la respuesta a la **selección dentro de familias**, σ_w es el desvío estándar fenotípico de los desvíos respecto a la media familiar y h_w^2 es la heredabilidad correspondiente a este tipo de selección.
- La idea detrás de las **pruebas de progenie** es evaluar el valor reproductivo a partir de los fenotipos de su progenie. Teóricamente parece una alternativa óptima a la evaluación, pero al tener que esperar que los hijos crezcan para realizar la evaluación de la característica, el IG se ve aumentado. Por otra parte, cuánto más individuos evaluemos aumentaremos la precisión del método, pero estaremos disminuyendo la cantidad de candidatos a evaluar y por lo tanto también bajaremos la intensidad de selección.

- Podemos calcular la **heredabilidad de la selección familiar** directamente como la relación entre las varianzas aditiva y fenotípica correspondientes: $h_f^2 = \left(\frac{1+(n-1)r}{1+(n-1)t}\right) h^2$
- De la misma manera, la **heredabilidad de las desviaciones intrafamiliares** será $h_w^2 = \frac{(1-r)}{(1-t)} h^2$
- En el caso de la **selección fraternal**, la heredabilidad podemos calcularla como $h_s^2 = \frac{nr}{1+(n-1)t} h^2$
- El **estimador del valor de cría del individuo**, evaluado **a través de sus parientes** será $\hat{A} = \frac{n\ r\ h^2}{1+(n-1)t}(z_i - \bar{z})$ y su correspondiente precisión será $r_{AC} = r\ h \sqrt{\frac{n}{1+(n-1)t}}$

13.3 Combinando la información proporcionada por diferentes tipos de parientes

Hasta ahora hemos visto **selección fenotípica individual** en el capítulo «Selección artificial I», **selección por una característica correlacionada** con fenotipo en el mismo individuo en el capítulo «Correlaciones y respuesta correlacionada» y en la sección precedente «Diferentes tipos de selección basados en estructuras de parentesco definidas». En el último caso, vimos que se podía tratar como la ponderación de la información fenotípica proveniente de medias familiares y de desvíos respecto a esas medias, pero nunca hicimos un esfuerzo por buscar la combinación óptima de esas dos fuentes. En esta sección vamos a comenzar a remediar ese punto, buscando combinaciones óptimas desde el punto de vista de la respuesta a la selección. Las mismas provienen de la aplicación de la teoría de índices de selección y vamos a basar buena parte de esta sección en las notas del curso de Werf (2000).

Un índice de selección (lineal) es una combinación (lineal) de diferentes valores, usualmente fenotípicos, tal que me permite asignar un valor reproductivo único a cada individuo. En el contexto de esta sección, la idea es cómo combinar de manera óptima la información, de tal manera que la correlación del índice con el valor de cría sea máxima. Dicho de otra forma, nuestro índice será

$$I = EBV = \hat{A} = b_1\ z_1 + b_2\ z_2 + ... + b_k\ z_k \tag{13.16}$$

donde utilizamos los z_i para identificar los fenotipos de cada tipo de fuente de información. Los tipos de fuentes de información pueden incluir el fenotipo propio del candidato, el promedio de hermanos enteros, promedio de medios hermanos, fenotipo de un progenitor, etcétera. De la teoría de modelos lineales, sabemos que los coeficientes correspondientes serán iguales a

$$b_i = \frac{Cov(z_i,A)}{Var(z_i)} \tag{13.17}$$

Por lo tanto, para estimar estos coeficientes la información que vamos a precisar es

1) Covarianzas entre los valores fenotípicos y el valor de cría que nos interesa, es decir $Cov(z_i,A)$.
2) Varianzas fenotípicas en las diferentes fuentes de información, es decir $Var(z_i)$.
3) Como las fuentes de información covarían, para no superponer información (es decir, contar dos veces la misma información), debemos tener en cuenta esas covarianzas, es decir $Cov(z_i,z_j)$.

La información de los puntos 2 y 3 la podemos organizar en una matriz de varianza-covarianza, como vimos en la sección «Matrices de varianza-covarianza» del capítulo «Correlaciones y respuesta correlacionada», es decir, la matriz $\mathbf{P}$.

Como los elementos b_i forman un vector de largo k, que llamaremos $\mathbf{b}$ y las informaciones fenotípicas correspondientes a las distintas fuentes de información forman un vector $\mathbf{z}$ con elementos z_i, la ecuación (13.16) podemos escribirla como

$$EBV = \hat{A} = b_1\ z_1 + b_2\ z_2 + ... + b_k\ z_k = \mathbf{b'z} \tag{13.18}$$

Más aún, las covarianzas entre las distintas fuentes de información fenotípica y el valor de cría, es decir las que describe el punto 1, $Cov(z_i,A)$, las podemos organizar en un vector. Como en realidad se trata de varianzas y covarianzas genéticas, usando la misma notación que usamos previamente en el capítulo anterior, llamaremos a esto $\mathbf{G}$, aunque teniendo en cuenta que en este caso particular se trata de un vector de varianzas-covarianzas genéticas.

Por lo tanto, usando la notación matricial, el vector de coeficientes del índice se reduce a

$$\mathbf{b = P^{-1}G} \tag{13.19}$$

por lo que debemos construir $\mathbf{G}$ y $\mathbf{P}$ para poder calcular los pesos de las distintas fuentes de información.

Cálculo de los elementos de la matriz de varianzas-covarianzas fenotípica

Recordemos que las características fenotípicas de las que estamos hablando en este contexto son en realidad siempre las mismas, pero provenientes de diferentes fuentes de información. Tenemos, por lo tanto, diferentes situaciones a considerar:

- Información generada a partir de una sola medición, donde la varianza de la misma está dada por

$$Var(z_i) = \sigma_P^2 \tag{13.20}$$

- Información generada a partir de la media de n individuos, relacionados entre sí de tal manera que el coeficiente de correlación intraclase sea t y donde por lo tanto la varianza del promedio será

$$Var(\bar{z}_i) = t\sigma_P^2 + \frac{(1-t)}{n}\sigma_P^2 = \sigma_P^2\left[t + \frac{(1-t)}{n}\right] = \sigma_P^2\left[\frac{nt + (1-t)}{n}\right] \therefore$$
$$Var(\bar{z}_i) = \sigma_P^2\left[\frac{1 + (n-1)t}{n}\right] \tag{13.21}$$

Esto es equivalente a lo que vimos en la sección «La reducción en la varianza fenotípica con varias medidas», resultado de promediar n observaciones, algo que vimos en el capítulo «Parámetros genéticos: heredabilidad y repetibilidad». El coeficiente de correlación intraclase, habíamos visto que podía ser igual a la repetibilidad, es decir $t = \rho$, cuando teníamos varias medidas repetidas en el mismo individuo o $t = r_{A_1A_2}\ h^2$ cuando teníamos medidas en distintos individuos con el mismo grado de parentesco.

- La covarianza entre dos fuentes de información que consiste cada una en una sola medida, tomadas en individuos diferentes, es igual a

$$Cov(z_i,z_j) = a_{ij}\ \sigma_A^2 = a_{ij}\ h^2\ \sigma_P^2 \tag{13.22}$$

En los casos en que las observaciones entre individuos compartan una fracción del ambiente en común, por ejemplo, si c^2 es la proporción de varianza en común que corresponde al ambiente compartido, entonces adicionamos $c^2\ \sigma_P^2$ (ignorando los efectos de dominancia).

- La covarianza entre dos fuentes de información, una de ellas una sola medida en el candidato a la selección y la otra el promedio de n medidas en individuos que tienen el mismo parentesco entre sí y también el mismo parentesco con el candidato, es igual a

$$Cov(z_i,\bar{z}_j) = a_{ij}\ \sigma_A^2 = a_{ij}\ h^2\ \sigma_P^2 \tag{13.23}$$

con a_{ij} el parentesco aditivo entre el candidato y cualquiera de los integrantes del grupo a promediar (todos tienen el mismo parentesco con el candidato).

Cálculo de los elementos del vector de varianzas-covarianzas genéticas

En este caso, solo tenemos dos posibilidades en el vector:

- La covarianza entre **una sola medida** fenotípica y el **valor de cría** del individuo, que está dada por

$$Cov(z_i,A) = a_{ij}\ \sigma_A^2 = a_{ij}\ h^2\ \sigma_P^2 \tag{13.24}$$

- La covarianza entre el promedio de n medidas fenotípicas y el **valor de cría** del individuo, que está dada por

$$Cov(\bar{z}_i,A) = a_{ij}\ \sigma_A^2 = a_{ij}\ h^2\ \sigma_P^2 \tag{13.25}$$

Nuevamente, es importante acá que todos los individuos con información fenotípica sean con el mismo grado de parentesco entre sí y a su vez, con el mismo grado de parentesco (puede ser diferente al anterior) con el candidato.

Información propia y de un progenitor

Supongamos que queremos construir un índice utilizando la información propia y de un progenitor a fin de predecir el valor de cría de un individuo. En ese caso, usando la ecuación (13.16), nuestro índice sería

$$I = EBV_{ind} = \hat{A} = b_{ind}\ z_{ind} + b_{pro}\ z_{pro} \tag{13.26}$$

Por otra parte, de acuerdo con la ecuación (13.19), el índice será igual a

$$\mathbf{b} = \mathbf{P^{-1}G} \tag{13.27}$$

por lo que debemos determinar dichas matrices. Aplicando las «reglas» que vimos antes para los elementos de $\mathbf{P}$, tenemos que la misma quedará determinada por

$$\mathbf{P} = \begin{pmatrix} Var(z_{ind}) & Cov(z_{ind},z_{pro}) \\ Cov(z_{pro},z_{ind}) & Var(z_{pro}) \end{pmatrix} = \begin{pmatrix} \sigma_P^2 & \frac{1}{2}h^2\sigma_P^2 \\ \frac{1}{2}h^2\sigma_P^2 & \sigma_P^2 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2}h^2 \\ \frac{1}{2}h^2 & 1 \end{pmatrix} \sigma_P^2 \tag{13.28}$$

Por otra parte, aplicando las reglas para el vector $\mathbf{G}$, dado que $a_{ind,ind} = 1$, $a_{ind,pro} = \frac{1}{2}$, tenemos que el mismo es igual a

$$\mathbf{G} = \begin{pmatrix} Cov(z_{ind},A) \\ Cov(z_{pro},A) \end{pmatrix} = \begin{pmatrix} a_{ind,ind}\ \sigma_A^2 \\ a_{ind,pro}\ \sigma_A^2 \end{pmatrix} = \begin{pmatrix} h^2\ \sigma_P^2 \\ \frac{1}{2}\ h^2\ \sigma_P^2 \end{pmatrix} = \begin{pmatrix} h^2 \\ \frac{1}{2}\ h^2 \end{pmatrix} \sigma_P^2 \tag{13.29}$$

Utilizando los resultados de las ecuaciones (13.28) y (13.29), tenemos que el índice estará dado por

$$\mathbf{b} = \mathbf{P^{-1}G} = \left[\begin{pmatrix} 1 & \frac{1}{2}h^2 \\ \frac{1}{2}h^2 & 1 \end{pmatrix} \sigma_P^2 \right]^{-1} \begin{pmatrix} h^2 \\ \frac{1}{2}\ h^2 \end{pmatrix} \sigma_P^2 = \begin{pmatrix} 1 & \frac{1}{2}h^2 \\ \frac{1}{2}h^2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} h^2 \\ \frac{1}{2}\ h^2 \end{pmatrix} \tag{13.30}$$

Más aún, la inversión de matrices de **2×2** es directa, ya que si la misma tiene coeficientes a,b,c,d, su inversa es

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{13.31}$$

por lo que aplicando esto a la ecuación (13.30) tenemos que

$$\mathbf{b} = \begin{bmatrix} 1 & \frac{1}{2}h^2 \\ \frac{1}{2}h^2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} h^2 \\ \frac{1}{2}\ h^2 \end{bmatrix} = \left(\frac{1}{1-\frac{1}{4}h^4} \right) \begin{bmatrix} 1 & -\frac{1}{2}h^2 \\ -\frac{1}{2}h^2 & 1 \end{bmatrix} \begin{bmatrix} h^2 \\ \frac{1}{2}\ h^2 \end{bmatrix} \therefore$$
$$\mathbf{b} = \left(\frac{4}{4-h^4} \right) \begin{bmatrix} h^2 - \frac{1}{4}h^4 \\ \frac{1}{2}h^2 - \frac{1}{2}h^4 \end{bmatrix} \tag{13.32}$$

Es decir, tenemos ahora los coeficientes del índice, para la información que aporta el individuo y para la información que aporta su progenitor (uno de ellos), que sería

$$EBV = \hat{A} = \mathbf{b'z} = \left(\frac{4}{4-h^4} \right) [(h^2 - \frac{1}{4}h^4)\ z_{ind} + (\frac{1}{2}h^2 - \frac{1}{2}h^4)\ z_{pro}] \tag{13.33}$$

En la figura 13.1 podemos apreciar el comportamiento de los dos coeficientes, el de color rojo para el coeficiente correspondiente a la información aportada por el individuo, el de color azul correspondiente a la información que aporta su progenitor. De la ecuación (13.33) resulta claro que tanto los coeficientes del índice, como sus pesos relativos, solo dependen del valor de la heredabilidad. A heredabilidades muy altas la predicción del índice es casi idéntica a la predicción a partir del fenotipo individual, mientras que a heredabilidades más bajas el papel de la información aportada por el progenitor puede ser mucho más relevante.

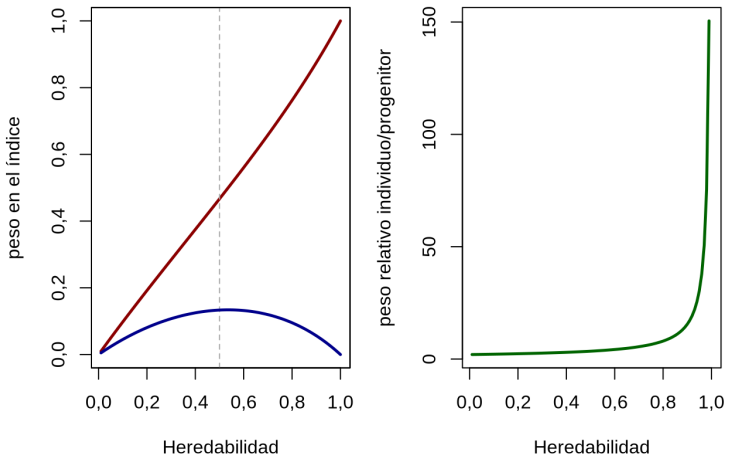


Figura 13.1. Pesos en el índice con información del individuo y un progenitor en función de la heredabilidad de la característica. A la izquierda, en rojo el coeficiente correspondiente al individuo, mientras que en azul el correspondiente al peso de la información en el progenitor. A la derecha (en verde), la relación entre los dos coeficientes, es decir b_{ind}/b_{pro}, donde se aprecia claramente que a mayor heredabilidad mayor es el peso relativo de la información individual respecto a la del progenitor. De hecho, si se observa el gráfico de la izquierda, el máximo del coeficiente para la información aportada por el progenitor está en $h^2 = \frac{1}{2}$, ya que la derivada de $\frac{1}{2}h^2 - \frac{1}{2}h^4$ respecto a la heredabilidad se anula en ese valor y la derivada segunda tiene signo negativo, aunque la relación entre el coeficiente del individuo y del progenitor es una función monotónica creciente.

Ejemplo 13.2

En el rebaño experimental de ovejas verdes se están buscando los mejores carneros para mejorar el índice de esfericidad. Para ello se cuenta con la información de algunos borregos candidatos y la misma información de sus padres, cuando tenían la misma edad, de acuerdo con la siguiente tabla

Caravana Borrego	I_e	Caravana Padre	I_e
J117	-17	H034	-11
J432	11	F404	4
J521	-7	G009	1
J707	4	E009	13

Si la heredabilidad de la característica Índice de Esfericidad I_e es de $h^2 = 0{,}50$, estimar un índice para los valores de cría esperados de estos borregos y ordenarlos de acuerdo con dicho índice.

De acuerdo con la ecuación (13.33), el índice correspondiente a esta característica sería

$$EBV = \hat{A} = \mathbf{b}'\mathbf{z} = \left(\frac{4}{4-h^4}\right)[(h^2 - \frac{1}{4}h^4)\, z_{ind} + (\frac{1}{2}h^2 - \frac{1}{2}h^4)\, z_{pro}] =$$

$$= \left(\frac{4}{4-0{,}50^2}\right)[(0{,}50 - \frac{1}{4}0{,}50^2)\, z_{ind} + (\frac{1}{2}0{,}50 - \frac{1}{2}0{,}50^2)\, z_{pro}] =$$

$$= \frac{4}{3{,}75}[0{,}4375\, z_{ind} + 0{,}125\, z_{pro}] \;\therefore$$

$$EBV = \hat{A} = 0{,}4667\, z_{ind} + 0{,}1333\, z_{pro}$$

Utilizando este índice, podemos calcularlo directamente para cada animal, o hacerlo para cada fuente de información por separado y luego sumar ambos valores de cada animal, como en la siguiente tabla.

Caravana Borrego	I_e	I_e Padre	$b_{ind}\, z_{ind}$	$b_{pro}\, z_{pro}$	EBV
J117	-17	-11	$-7{,}93$	$-1{,}47$	$-9{,}4$
J432	11	4	5,13	0,53	5,7
J521	-7	1	$-3{,}26$	0,13	$-3{,}1$
J707	4	13	1,87	1,73	3,6

De acuerdo con la tabla, recordando que queremos disminuir la esfericidad del rebaño, el orden de los borregos de acuerdo con su valor de cría esperado (EBV) es **J117**, **J521**, **J707** y **J432**.

PARA RECORDAR

- Un **índice de selección** es una combinación de diferentes valores, usualmente fenotípicos, tal que permiten asignar un valor reproductivo único a cada individuo, podemos construirlo tal que $I = EBV = \hat{A} = b_1\ z_1 + b_2\ z_2 + ... + b_k\ z_k$ donde utilizamos z_i para identificar los fenotipos de cada tipo de fuente de información que pueden incluir el fenotipo propio del candidato, el promedio de hermanos enteros, promedio de medios hermanos, fenotipo de un progenitor, etcétera.
- Podemos describir el vector de coeficientes del índice usando la notación matricial como $b = P^{-1}G$
- Calcularemos los elementos de la matriz P dependiendo de la información de la que dispongamos, tal que 1)Si disponemos de información generada a partir de **una sola medición, la varianza** de la misma está dada por $Var(z_i) = \sigma^2_P$

2) Si disponemos de **información generada a partir de la media de n individuos**, relacionados entre sí de tal manera que el coeficiente de correlación intraclase sea t, la **varianza del promedio** será $Var(\bar{z}_i) = \sigma^2_P \left[\frac{1+(n-1)t}{n}\right]$
3) Además, la **covarianza entre dos fuentes de información que consiste cada una en una sola medida**, tomadas en individuos diferentes, es igual a $Cov(z_i,z_j) = a_{ij}\ h^2\ \sigma^2_P$
4) Consiguientemente, la **covarianza entre dos fuentes de información, una de ellas una sola medida en el candidato a la selección y la otra el promedio de n medidas en individuos que tienen el mismo parentesco entre sí y también el mismo parentesco con el candidato**, es igual a $Cov(z_i,\bar{z}_j) = a_{ij}\ h^2\ \sigma^2_P$ siendo a_{ij} el parentesco aditivo entre el candidato y cualquiera de los integrantes del grupo a promediar (todos tienen el mismo parentesco con el candidato).

- De la misma manera, calcularemos los elementos de la matriz G dependiendo de la información de la que dispongamos, tal que

1) La covarianza entre **una sola medida** fenotípica y el **valor de cría** del individuo, está dada por $Cov(z_i,A) = a_{ij}\ h^2\ \sigma^2_P$
2) La **covarianza entre el promedio de n medidas fenotípicas** y el **valor de cría** del individuo, está dada por $Cov(\bar{z}_i,A) = a_{ij}\ h^2\ \sigma^2_P$

- En el caso que quisiéramos construir un índice utilizando la información propia y de un progenitor a fin de predecir el valor de cría de un individuo. El mismo sería $I = EBV_{ind} = \hat{A} = b_{ind}\ z_{ind} + b_{pro}\ z_{pro}$ donde podemos expresar los coeficientes como $EBV = \hat{A} = \mathbf{b'z} = \left(\frac{4}{4-h^4}\right)[(h^2 - \frac{1}{4}h^4)\ z_{ind} + (\frac{1}{2}h^2 - \frac{1}{2}h^4)\ z_{pro}]$
- Podemos notar que a **heredabilidades muy altas** la **predicción del índice es casi idéntica a la predicción a partir del fenotipo individual**, mientras que a heredabilidades más bajas el papel de la información aportada por el progenitor puede ser mucho más relevante.

Información propia y de varios hermanos enteros

En este caso, la información de la que disponemos es, además del registro fenotípico en el individuo, la media de n hermanos enteros del candidato (y por lo tanto, hermanos enteros entre sí, ya que los padres de todos deben ser los mismos que los del candidato). Como antes, vamos a armar primero las matrices **P** y (el vector) **G** para luego proceder a obtener los coeficientes.

Para obtener la matriz **P** alcanza con utilizar las reglas, recordando que ahora tenemos, la varianza de la información individual, igual a

$$Var(z_{ind}) = \sigma_P^2 \tag{13.34}$$

mientras que la varianza del promedio de n medidas es

$$Var(z_{HE}) = t\,\sigma_P^2 + \frac{(1-t)}{n}\sigma_P^2 = \left[\frac{1+(n-1)t}{n}\right]\sigma_P^2 \tag{13.35}$$

con $t = a_{ind,HE}\; h^2 + c^2 = \frac{1}{2}\; h^2 + c^2$ Por otra parte, la covarianza entre ambas medidas (la individual y el promedio de n hermanos enteros) será

$$Cov(z_{ind},z_{HE}) = t\sigma_P^2 \tag{13.36}$$

Poniendo todo junto, la matriz **P** nos queda

$$\begin{aligned}\mathbf{P} &= \begin{bmatrix} Var(z_{ind}) & Cov(z_{ind},z_{HE}) \\ Cov(z_{HE},z_{ind}) & Var(z_{HE}) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_P^2 & t\,\sigma_P^2 \\ t\,\sigma_P^2 & t\,\sigma_P^2 + \frac{(1-t)}{n}\sigma_P^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & t \\ t & \frac{1+(n-1)t}{n} \end{bmatrix}\sigma_P^2 \end{aligned} \tag{13.37}$$

Por otro lado, los elementos del vector **G** son

$$Cov(z_{ind},A) = a_{ind,ind}\;\sigma_A^2 = \sigma_A^2 \tag{13.38}$$

y

$$Cov(z_{HE},A) = a_{HE,ind}\;\sigma_A^2 = \frac{1}{2}\sigma_A^2 \tag{13.39}$$

por lo que

$$\mathbf{G} = \begin{bmatrix} Cov(z_{ind},A) \\ Cov(z_{HE},A) \end{bmatrix} = \begin{bmatrix} \sigma_A^2 \\ \frac{1}{2}\sigma_A^2 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}\sigma_A^2 \tag{13.40}$$

Ahora, poniendo las dos matrices juntas para armar el índice, recordando que tenemos que invertir $\mathbf{P}$, tenemos que

$$\begin{aligned}\mathbf{b} = \mathbf{P^{-1}G} = \left(\begin{bmatrix}1 & t \\ t & \frac{1+(n-1)t}{n}\end{bmatrix}\sigma_P^2\right)^{-1}\begin{bmatrix}1 \\ \frac{1}{2}\end{bmatrix}\sigma_A^2 = \\ = \begin{bmatrix}1 & t \\ t & \frac{1+(n-1)t}{n}\end{bmatrix}^{-1}\begin{bmatrix}1 \\ \frac{1}{2}\end{bmatrix}\frac{\sigma_A^2}{\sigma_P^2} = \begin{bmatrix}1 & t \\ t & \frac{1+(n-1)t}{n}\end{bmatrix}^{-1}\begin{bmatrix}1 \\ \frac{1}{2}\end{bmatrix}h^2\end{aligned} \tag{13.41}$$

Haciendo la inversión de la matriz $\mathbf{P}$ escalada por $1/\sigma_P^2$ (que sacamos de factor común) y multiplicando por $\mathbf{G}$, tenemos entonces que

$$\mathbf{b} = \begin{bmatrix}1 & t \\ t & \frac{1+(n-1)t}{n}\end{bmatrix}^{-1}\begin{bmatrix}1 \\ \frac{1}{2}\end{bmatrix}h^2 = \left(\frac{n}{1+(n-1)t-nt^2}\right)\begin{bmatrix}\frac{1+(n-1)t}{n} & -t \\ -t & 1\end{bmatrix}\begin{bmatrix}1 \\ \frac{1}{2}\end{bmatrix}h^2 \therefore$$
$$\mathbf{b} = \left(\frac{nh^2}{1+(n-1)t-nt^2}\right)\begin{bmatrix}\frac{2+(n-2)t}{2n} \\ \frac{1}{2}-t\end{bmatrix} \tag{13.42}$$

Tenemos por lo tanto el índice, como aparece en la ecuación (13.42), para el caso en que contamos con la información propia del individuo y del promedio de n hermanos enteros del mismo. Ahora, a diferencia de lo que ocurría en el caso de individuo y progenitor (que solo dependía de h^2), ahora nuestro índice depende de varios parámetros: h^2, n y t (que a su vez depende de h^2 y c^2). Por lo tanto, resulta algo complejo el estudio analítico de esta función y vamos solamente a realizar una exploración gráfica del comportamiento de los coeficientes en distintas situaciones.

Como se ilustra en la figura 13.2, el aporte relativo y absoluto de los coeficientes de aporte individual y del promedio de n hermanos enteros varía ampliamente de acuerdo con la heredabilidad, la proporción de varianza debida al ambiente común y el número de hermanos evaluados. Las gráficas de la parte superior representan la situación donde la proporción de la varianza que corresponde al ambiente común es cero, es decir, los hermanos enteros no comparten ambiente especial (materno, etcétera). En este caso, ambos coeficientes son positivos siempre; mientras que el correspondiente al fenotipo individual del candidato crece en forma monotónica, el aporte del promedio de los n hermanos enteros posee máximos a frecuencias intermedias y que dependen de n. Por ejemplo, se observa claramente que a mayor n, el máximo se desplaza hacia la izquierda, es decir, se da con menores heredabilidades.

Por otra parte, cuando la proporción de la varianza debida al ambiente común es mayor a cero, es decir $c^2 > 0$, el signo del coeficiente correspondiente al aporte del promedio de los hermanos aumenta hasta cierta heredabilidad y luego decrece, pasando a ser negativo incluso. Este efecto se debe a que si una mayor parte de la similaridad entre hermanos se debe al ambiente, entonces una menor proporción será genética. De hecho, cuando el signo es negativo la covarianza entre hermanos enteros sirve como una corrección por el ambiente común.

En este punto resulta claro que si usamos los índices de selección para incorporar la información de dos fuentes, una individual y la otra familiar, estamos dándole mayor o menor

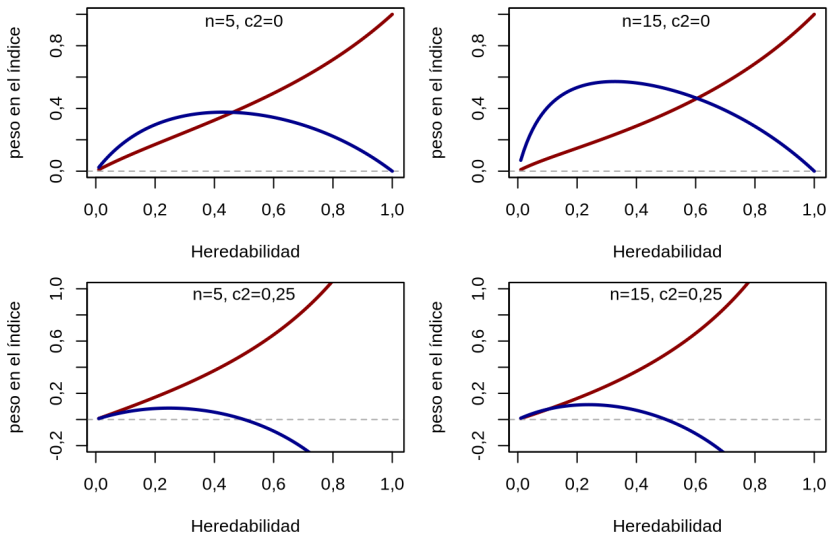


Figura 13.2. Comportamiento de los coeficientes para el aporte del fenotipo individual (en azul) y del promedio de n hermanos enteros del candidato, de acuerdo con la heredabilidad h^2 y la proporción de varianza del ambiente común c^2. A medida que crece la heredabilidad, crece el coeficiente del aporte individual, mientras que el aporte de los hermanos tiene un máximo a heredabilidades intermedias. A medida que crece el número de hermanos evaluados (gráficos de la izquierda versus los correspondientes a la derecha) aumenta el peso relativo del aporte de la media de hermanos enteros. Mientras que cuando $c^2 = 0$ ambos coeficientes son siempre positivos, cuando $c^2 > 0$ a partir de determinados valores de h^2 el aporte del promedio de los hermanos enteros comienza a ser negativo (gráficos de la parte inferior).

importancia a la información individual o a la familiar. En el caso de **heredabilidades altas** le daremos mayor importancia a la **información individual** (el peso relativo del coeficiente correspondiente es mayor), mientras que, si la **heredabilidad es baja**, tenderemos a favorecer la **información familiar.** En este último caso, además, al seleccionar basados en información familiar (que es casi idéntica para todos los hermanos que componen el grupo) estaremos tendiendo a **seleccionar como reproductores individuos fuertemente emparentados**, lo que puede llevar a un importante **aumento de la consanguinidad promedio** de la población y de la consiguiente **depresión endogámica** (que veremos en el capítulo «Endocría, exocría, consanguinidad y depresión endogámica»).

PARA RECORDAR

- Cuando disponemos además del registro fenotípico en el individuo, la media de n hermanos enteros del candidato, las matrices P y G serán

$$\mathbf{P} = \begin{bmatrix} 1 & t \\ t & \frac{1+(n-1)t}{n} \end{bmatrix} \sigma_P^2 \tag{13.43}$$

y

$$\mathbf{G} = \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} \sigma_A^2 \tag{13.44}$$

- Sabiendo además que $\mathbf{b} = \mathbf{P^{-1}G}$, entonces

$$\mathbf{b} = \left(\frac{nh^2}{1+(n-1)t-nt^2} \right) \begin{bmatrix} \frac{2+(n-2)t}{2n} \\ \frac{1}{2} - t \end{bmatrix} \tag{13.45}$$

- Como **diferencia respecto a lo que ocurría en el caso de individuo y progenitor** (que solo dependía de h^2), ahora nuestro índice depende de varios parámetros: h^2, n y t (que a su vez depende de h^2 y c^2).
- Podemos concluir que si usamos los índices de selección para incorporar la información de dos fuentes, una individual y la otra familiar, estamos dándole mayor o menor importancia a la información individual o a la familiar. En el caso de **heredabilidades altas** le daremos mayor importancia a la **información individual**, mientras que, si la **heredabilidad es baja**, tenderemos a favorecer la **información familiar.** En este último caso, además, al seleccionar basados en información familiar estaremos tendiendo a **seleccionar como reproductores individuos fuertemente emparentados**, lo que puede llevar a un importante **aumento de la consanguinidad promedio** de la población y de la consiguiente **depresión endogámica.**

Pruebas de progenie

Muchas veces, la mejor alternativa para evaluar a un futuro reproductor es a través de su progenie. Un caso típico de esto es cuando la característica no se expresa en el sexo del reproductor, como por ejemplo en la producción lechera. Más importante aún, como veremos más adelante, las pruebas de progenie son el único mecanismo que permite alcanzar una precisión cercana a uno, al menos cuando el número de progenies a evaluar no es limitante. En las pruebas de progenie se aparea cada uno de los candidatos a evaluar (sin pérdida de generalidad, supongamos que son machos) con un número determinado de hembras, escogidas al azar de la población. Esto es un punto fundamental en las derivaciones, por lo que supongamos no tenemos problema en resolverlo.

La idea es sencilla: para obtener estimados de los futuros valores de cría de los candidatos vamos a regresar el valor de cría del progenitor en el promedio de los valores fenotípicos de la progenie, es decir, nuestro índice será

$$EBV_{sire} = b_1 \, P_1 \tag{13.46}$$

con P_1 igual al promedio de la progenie expresado como desvío de la media de la población y b_1 el coeficiente de regresión del valor de cría del progenitor en la media del fenotipo de su progenie.

Como vimos en la sección «Semejanza entre parientes» del capítulo «Parentesco y semejanza entre parientes», como un padre con sus hijos comparte solo la mitad del genoma (siempre, es algo «determinístico», del mecanismo de herencia, despreciando la contribución de los genomas de organelos, etcétera), la covarianza entre el valor aditivo del progenitor y el promedio de la progenie será igual a la covarianza aditiva entre padre y la progenie, que es

$$Cov(A,P_1) = Cov(A,\frac{1}{2}A) = \frac{1}{2}V_A \tag{13.47}$$

Por otra parte, ya que en el promedio de n medidas realizadas en la progenie va a aparecer n el efecto aditivo $\frac{1}{2}A$, el aporte a la varianza va a ser $Var(\frac{1}{2}A) = \frac{1}{4}V_A$. El resto de la varianza fenotípica, es decir $(V_P - \frac{1}{4}V_A)$ va a estar promediado entre n observaciones, por lo que sumando ambos aportes (que son ortogonales), tenemos que **la varianza en la media de la progenie** será igual a

$$Var(P_1) = \frac{1}{4}V_A + \frac{(V_P - \frac{1}{4}V_A)}{n} \tag{13.48}$$

Tenemos entonces lo que precisamos para derivar el coeficiente de regresión, que será

$$b_1 = \frac{Cov(A,P_1)}{Var(P_1)} = \frac{\frac{1}{2}V_A}{\frac{1}{4}V_A + \frac{(V_P - \frac{1}{4}V_A)}{n}} = \frac{\frac{1}{2}V_A \, n}{\frac{1}{4}V_A \, n + V_P - \frac{1}{4}V_A} =$$

$$\frac{\frac{1}{2}V_A \, n}{\frac{1}{4}V_A \, (n-1) + V_P} = \frac{\frac{1}{2}V_A \, n}{\frac{1}{4}V_P \, h^2 \, (n-1) + V_P} =$$

$$\frac{V_A}{V_P} \frac{\frac{1}{2} \, n}{\frac{1}{4} \, h^2 \, (n-1) + 1} = h^2 \frac{1}{2} \frac{4n}{h^2(n-1) + 4} \; \therefore$$

$$b_1 = \frac{2\,n\,h^2}{h^2(n-1)+4} \tag{13.49}$$

13.3.1 Límites de la eficiencia en pruebas de progenie

Las pruebas de progenie tienen un uso muy importante cuando se evalúan reproductores que van a ser utilizados en forma masiva. De hecho, es el único método que permite alcanzar precisiones cercanas a 1, cuando el número de hijos a evaluar no es una limitante (es decir, es muy grande). Esto puede ser un factor importante para la venta de reproductores o en algunos casos una de las pocas posibilidades si la característica se expresa en hembras, por ejemplo. Sin embargo, el uso en establecimientos que generan sus propios reproductores puede presentar importantes limitaciones. Una de ellas es la proveniente del incremento en el intervalo generacional asociado a tener que esperar a que la progenie pueda ser evaluada antes de usar el futuro reproductor en forma masiva. Este intervalo suele ser de más de 2 y usualmente más de 3 años. Por otro lado, la intensidad de selección suele verse disminuida ya que necesariamente se ensayan pocos animales. Por lo tanto, la pregunta que nos hacemos es **¿cuándo es mejor la prueba de progenie que la selección fenotípica individual?**

En principio, igual que para respuesta correlacionada en general, podemos evaluar la eficiencia de un método respecto al otro evaluando el cociente entre ambos, que suele eliminar algunos términos de la comparación (por ejemplo, la intensidad de selección o el desvío aditivo). Un caso particular es la comparación entre un método de selección directa y uno basado en información de parientes, por ejemplo, la prueba de progenie. Sin pérdida de generalidad, pongamos que la característica por la que nos interesa seleccionar es una que se mide antes del primer año de vida del animal, por ejemplo, peso al destete. En ese caso, el intervalo generacional para selección fenotípica individual es de 3 años, mientras en el caso de la prueba de progenie es de 5 años. Para que la prueba de progenie sea más fenotípica individual, el cociente deber ser mayor que 1. Es decir,

$$\frac{r_{AA'}h\sqrt{\frac{n}{1+(n-1)r_{A_1A_2}h^2}}}{h}\frac{IG_1}{IG_m} \geqslant 1 \tag{13.50}$$

Simplificando y elevando al cuadrado ambos lados (numerador y denominador de la izquierda son necesariamente positivos), tenemos

$$\begin{aligned} r^2_{AA'}\frac{n}{1+(n-1)r_{A_1A_2}h^2}\left(\frac{IG_1}{IG_m}\right)^2 &\geqslant 1 \\ r^2_{AA'}\,n\,\left(\frac{IG_1}{IG_m}\right)^2 &\geq 1+(n-1)r_{A_1A_2}h^2 \\ \frac{r^2_{AA'}\,n\,\left(\frac{IG_1}{IG_m}\right)^2-1}{(n-1)r_{A_1A_2}} &\geqslant h^2 \end{aligned} \tag{13.51}$$

Pero en nuestro caso $r^2_{AA'} = 1/4$ y $r_{A_1A_2} = 1/4$, por lo que

$$\frac{\frac{1}{4}\, n \left(\frac{IG_1}{IG_m}\right)^2 - 1}{\frac{1}{4}(n-1)} = \frac{n \left(\frac{IG_1}{IG_m}\right)^2 - 4}{(n-1)} \geqslant h^2 \qquad (13.52)$$

Cuando n tiende a infinito, o sea que el número de hijos por candidato no es limitante,

$$\lim_{n \to \text{inf}} \frac{n \left(\frac{IG_1}{IG_m}\right)^2 - 4}{(n-1)} = \left(\frac{IG_1}{IG_m}\right)^2 \geq h^2 \qquad (13.53)$$

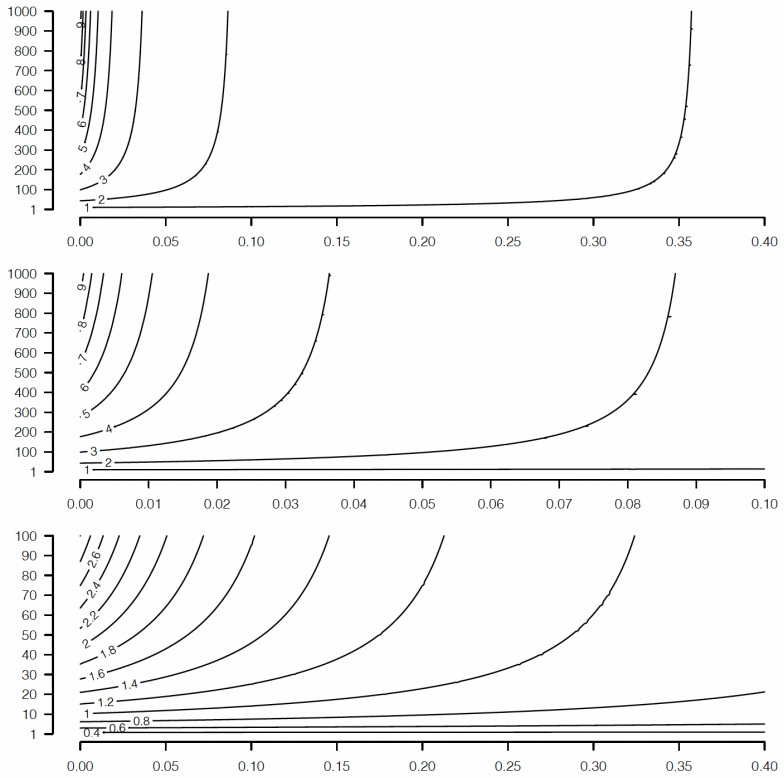


Figura 13.3. Eficiencia relativa de la prueba de progenie sobre la selección fenotípica individual. Las curvas de nivel mayores a 1 indican mayor eficiencia de la prueba de progenie. La heredabilidad se muestra en las abscisas y el número de hijos por candidato en las ordenadas.

Por lo tanto, si bien la precisión de la prueba de progenie se incrementa con el número de hijos, la misma nunca superará en eficiencia a la selección fenotípica individual cuando la heredabilidad sea superior a $\left(\frac{IG_1}{IG_m}\right)^2$. Con los intervalos generacionales de 3 y 5 años, esto significa que la prueba de progenie solo podrá ser más eficiente para características en que $h^2 \leqslant \left(\frac{IG_1}{IG_m}^2\right) = \left(\frac{3}{5}\right)^2 = 0{,}36$.

Otra forma de ver esto es a partir de un gráfico de curvas de nivel de la eficiencia relativa, de acuerdo con las diferentes heredabilidades y número de hijos por candidato. En la figura 13.3 podemos ver qué ocurre a diferentes escalas del problema. El panel superior muestra el comportamiento de la eficiencia relativa en toda la escala de heredabilidad donde puede ser igual o mayor a 1, donde se aprecia claramente como la eficiencia aumenta con el número de hijos (para eso puedes imaginar una línea vertical que a medida que subes va cortando curvas de nivel cada vez más altas) y disminuye con h^2 (lo mismo, ahora con una línea horizontal). En el panel del medio redujimos el límite superior de la heredabilidad para poder apreciar claramente lo que ocurre a heredabilidades muy bajas, donde la eficiencia relativa puede ser muy importante. Finalmente, en el panel de abajo mostramos lo que ocurre con un número de hijos por candidatos bastante más acordes a la realidad, donde se aprecia que las eficiencias relativas razonables están entre 1 y 3.

Precisión del valor de cría a partir del índice

La precisión de los valores de cría esperados a partir de la información en el índice la podemos estimar como la correlación entre los valores del índice y el valor de cría, es decir

$$r_{IA} = \frac{Cov(I,A)}{\sqrt{\sigma_I^2 \sigma_A^2}} \tag{13.54}$$

La covarianza entre el índice y el valor de cría, el numerador de la ecuación 13.54, usando notación matricial es igual a

$$Cov(I,A) = Cov(\mathbf{b}'\mathbf{z},A) = \mathbf{b}'Cov(\mathbf{z},A) = \mathbf{b}'\mathbf{G} \tag{13.55}$$

Por otra parte, la varianza en el índice (recordar que el índice es un escalar), usando nuevamente la notación matricial, la obtenemos como

$$\sigma_I^2 = Var(\mathbf{b}'\mathbf{z}) = \mathbf{b}'Var(\mathbf{z})\mathbf{b} = \mathbf{b}'\mathbf{P}\mathbf{b} \tag{13.56}$$

Observar que $\mathbf{b}'\mathbf{P}\mathbf{b}$ es un escalar (un número, no una matriz o un vector). De acuerdo con la ecuación 13.19

$$\mathbf{b} = \mathbf{P}^{-1}\mathbf{G} \therefore \mathbf{P}\mathbf{b} = \mathbf{P}\mathbf{P}^{-1}\mathbf{G} = \mathbf{I}\mathbf{G} = \mathbf{G} \tag{13.57}$$

y por lo tanto

$$\mathbf{b}'\mathbf{G} = \mathbf{b}'\mathbf{P}\mathbf{b} \tag{13.58}$$

lo que es idéntico al resultado obtenido en la ecuación 13.56. Por lo tanto, combinando las ecuaciones 13.55, 13.56 y 13.58, tenemos que

$$Cov(A,I) = \sigma_I^2 \tag{13.59}$$

Si ahora volvemos al punto inicial, la precisión en el valor de cría obtenido a partir del índice, de acuerdo con la ecuación (13.54) será

$$r_{IA} = \frac{Cov(I,A)}{\sqrt{\sigma_I^2 \sigma_A^2}} = \frac{\sigma_I^2}{\sqrt{\sigma_I^2 \sigma_A^2}} = \frac{\sigma_I^2}{\sigma_I \sigma_A} = \frac{\sigma_I}{\sigma_A} \tag{13.60}$$

PARA RECORDAR

- Muchas veces, la mejor alternativa para evaluar a un futuro reproductor es a través de su progenie. El principio básico del método es que, para **obtener los futuros valores de cría** de los candidatos, se **regresa el valor de cría del progenitor en el promedio de los valores fenotípicos de la progenie**, es decir, el índice será $EBV_{sire} = b_1\ P_1$ con P_1 igual al promedio de la progenie expresado como desvío de la media de la población y b_1 el coeficiente de regresión del valor de cría del progenitor en la media del fenotipo de su progenie.
- El coeficiente de la regresión será $b_1 = \frac{2\ n\ h^2}{h^2(n-1)+4}$
- Para que la prueba de progenie sea más eficiente que la relección individual: $\frac{r_{AA'}h\sqrt{\frac{n}{1+(n-1)r_{A_1A_2}h^2}}}{h}\frac{IG_1}{IG_m} \geqslant 1$
- Si bien la precisión de la prueba de progenie se incrementa con el número de hijos, la misma nunca superará en eficiencia a la selección fenotípica individual cuando la heredabilidad sea superior a $\left(\frac{IG_1}{IG_m}\right)^2$.
- La **precisión en el valor de cría obtenido a partir del índice** será $r_{IA} = \frac{\sigma_I}{\sigma_A}$

13.4 Métodos de selección para varias características

Hasta ahora nos hemos manejado en el presente capítulo con la selección de una característica sola, aportando información de otras fuentes que son los parientes. Sin embargo, en la práctica del mejoramiento genético es difícil encontrar situaciones donde exista un solo objetivo a mejorar desde el punto de vista genético. Por ejemplo, en producción de lana siempre queremos aumentar el peso de vellón por animal, ya que es lo que producimos, pero al mismo tiempo el diámetro de las fibras (el menor es el mejor pago) es crucial para el precio y apenas podremos comercializarlas a partir de determinado punto. En la producción lechera, vamos a procurar aumentar la producción de sólidos (proteína y grasa) ya que el pago está directamente asociado a los volúmenes de los mismos (de hecho, proteína y grasa son dos características con pago diferente), pero al mismo tiempo vamos a querer reducir el volumen de leche producida (porque es castigado por la industria, que debe concentrar) y reducir el recuento de células somáticas en leche (que es un indicador de la calidad sanitaria de la misma). En producción de carne vamos a querer, como siempre, aumentar la producción de kilos por animal, pero también mejorar características de palatabilidad de la misma. En todos estos casos se trata de características que tienen una fuerte base genética y por lo tanto son buenos objetivos para plantearnos la mejora de los mismos por vía genética.

La selección por más de una característica presenta complejidades importantes, como veremos en el resto del capítulo, tanto desde el punto de vista conceptual como desde el punto de vista práctico. Afortunadamente, para las razas que poseen una buena organización y buenos registros, los métodos «modernos» de evaluación genética nos permiten realizar selección por varias características (las características evaluadas) de manera muy sencilla a partir de los catálogos.

En lo que resta de esta sección veremos tres métodos alternativos de selección, **selección en tándem**, **selección por niveles independientes de rechazo** e **índices de selección**, mientras que dejaremos para las secciones finales el desarrollo de las bases teóricas de los índices de

selección aplicados a muchas características y a una introducción del **BLUP** y los métodos modernos de evaluación.

Selección en tándem

El primer método que vamos a presentar es el más sencillo de implementar en la práctica, la **selección en tándem** y de hecho no presenta ninguna diferencia con las estrategias de selección que usamos para una sola característica. En este método, la idea es muy sencilla: seleccionar cada vez por una sola característica, como haríamos usualmente, y luego de un tiempo cambiar a otra característica de interés. Si bien el método se representa usualmente con los métodos de selección a partir del fenotipo del individuo, en principio no hay restricciones mayores a aplicarlo con fuentes de información diferentes, como las provenientes de parientes.

La lógica de este método se presenta en la figura 13.4, donde los ejes coordenados representan dos características de interés. En la imagen de la izquierda se representa la población total (círculo azul), que sin pérdida de generalidad podría tener una distribución del tipo normal bivariada y sobre la que vamos a hacer selección por umbral en la característica 1 primero. Por lo tanto, definimos un valor de la misma a partir del cual los animales que lo superen serán los reproductores (área roja). Otra posibilidad análoga es seleccionar determinada proporción de animales, de acuerdo con los requisitos para generar los reemplazos necesarios y ambas opciones son equivalentes en caso de normalidad de la característica. Luego de algún número de generaciones o de años, después de haber observado el desplazamiento de la media de la característica lo suficiente como para «descuidarla» por un tiempo, pasamos a mejorar la segunda característica. Decimos «descuidar» porque no le vamos a prestar atención durante el siguiente período, donde el foco estará en la segunda característica, como se muestra en la imagen de la derecha en la figura 13.4. Ahora le prestaremos atención a la característica 2, donde realizaremos selección nuevamente por el método de elección (umbral, por ejemplo). Obviamente, todo el ciclo se puede repetir el número de veces que sea necesario si los objetivos de selección continúan siendo los mismos, o se pueden agregar más características en el ciclo.

La complejidad de este método es muy baja ya que no requerimos nada adicional a lo que ya acostumbramos a trabajar en selección por una sola característica, alcanza con definir los momentos en que realizaremos los cambios de objetivos de selección, lo que puede realizarse *a priori* o luego de observar la respuesta a la selección. Claramente la eficiencia de este método es generalmente baja o muy baja, en particular si las características poseen una fuerte correlación genética en sentido desfavorable. Por ejemplo, si mejoramos primero peso de vellón durante años, solo prestando atención a esa característica, lo más probable es que el grosor de la lana se incremente en forma sustantiva. Pasamos ahora a seleccionar por diámetro de las fibras y lentamente recuperamos la finura que habíamos perdido, pero ahora el peso del vellón volvió a bajar. Luego de un ciclo entero posiblemente estemos muy cerca del punto de partida, lo que no parece una buena estrategia. Por otro lado, para características donde la correlación genética es favorable, como vimos en el capítulo «Correlaciones y respuesta correlacionada», la selección por una característica «arrastrará» a la otra en sentido correcto y luego, al cambiar a seleccionar por la segunda característica, también la selección por la segunda mejorará la primera. En este caso, lo que consigue este método de selección es, de alguna manera, balancear el avance en las dos características, con relación a seleccionar solo por una de ellas.

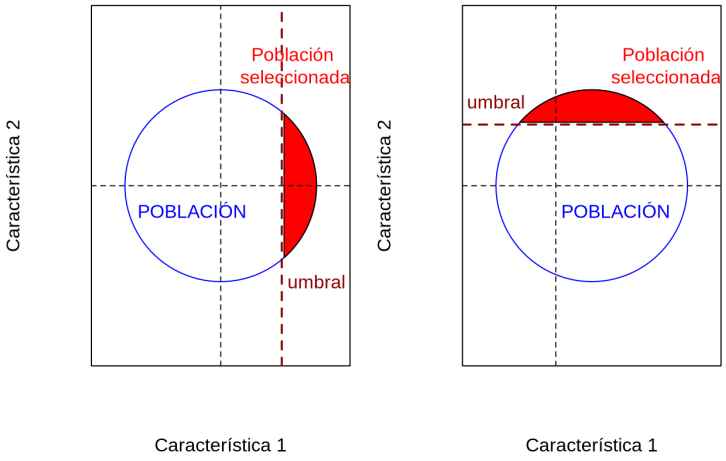


Figura 13.4. Selección en tándem. Durante la primera fase (imagen de la izquierda) se selecciona por la característica 1, de la manera usual en selección fenotípica individual para una sola característica (por ejemplo, selección umbral). El proceso se repite durante un determinado período de generaciones o años, para luego pasar a seleccionar por la segunda característica (imagen de la derecha), de la misma manera que lo hicimos en la primera fase. Notemos el desplazamiento de la media de la población de la característica 1 hacia la derecha, antes de empezar el proceso de selección por la segunda característica. El proceso entero se repite todas las veces que sea necesario.

Niveles independientes de rechazo

Una alternativa al método de selección por tándem es la selección por el método de **niveles independientes de rechazo** (o **NIR**, por sus siglas). La idea, esquematizada en la figura 13.5, es muy sencilla: establecer un umbral independiente para cada característica y seleccionar como reproductores solamente a aquellos individuos que superen los dos criterios (el área en rojo en la figura).

A diferencia del método de selección en tándem, en el método **NIR** el foco se pone siempre en las dos características, lo que evita el comportamiento «en dientes de sierra» que presenta el método del tándem (cuando funciona). El principal problema que aparece en

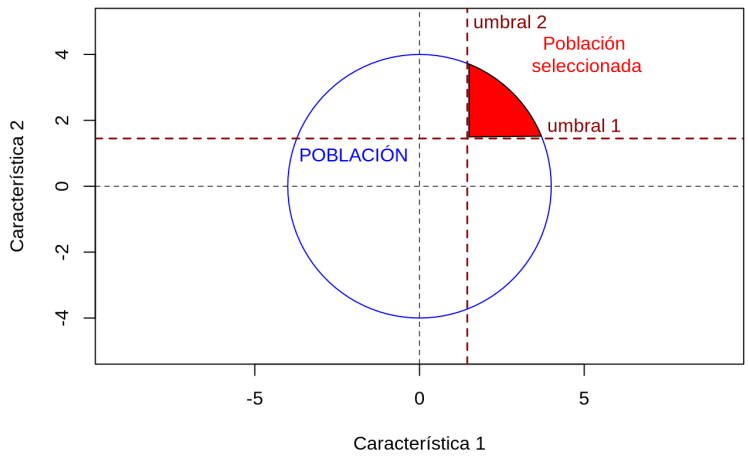


Figura 13.5. Niveles Independientes de Rechazo (NIR). Para dos características se fijan previamente los umbrales (líneas a trazos rojos) en forma independiente y los animales que «pasen» dichos umbrales serán los futuros reproductores. La proporción de animales seleccionados disminuye rápidamente a medida que incorporamos más características independientes.

forma obvia con este método es que a medida que establecemos umbrales para diferentes características, el número de individuos que pasa todos los filtros es cada vez menor. De hecho, si en todas las características seleccionamos la misma proporción de individuos y las características son independientes entre sí, la proporción total seleccionada estará dada por $p = p_i^k$. Por lo tanto, si definimos una proporción p determinada de individuos que necesitamos como reproductores (algo que estará dado por la eficiencia reproductiva alcanzada y por la estructura de edades de la población), la proporción p_i seleccionada para cada característica estará dada por

$$p = p_i^k \therefore p_i = \sqrt[k]{p} \tag{13.61}$$

con $p_i = p_j \; \forall \; i,j$

Veamos un ejemplo que ilustra las implicancias prácticas de esto. Supongamos que tenemos 4 características de nuestro interés, que son totalmente independientes entre ellas y

con distribución normal todas. Precisamos retener una proporción de 20 % de los animales como reproductores, por lo que $p = 0{,}20$. Aplicando la ecuación (13.61), tenemos que la proporción seleccionada en cada características sería

$$p_i = \sqrt[k]{p} = \sqrt[4]{p} = p^{\frac{1}{4}} = 0{,}6687 \tag{13.62}$$

Por lo tanto, reteniendo una proporción $p = 0{,}20$ de animales, que en una característica sola me conduciría a una intensidad de selección de $i = 1{,}40$, en cada una de las características solo selecciono una proporción $p_i = 0{,}6687$, lo que me conduce a una intensidad en cada una de $i \approx 0{,}545$.

Por otro lado, como veremos enseguida, este método puede eliminar individuos excepcionalmente buenos en una característica, pero que estén levemente por debajo del umbral en cualquiera de las otras características, o aún individuos excelentes en todas las características excepto una.

Índices de selección

El comentario precedente sobre los inconvenientes del método **NIR** abre las puertas para considerar algún método que no presente el inconveniente de excluir buenos candidatos en muchas características por el resultado relativamente malo en una sola de ellas. Esto nos lleva al concepto de mérito global o mérito agregado del candidato. Es decir, debemos encontrar una combinación de los valores en las distintas características que pueda resumirse en un solo valor, el cuál será entonces nuestro nuevo valor fenotípico del candidato.

Dicho de otra manera, construimos un nuevo fenotipo, de forma relativamente arbitraria y **vamos a seleccionar por ese nuevo fenotipo que es el índice**. En general, por simplicidad, se utilizan índices que son una combinación lineal de los fenotipos de las características de interés, aunque también existen índice no-lineales en las características. Vamos a dedicar toda la sección «Índices de selección para múltiples características» a describir las bases de la teoría de índices de selección aplicados al mejoramiento de varias características, pero antes de pasar a eso vamos a describir brevemente la idea y su aplicación en mejoramiento genético.

Como vimos antes, uno de los problemas de considerar los caracteres como independientes es la reducción en la eficiencia producto de la necesidad de definir límites arbitrarios (los umbrales) para las distintas características. Si encontramos una forma racional de ponderar la incidencia de cada una de las k características en la función de mérito global del individuo, entonces el fenotipo agregado del individuo será igual a

$$I = b_1\, z_1 + b_2\, z_2 + \ldots + b_k\, z_k = \sum_{i=1}^{k} b_i\, z_i = \mathbf{b^T z} \tag{13.63}$$

Esto es una función lineal de todas las características, con pesos fenotípicos b_i. Supongamos que resolvimos el problema de cómo determinar los pesos, cosa que veremos cómo hacer en la sección Índices de selección para múltiples características. En ese caso, si tenemos los fenotipos como desvío de la media para cada característica, alcanza con multiplicar cada

fenotipo z_i por su correspondiente ponderación b_i y sumar todos los resultados para tener el mérito agregado del individuo, que será su nuevo fenotipo. A partir de esto podemos aplicar el criterio de selección individual que sea de nuestro agrado.

En la figura 13.6 se representa el funcionamiento del índice. Para dos características el mismo será una combinación lineal de ambas, cuya dirección está descrita por la recta a trazos de color rojo. Conociendo la proporción de animales que necesitamos seleccionar, alcanza con ordenar los individuos por el índice y determinar el corte en el punto en que haya esa proporción con índice igual o mayor,

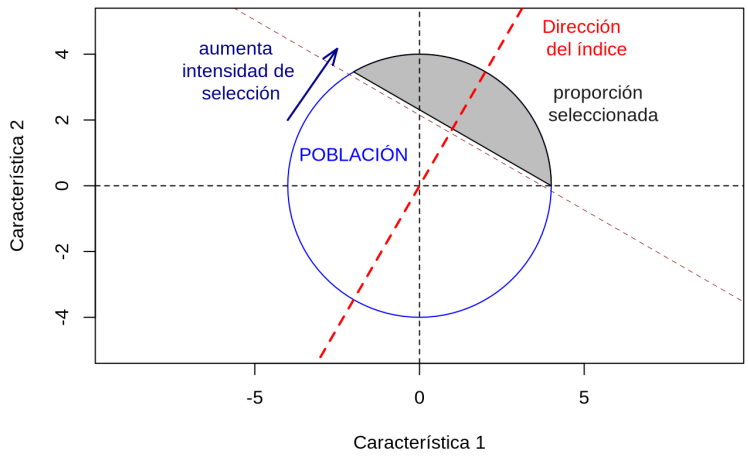


Figura 13.6. Índice de selección para dos características. El índice es una combinación lineal de ambas características y tiene la dirección de la línea roja a trazos gruesos. Pasamos de tener dos valores fenotípicos para cada individuo a un solo valor que representa el mérito global o agregado del mismo.

Dentro de las ventajas que posee esta combinación lineal está la de no dejar afuera fenotipos (genotipos) de alto valor en una característica porque no pasan en otra. Si nos fijamos en la figura 13.6 nuevamente, veremos que hay una proporción de individuos seleccionados (sombreado en gris) que se encuentran a la izquierda del eje de las ordenadas. Es decir, para la característica 1 son peores que la media de los individuos de la población. Sin embargo, para la característica 2 son excepcionalmente valiosos. Posiblemente, en **NIR** hubiesen quedado por el camino. Sin embargo, el mérito global de los mismos es suficientemente elevado como para que justifique retenerlos, cosa que hace el índice.

Otra de las ventajas que presenta un índice es la simplicidad de uso. En lugar de tener que inspeccionar para cada posible reproductor el mérito en k características, alcanza con el valor en una sola. En el caso de que se trate de semen, por ejemplo, donde existe una gran oferta de reproductores, dentro del mismo intervalo de valores del índice es posible luego afinar el criterio y elegir los reproductores mejores en las características que más nos interesan. Por ejemplo, si un rodeo presenta problema de baja cantidad de proteínas, dentro de los

toros con determinado valor del índice puedo elegir aquellos que tengan ese valor porque son excelentes en proteína (en general puedo hacerlo sin empeorar otras características).

En la actualidad, existen índices de selección para muchas razas, lo que permite obtener un avance eficiente en la selección. Para esto se requiere, no solamente contar con grandes volúmenes de datos que permitan obtener estimaciones adecuadas de los parámetros genéticos (algo que veremos enseguida), sino también alcanzar un acuerdo entre los productores acerca de tipos de sistemas de producción para los que se desea que el índice de selección sea una fiel representación del sistema de ingresos y egresos de los establecimientos. Como suele ser imposible que un solo índice represente a diferentes tipos de establecimientos, en muchos casos se tendrán tantos índices como situaciones los datos y el esfuerzo permitan generar, cada uno de ellos eficiente para su propio sistema de producción. Un ejemplo típico de esto es en producción lechera, donde dependiendo del tipo de producción (estabulada, pastoril, etcétera) y dependiendo del destino de la leche (industria, producción de quesos, etcétera) se deben desarrollar diferentes índices que representen esas combinaciones. En lo que sigue veremos la teoría que sustenta el desarrollo de índices de selección como forma de mejorar la eficiencia y simplificar la selección.

Ejemplo 13.3

Supongamos que tenemos rodeo lechero y queremos mejorar simultáneamente las siguientes características: producción de leche, producción de grasa, producción de proteína, puntaje de células somáticas y tasa de preñez en hijas. Para esto, nos ofrecen a la venta el semen de 7 toros disponibles, con la siguiente información:

Toro	Leche		Grasa		Proteína		cs[1]		THP[2]		Índice de selección[3]
	DEP[4]	r_{AC}	DEP[4]	r_{AC}	DEP[4]	r_{AC}	DEP[4]	r_{AC}	DEP[4]	r_{AC}	
HOLDNKM000000260548	57	0,59	12,8	0,57	13,2	0,57	−0,50	0,62	0,67	0,56	51,75
HOL840M003147224052	288	0,58	19,6	0,57	17,5	0,57	−0,49	0,60	−0,42	0,58	50,30
HOLNLDM000580616806	31	0,60	10,1	0,59	9,4	0,59	−0,62	0,63	−1,21	0,60	45,90
HOLFRAM002933738576	48	0,63	18,2	0,62	8,9	0,62	−0,35	0,67	−0,82	0,64	36,42
HOLNLDM000694388754	154	0,62	15,4	0,61	11,5	0,62	−0,36	0,67	−2,85	0,65	33,78
HOL840M003150687270	328	0,56	17,7	0,55	17,6	0,55	−0,04	0,58	0,23	0,57	30,93
HOLCANM000012529241	308	0,59	21,0	0,59	14,5	0,59	−0,10	0,65	−1,85	0,59	25,69

[1] células somáticas; [2] tasa de preñez en hijas; [3] $= -0{,}05 \times DEP(L) + 0{,}33 \times DEP(G) + 2{,}25 \times DEP(P) + 1{,}31 \times DEP(TPH) - 39{,}6 \times DEP(SCS)$; [4] $DEP = \frac{1}{2}\hat{A}$

Para tomar la decisión de selección, nos planteamos los siguientes niveles de exigencia:

- leche: DEP menor a 100 litros
- grasa: DEP mayor a 10 kg
- proteína: DEP mayor a 9 kg
- células somáticas: DEP menor a $-0{,}40$
- tasa de preñez en hijas: DEP mayor a $-1{,}5$

a. De acuerdo con el método de selección **NIR**, ¿a cuáles toros seleccionaríamos? **b.** Si en lugar de aplicar el método de selección **NIR**, aplicáramos el método de **Índice de selección**, ¿seleccionaríamos a los mismos toros? **c.** Interpretar el índice de selección utilizado.

a. El método **NIR** implica que cada criterio es excluyente de por sí. Es decir, alcanza con que no cumpla uno solo como para que el candidato quede eliminado.

Por *leche*, eliminaríamos a los toros HOL840M003147224052, HOLNLDM000694388754, HOL840M003150687270 y HOLCANM000012529241. Luego, por *grasa* no eliminaríamos a ningún toro.

Por *proteína* eliminaríamos al toro HOLFRAM002933738576. Por *células somáticas* y *tasa de preñez en hijas* no eliminaríamos a ningún toro (los que no cumplen con el nivel de exigencia, ya fueron eliminados por las características anteriores).

Por lo tanto, seleccionaríamos a los toros **HOLDNKM000000260548** y **HOLNLDM000580616806**.

b. Si aplicáramos el método de **índice de selección**, los dos toros que seleccionaríamos serían **HOLDNKM000000260548** y **HOL840M003147224052**.

La diferencia se da en que mediante el método de selección **NIR** el toro HOL840M003147224052 es eliminado debido a su mal desempeño en producción de leche. Sin embargo, el método de **índice de selección** compensa el mal desempeño en unos caracteres con el buen desempeño en otros caracteres, además de tener en cuenta el coeficiente de ponderación de cada carácter.

c. El **índice de selección** utilizado es igual a $-0{,}05 \times DEP(L) + 0{,}33 \times DEP(G) + 2{,}25 \times DEP(P) + 1{,}31 \times DEP(TPH) - 39{,}6 \times DEP(SCS)$, por lo tanto:

- Pondera en forma **negativa** el **volumen de leche** (coeficiente $-0{,}05$), lo que es típico de mercados que priorizan los sólidos (destino queso, manteca, leche en polvo, etcétera).
- Favorece la producción de Grasa (coeficiente $+0{,}33$) y Proteína (coeficiente $+2{,}25$). La relación entre ambas ponderaciones refleja una visión «antigua» que no se condice en l actualidad con los precios de ambos productos (el valor de la grasa de leche se ha incrementado notoriamente en los últimos años).
- La tasa de preñez en hijas del candidato es fundamental para mantener el nivel de reemplazos en el rodeo y para poder hacer selección en hembras (coeficiente $+1{,}31$).
- El recuento de células somáticas representa, junto con el conteo microbiológico, uno de los parámetros claves en la calidad de la leche y es un factor clave en el pago (con bonificaciones por número reducido y castigo por elevado). El score de células somáticas es la medida que se utiliza en las evaluaciones y por lo tanto su ponderación es fuertemente negativa (coeficiente $-39{,}6$).

PARA RECORDAR

- La **selección en tándem** se basa en seleccionar primero por una sola característica y luego de un tiempo cambiar a otra característica de interés. Es de destacar que la complejidad de este método es baja, aunque también lo es su eficiencia, particularmente si las características poseen una fuerte correlación genética en sentido desfavorable.
- La selección por **niveles independientes de rechazo (NIR)** se basa en establecer un umbral independiente para cada característica y seleccionar como

reproductores solamente a aquellos individuos que superen todos los criterios. El principal problema que aparece con este método es que a medida que establecemos umbrales para diferentes características, el número de individuos que pasa todos los filtros es cada vez menor. A su vez, este método puede eliminar individuos excepcionalmente buenos en una característica, pero que estén levemente por debajo del umbral en cualquiera de las otras características, o aún individuos excelentes en todas las características excepto una.

- Los **índices de selección** evalúan el mérito global o mérito agregado de los candidatos. Si encontramos una forma racional de ponderar la incidencia de cada una de las k características en la función de mérito global del individuo, entonces el fenotipo agregado del individuo será igual a $I = b_1\ z_1 + b_2\ z_2 + \ldots + b_k\ z_k\ =\ \sum_{i=1}^{k} b_i\ z_i\ =\ \mathbf{b^T z}$. Esto es una función lineal de todas las características, con pesos fenotípicos b_i. Dentro de las ventajas que posee esta combinación lineal está la de no dejar afuera fenotipos (genotipos) de alto valor en una característica porque no pasan en otra además de la simplicidad del uso del índice una vez calculado.

13.5 Índices de selección para múltiples características

Previamente, en la sección «Combinando la información proporcionada por diferentes tipos de parientes», discutimos acerca del uso de los índices de selección como una manera de combinar información fenotípica proveniente de diferentes relaciones de parentesco a fin de obtener una mejor estimación del valor de cría de un individuo. En esta sección vamos a ver el tema de los índices de selección nuevamente, pero desde una perspectiva diferente: la combinación de la información de diferentes características en un solo valor a partir del cual elegir los mejores reproductores.

Vamos a usar la misma notación que usamos al final del capítulo anterior para las matrices de varianza-covarianza genética ($\mathbf{G}$) y fenotípica ($\mathbf{P}$), mientras que usaremos la notación, que ya usamos antes, de $\sigma(\cdots, \cdots)$ para la covarianza, tanto escalar como vectorial. Las ideas de esta sección se basan en el libro de Bruce Walsh y Michael Lynch (2018), que constituyen un excelente material para ampliar lo tratado aquí.

Nuevamente, al tratarse de un índice lineal sobre medidas fenotípicas, nuestro índice será de la forma

$$I = b_1\ z_1 + b_2\ z_2 + \ldots + b_k\ z_k = \mathbf{b'z} \tag{13.64}$$

donde ahora los distintos z_i serán diferentes características fenotípicas. Por ejemplo, en producción lechera es de nuestro interés aumentar la producción de grasa y proteína, al mismo tiempo que reducir el volumen de leche y el recuento de células somáticas. Todas estas características (grasa, proteína, volumen de leche, recuento de células somáticas) serían los z_i.

Al ser el índice una combinación lineal de otras características, podemos considerarlo como una nueva característica fenotípica. Como ya vimos antes (en otro contexto) en la sección «Precisión del valor de cría a partir del índice», la varianza del índice es igual a

$$\sigma_I^2 = \sigma(\mathbf{b}'\mathbf{z},\mathbf{b}'\mathbf{z}) = \mathbf{b}'\sigma(\mathbf{z},\mathbf{z})\mathbf{b} = \mathbf{b}'\mathbf{P}\mathbf{b} \tag{13.65}$$

mientras que la varianza genético-aditiva del mismo sería

$$\sigma_{A_I}^2 = \sigma_A(\mathbf{b}'\mathbf{z},\mathbf{b}'\mathbf{z}) = \mathbf{b}'\sigma_A(\mathbf{z},\mathbf{z})\mathbf{b} = \mathbf{b}'\mathbf{G}\mathbf{b} \tag{13.66}$$

Por lo tanto, combinando las ecuaciones (13.65) y (13.66), tenemos que la heredabilidad del índice es igual a

$$h_I^2 = \frac{\sigma_{A_I}^2}{\sigma_I^2} = \frac{\mathbf{b}'\mathbf{G}\mathbf{b}}{\mathbf{b}'\mathbf{P}\mathbf{b}} \tag{13.67}$$

A riesgo de mencionar lo obvio, tanto $\mathbf{b}'\mathbf{G}\mathbf{b}$ como $\mathbf{b}'\mathbf{P}\mathbf{b}$ son escalares (números), ya que si tenemos k características las dimensiones de ambos productos son $[1 \times k][k \times k][k \times 1]$ y de acuerdo con las reglas de multiplicación de matrices los índices internos «colapsan», quedando las dimensiones del resultado $[1 \times 1]$, es decir un escalar.

Asumiendo que la distribución de los valores de cría $\mathbf{g}$ y los fenotipos $\mathbf{z}$ es conjuntamente normal multivariada, como ya vimos previamente, entonces se aplica la **ecuación del criador** a la respuesta esperada del índice y utilizando los resultados de las ecuaciones (13.65) y (13.67), con una intensidad de selección i, tenemos

$$R_I = i\, h_I^2\, \sigma_I = i\, \frac{\mathbf{b}'\mathbf{G}\mathbf{b}}{\mathbf{b}'\mathbf{P}\mathbf{b}}\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}} = i\, \frac{\mathbf{b}'\mathbf{G}\mathbf{b}}{\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}}} \tag{13.68}$$

La pregunta que nos surge ahora es **¿cómo cambian las diferentes características que conforman el índice cuando seleccionamos por este?** Para eso, bajo los supuestos usuales para que se cumpla la ecuación del criador, tenemos que el cambio en la media será igual a

$$\Delta\mu = \mathbf{G}\mathbf{P}^{-1}\mathbf{s} \tag{13.69}$$

Debemos pues encontrar el vector del diferencial de selección $\mathbf{s}$. Consideremos un carácter, por ejemplo, el carácter j y su diferencial s_j. Como la correlación entre el **fenotipo del carácter** j y el **fitness** lo podemos expresar como

$$\rho_{z_j,w} = \rho_{z_j,I} \times \rho_{I,w} \tag{13.70}$$

entonces, expresado en términos de varianzas y covarianzas, tenemos que

$$\frac{\sigma(z_j,w)}{\sigma_{z_j}\sigma_w} = \left(\frac{\sigma(z_j,I)}{\sigma_{z_j}\sigma_I}\right)\left(\frac{\sigma(I,w)}{\sigma_I\sigma_w}\right) \tag{13.71}$$

Al mismo tiempo, si recordamos la identidad de Robertson-Price que vimos en la sección «El diferencial de selección direccional y la identidad de Robertson-Price», la misma establecía

que **el diferencial de selección era igual a la covarianza entre el fenotipo y el fitness** correspondiente a ese fenotipo, es decir $\sigma(z_j,w = s_j$. En forma análoga, si el fenotipo es el índice, entonces $\sigma(I,w) = s_I = i\sigma_I$, con i la intensidad de selección en el índice. La última igualdad proviene, obviamente, de $i = \frac{s}{\sigma_P} \Leftrightarrow s = i\,\sigma_P$.

Utilizando estos resultados y resolviendo para $\sigma(z_j,w)$ en la ecuación (13.71), que como vimos es igual al diferencial de selección que nos interesa, tenemos

$$\sigma(z_j,w) = \sigma_{z_j}\sigma_w \left(\frac{\sigma(z_j,I)}{\sigma_{z_j}\sigma_I}\right)\left(\frac{\sigma(I,w)}{\sigma_I\sigma_w}\right) = \frac{\sigma(z_j,I)\sigma(I,w)}{\sigma_I^2} = i\,\frac{\sigma(z_j,I)}{\sigma_I} \tag{13.72}$$

Por otro lado, también podemos escribir

$$\sigma(z_j,I) = \sigma(z_j,\sum_k b_k z_k) = \sum_k b_k P_{jk} \tag{13.73}$$

donde P_{ij} es el elemento de la fila i y columna j en la matriz $\mathbf{P}$. De esta forma, podemos ahora escribir el resultado de la ecuación (13.71) como

$$\sigma(z_j,w) = s_j = \left(\frac{i}{\sigma_I}\right) \times \sum_k b_k P_{jk} \tag{13.74}$$

lo que para el conjunto de características pasa a ser entonces el vector de diferenciales de selección

$$\mathbf{s} = \left(\frac{i}{\sigma_I}\right) \times \mathbf{Pb} \tag{13.75}$$

Finalmente, combinando el resultado de la ecuación (13.69) con este vector de diferenciales de selección, tenemos que la respuesta a la selección será

$$\Delta\mu = \mathbf{GP}^{-1}\mathbf{s} = \left(\frac{i}{\sigma_I}\right) \times \mathbf{GP}^{-1}\mathbf{Pb} = \left(\frac{i}{\sigma_I}\right) \times \mathbf{Gb} \tag{13.76}$$

y como $\sigma_I = \sqrt{\sigma_I^2} = \sqrt{\mathbf{b}^T\mathbf{Pb}}$, la ecuación anterior se puede escribir como

$$\Delta\mu = \left(\frac{i}{\sigma_I}\right)\frac{\mathbf{Gb}}{\sqrt{\mathbf{b}^T\mathbf{Pb}}} \tag{13.77}$$

El índice de Smith-Hazel

En esta sección vamos a introducir un cambio en la notación para indicar la transpuesta de una matriz o un vector, ya que usaremos ^{T}p, dejando el uso de $'$ para las derivadas. Normalmente, la idea de usar índices de selección es resumir en un valor el mérito general del individuo. Si tenemos un conjunto de valores económicos para cada una de las características, entonces podemos ordenar estos pesos relativos en un vector $\mathbf{a}$ y por lo tanto, el mérito general del individuo (a partir de su fenotipo en las distintas características) es

$$J = \mathbf{a^T z} \tag{13.78}$$

Claramente, J nos habla del mérito fenotípico del individuo, pero nuestro objetivo es seleccionar a los mejores reproductores, por lo que si llamamos g_i al valor de cría en la característica $_i$, entonces el correspondiente **mérito genético agregado** será

$$H = \sum_i a_i g_i = \mathbf{a^T g} \tag{13.79}$$

y

$$H - \mu_H = H - \mathbf{a^T}\mu \tag{13.80}$$

será el valor de cría del individuo. Debido a las estructuras de dependencias entre características (varianzas y covarianzas genéticas y fenotípicas), por lo general J no es el mejor predictor del mérito genético agregado H.

Como vimos antes, bajo los supuestos tradicionales de ausencia de interacción genotipo-ambiente, epistasis, etcétera, la media de valores de cría de los reproductores seleccionados será igual a la media esperada en los valores fenotípicos de la progenie. Es decir, el cambio en H dentro de la misma generación será igual al cambio esperado en J en la siguiente generación. Este cambio se maximiza seleccionando entonces a los individuos con mayores valores de H y nuestra tarea será entonces encontrar un índice $I_s = \mathbf{b_s^T z}$ tal que maximice la correlación con H.

La correlación entre un índice arbitrario $I = \mathbf{b^T z}$, y el mérito genético agregado H es

$$\rho_{I,H} = \frac{\sigma_{I,H}}{\sigma_I \sigma_H} \tag{13.81}$$

La varianza en el mérito genético agregado será igual a

$$\sigma^2_H = \sigma(\mathbf{a^T z}, \mathbf{a^T z}) = \mathbf{a^T G a} \tag{13.82}$$

mientras que, como ya vimos antes en la ecuación (13.65), la varianza del índice es igual a

$$\sigma^2_I = \mathbf{b^T P b} \tag{13.83}$$

Por otra parte, la covarianza entre el mérito genético agregado y el índice es

$$\sigma_{I,H} = \sigma(\mathbf{a^T g}, \mathbf{b^T z}) = \mathbf{a}^T \sigma(\mathbf{g},\mathbf{z})\mathbf{b} = \mathbf{a}^T \sigma(\mathbf{g}, \mathbf{g}+\mathbf{e})\mathbf{b} = \mathbf{a}^T \sigma(\mathbf{g},\mathbf{g})\mathbf{b} = \mathbf{a^T G b} \tag{13.84}$$

Poniendo todo junto, tenemos que la correlación entre el índice arbitrario y el mérito genético agregado es igual a

$$\rho_{I,H} = \frac{\sigma_{I,H}}{\sigma_I \sigma_H} = \frac{\mathbf{a^T G b}}{\sqrt{\mathbf{b^T P b}}\sqrt{\mathbf{a^T G a}}} \tag{13.85}$$

Para hallar el máximo de esta función (función de $\mathbf{b}$) debemos hacer la derivada e igualarla a cero. Para esto, una ayuda importante viene de notar que el término $\mathbf{a^T G a}$ es una constante, por lo que la optimización no dependerá de él, es decir, debemos optimizar $\frac{\mathbf{a^T G b}}{\sqrt{\mathbf{b^T P b}}}$. Además, tanto el numerador como el denominador son productos cuadráticos que como ya vimos son simples escalares, por lo que aplicaremos las técnicas estándar de derivación sobre ellos y solo luego utilizaremos derivación sobre vectores y matrices. En primer lugar, podemos usar la regla de la cadena para el cociente en la ecuación (13.85), es decir $\left(\frac{u(x)}{v(x)}\right)' = \frac{u'v - v'u}{v^2}$. Además, sabemos que la derivada de una forma cuadrática $x^T Ax$ donde además la matriz es simétrica es igual a $2x^T A$. Por lo que, utilizando la regla de la cadena y la derivada de una forma cuadrática, tenemos que $(\sqrt{\mathbf{b^T P b}})' = \frac{\mathbf{b^T P}}{\sqrt{\mathbf{b^T P b}}}$ y entonces, luego de igualar a cero la derivada, si llamamos $\tilde{\mathbf{b}}$ a las soluciones para $\mathbf{b}$, entonces

$$\left(\frac{\mathbf{a^T G b}}{\sqrt{\mathbf{b^T P b}}}\right)' = \frac{\mathbf{a}^T\mathbf{G}(\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}) - [\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}]'(\mathbf{a}^T\mathbf{G}\mathbf{b})}{\mathbf{b}^T\mathbf{P}\mathbf{b}} \tag{13.86}$$

La derivada que tenemos en el numerador, como ya vimos, es la derivada de una forma cuadrática, por lo que

$$[\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}]' = \frac{2\mathbf{b}^T\mathbf{P}}{2\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}} = \frac{\mathbf{b}^T\mathbf{P}}{\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}} \tag{13.87}$$

por lo que sustituyendo en la ecuación (13.86) e igualando a cero, teniendo solo en cuenta el numerador porque para que sea cero el denominador no importa, tenemos

$$\mathbf{a}^T\mathbf{G}(\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}) - \frac{\mathbf{b}^T\mathbf{P}}{\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}}(\mathbf{a}^T\mathbf{G}\mathbf{b}) = 0 \Leftrightarrow$$

$$\mathbf{a}^T\mathbf{G}(\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}})(\sqrt{\mathbf{b}^T\mathbf{P}\mathbf{b}}) = \mathbf{b}^T\mathbf{P}(\mathbf{a}^T\mathbf{G}\mathbf{b}) \Leftrightarrow$$

$$\mathbf{G}^T\mathbf{a}(\mathbf{b}^T\mathbf{P}\mathbf{b}) = \mathbf{P}^T\mathbf{b}(\mathbf{a}^T\mathbf{G}\mathbf{b})$$

Si llamamos $\tilde{\mathbf{b}}$ al vector en su solución, entonces

$$(\tilde{\mathbf{b}}^T\mathbf{P}\tilde{\mathbf{b}})\mathbf{G}\mathbf{a} = (\mathbf{a}^T\mathbf{G}\tilde{\mathbf{b}})\mathbf{P}\tilde{\mathbf{b}} \tag{13.88}$$

Por lo tanto, como $(\tilde{\mathbf{b}}^T\mathbf{P}\tilde{\mathbf{b}})$ y $(\mathbf{a}^T\mathbf{G}\tilde{\mathbf{b}})$ son escalares, la solución para $\mathbf{b}$ será de la forma

$$\mathbf{P}\tilde{\mathbf{b}} = c \times \mathbf{G}\mathbf{a} \tag{13.89}$$

con $c = \frac{\tilde{\mathbf{b}}^T\mathbf{P}\tilde{\mathbf{b}}}{\mathbf{a}^T\mathbf{G}\tilde{\mathbf{b}}}$ y por lo tanto

$$\tilde{\mathbf{b}} = c \times \mathbf{P}^{-1}\mathbf{G}\mathbf{a} \tag{13.90}$$

donde c es un escalar que es **irrelevante para el peso relativo** de los distintos componentes en el índice. Finalmente, utilizando este resultado, el índice de Smith-Hazel es

$$I_s = \mathbf{b}_s^T \mathbf{z} = (\mathbf{P}^{-1}\mathbf{G}\mathbf{a})^T \mathbf{z} \tag{13.91}$$

El índice de Smith-Hazel necesita conocer con precisión ambas matrices de varianza-covarianza, **G** y **P**, así como el vector de pesos **a**. En general las matrices de varianza-covarianza deben estimarse a partir de los datos y por lo tanto no son conocidas sin error, lo que tiene implicancias para la optimalidad del índice. Por otra parte, el índice parece ser relativamente resistente a los errores de estimación en **a**.

El índice de Smith-Hazel posee algunas propiedades que lo hacen muy interesante:

1. Provee la mayor respuesta posible en el mérito cuando la intensidad de selección es fija.
2. El índice está estrechamente relacionado con la solución de mínimos cuadrados para la regresión del mérito en los valores fenotípicos z y con las soluciones del método BLUP, que veremos más adelante en este capítulo.
3. Para una muestra dada, el índice de Smith-Hazel da la probabilidad máxima de seleccionar al individuo con mayor mérito agregado.

La respuesta a la selección utilizando el índice de Hazel-Smith estará dada por

$$R_H = i\, \frac{\mathbf{a}^T\mathbf{G}\mathbf{b}_s}{\sqrt{\mathbf{b}_s^T\mathbf{P}\mathbf{b}_s}} = i\, \sqrt{\mathbf{a}^T\mathbf{G}\mathbf{b}_s} = i\, \sqrt{\mathbf{a}^T\mathbf{G}\mathbf{P}^{-1}\mathbf{G}\mathbf{a}} \tag{13.92}$$

El índice de Smith-Hazel se puede extender fácilmente a un número diferentes de características fenotípicas que las consideradas en el mérito genético agregado y de hecho ni siquiera hace falta que tengan solapamiento las mismas, aportando la información la estructura de varianza-covarianza genética entre características.

Para esto definimos **P** como la matriz de varianza-covarianza fenotípica para el vector **z**, de dimensiones $[m \times m]$ y **G**, de dimensiones $[m \times n]$, cuyos elementos son $G_{kj} = \sigma(z_k, g_j)$ para $1 \leqslant k \leqslant m$, $1 \leqslant j \leqslant n$, a la matriz de covarianzas genéticas entre los m valores fenotípicos y los n caracteres comprendidos en H.

Esto nos lleva a un mérito genético agregado igual a

$$H = \sum_{j=1}^{n} a_j\, g_j \tag{13.93}$$

y un índice fenotípico

$$I_s = \sum_{k=1}^{m} b_k\, z_k \tag{13.94}$$

A partir de estas definiciones, la solución para el índice sigue siendo de la forma $\mathbf{b}_s = \mathbf{P}^{-1}\mathbf{Ga}$, pero dado que $\mathbf{G}$ ya no es más simétrica, en este caso la respuesta estará dada por

$$R_H = i\sqrt{\mathbf{a}^T\mathbf{G}^T\mathbf{P}^{-1}\mathbf{Ga}} \tag{13.95}$$

PARA RECORDAR

- La heredabilidad del índice es igual a $h_I^2 = \frac{\sigma_{A_I}^2}{\sigma_I^2} = \frac{\mathbf{b}'\mathbf{Gb}}{\mathbf{b}'\mathbf{Pb}}$ con una respuesta a la selección equivalente a $\Delta\mu = \left(\frac{i}{\sigma_I}\right)\frac{\mathbf{Gb}}{\sqrt{\mathbf{b}^T\mathbf{Pb}}}$
- Si tenemos un conjunto de valores económicos para cada una de las características, entonces podemos ordenar estos pesos relativos en un vector $\mathbf{a}$ y por lo tanto, el mérito general del individuo (a partir de su fenotipo en las distintas características) es $J = \mathbf{a^Tz}$
- El índice de Smith-Hazel $I_s = \mathbf{b}_s^T\mathbf{z} = (\mathbf{P}^{-1}\mathbf{Ga})^T\mathbf{z}$ necesita conocer con precisión ambas matrices de varianza-covarianza, $\mathbf{G}$ y $\mathbf{P}$, así como el vector de pesos $\mathbf{a}$.

1) Provee la mayor respuesta posible en el mérito cuando la intensidad de selección es fija.
2) Está estrechamente relacionado con la solución de mínimos cuadrados para la regresión del mérito en los valores fenotípicos z y con las soluciones del método BLUP.
3) Para una muestra dada, da la probabilidad máxima de seleccionar al individuo con mayor mérito agregado.

- La **respuesta a la selección utilizando el índice de Smith-Hazel** estará dada por $R_H = i\sqrt{\mathbf{a}^T\mathbf{GP}^{-1}\mathbf{Ga}}$

13.6 Métodos y técnicas avanzadas de selección

En la primera parte de este capítulo discutimos diferentes tipos de selección utilizando información de parientes, **selección familiar**, **selección intrafamiliar**, **selección fraternal**, **pruebas de progenie**, todas utilizando un solo tipo de relaciones de parentesco único. Más adelante, extendimos los tipos posibles de fuentes de información a través de la integración en índices de selección y vimos que era necesario desarrollar un índice para cada tipo de combinación. Esto, sin dudas, resulta extremadamente impráctico ya que es necesario estudiar individuo a individuo las fuentes de información disponibles u optar por no utilizar todas las fuentes disponibles.

Por otra parte, en la segunda parte del capítulo vimos que usualmente debemos trabajar con varias características a la vez para obtener una respuesta que se adecue a las ecuaciones de ingresos y egresos de los establecimientos. Si nos vamos un momento a la sección «La forma generalizada de la ecuación del criador» del capítulo «Correlaciones y respuesta correlacionada», recordaremos que la extensión de la **ecuación del criador** a la selección multivariada involucraba el uso de las matrices de varianza-covarianza genética y fenotípica

y vimos en este capítulo el papel de traspaso de información de una característica a otra a través de dichas matrices.

Afortunadamente, desde mediados del siglo XX contamos con los desarrollos, primero teóricos y luego gracias a la computación prácticos, que nos permiten resolver casi todos estos problemas de manera «automática». En ese sentido, uno de los avances más importantes se debe a Charles R. Henderson, que en una serie de artículos sentó las bases de lo que actualmente conocemos como las ecuaciones del modelo mixto lineal o el **BLUP** directamente y que es la metodología de base en las evaluaciones genéticas.

BLUP: las ecuaciones del modelo mixto lineal

Un modelo mixto lineal es un modelo mixto que combina efectos fijos y efectos aleatorios. Existen muchas definiciones para estos dos tipos de efectos, pero en nuestro contexto unas pocas consideraciones prácticas nos permitirán distinguirlos:

- **Efectos fijos** son aquellos que nos puede interesar contrastar sus niveles, que usualmente tienen pocos niveles y que en otros casos usaremos para compensar diferencias sistemáticas entre niveles. Por ejemplo, en sexo de los animales suele ser considerado un efecto fijo con dos niveles y su inclusión en los modelos compensa las diferencias entre machos y hembras. Normalmente, queremos **estimar** su valor.
- **Efectos aleatorios** son efectos donde asumimos que los datos observados son una muestra aleatoria de una población mayor. Normalmente poseen muchos niveles y los usamos, en nuestro caso, para modelar el valor de cría de los animales, los efectos maternales, etcétera y obviamente, los efectos ambientales y del error. En general, el objetivo es **predecir** su valor ya que se trata de un efecto aleatorio.

La idea de combinar las ecuaciones del modelo mixto en forma matricial aparece por primera vez en Charles Roy Henderson (1949), Charles Roy Henderson (1950) y fue posteriormente extendida por Charles Roy Henderson (1959), entre otros, demostrando alguna de las «propiedades» de este método. Más adelante, el propio Henderson bautizó el método con el acrónimo **BLUP**, por Best Linear Unbiased Predictor. El significado de cada una de estas palabras, de acuerdo con Henderson sería

- **Mejor**: Significa que la varianza del error de predicción $Var(a - \hat{a})$ (que es la diferencia entre los valores «verdaderos» y los predichos) se minimiza, o que, en forma equivalente, la correlación entre ellos se maximiza
- **Lineal:** Significa que los EBV predichos son funciones lineales de los datos (que normalmente se denota como $L'y$)
- **Insesgado**: Significa que el valor esperado de los EBV predichos y de los EBV «verdaderos» son iguales, es decir $E(a = \hat{a})$
- **Predictor**: Como ya vimos, significa que implica una **predicción del verdadero valor de cría** (ya que este es tratado como un efecto aleatorio).

Varias de estas propiedades solo se cumplen si el modelo asumido es el modelo verdadero, un problema filosófico difícil de resolver (por ejemplo, la condición de **Mejor**), aunque se trata de un modelo robusto frente a este supuesto, o el **Insesgado**, que solo es condicional a los hiperparámetros del modelo. Sin embargo, pese a esto, se trata de un modelo que ha sido muy exitoso como forma de modelar las evaluaciones genéticas y de incrementar la respuesta a la selección.

Nosotros no vamos a extendernos en la base teórica, que corresponde a cursos más avanzados, pero comenzaremos a dejar las bases acerca de cómo funciona para que en próximos cursos puedas usar los resultados y entender mejor la teoría.

En mejoramiento genético tenemos información de un número determinado de animales, tanto el fenotipo de las características de interés, como las relaciones de parentesco entre los animales (usualmente en un pedigrí) y de los distintos efectos que incidieron en la manifestación del fenotipo. Sin pérdida de generalidad, asumamos que poseemos información completa para n animales (incluyendo su pedigrí completo), que serán los que pretendemos evaluar. Los efectos fijos vamos a codificarlos en una matriz $\mathbf{X}$, que tendrá tantas filas como individuos con fenotipo, es decir n y tantas columnas como el número total de niveles de los factores considerados, menos el número de factores (para eliminar la dependencia lineal entre columnas). También en la matriz $\mathbf{X}$ podemos incluir las covariables (variables continuas) que sean necesarias (por ejemplo, kilogramos de alimento que recibió, temperaturas promedio o lo que se nos ocurra que explica el fenotipo). Por otra parte, vamos a considerar por simplicidad solamente dos efectos aleatorios, el genético-aditivo de los individuos y el error. Teniendo en cuenta la matriz $\mathbf{A}$ de parentesco aditivo entre los individuos (que si no consideramos a los progenitores de los individuos con registros será de dimensiones $[n \times n]$), la forma usual de especificar nuestro modelo mixto será

$$\begin{aligned} \mathbf{y} &= \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \\ \mathbf{u} &\sim (0, \mathbf{G}), \mathbf{G} = \mathbf{A}\sigma_A^2, \\ \mathbf{e} &\sim (0, \mathbf{R}), \mathbf{R} = \mathbf{I}\sigma_E^2, \\ cov(\mathbf{u}, \mathbf{e}') &= 0 \end{aligned} \tag{13.96}$$

El aporte sustantivo de Henderson fue demostrar que las ecuaciones correspondientes a este modelo se podían escribir en forma matricial como

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{13.97}$$

La solución de dicho sistema de ecuaciones se corresponde con la estimación de los efectos fijos (**BLUE**) y aleatorios (**BLUP**) del modelo mixto.

De hecho, en sistemas de tamaño pequeño, para obtener la solución alcanza con invertir la matriz del lado izquierdo de la ecuación (13.97) y luego premultiplicar por dicha inversa a ambos lados de la ecuación, es decir

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \tag{13.98}$$

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix} \therefore \tag{13.99}$$

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix} \tag{13.100}$$

En las evaluaciones genéticas los sistemas de ecuaciones suelen tener cientos, miles o aún millones de líneas, por lo que los métodos computacionales involucrados son mucho más sofisticados que la inversión directa. Además, para obtener las soluciones de acuerdo con las ecuaciones del **BLUP** es necesario conocer las varianzas genéticas y del error, cosa que usualmente se desconoce. Como forma de solucionar esto se suele recurrir a métodos iterativos que ya discutimos en capítulos previos y que permiten la resolución iterativa de los dos problemas a la vez.

Finalmente, desde el punto de vista conceptual, el **BLUP** posee conexión directa con dos temas que trabajamos en este capítulo. Por un lado, la integración de la información de parientes se realiza en forma «automática», en lugar de considerar cada situación particular (de combinación de distintas fuentes de información). Por el otro, el método se extiende fácilmente a la predicción de los valores de cría en diferentes características, Henderson y Quaas (1976), «apilando» los vectores de características fenotípicas y usando el producto Kronecker entre la matriz de parentesco y la matriz de varianza-covarianza genética entre características.

Ejemplo 13.4

Supongamos que contamos con la información de peso al destete de 6 animales pertenecientes a 2 rodeos:

Animal	Padre	Madre	Rodeo	Peso al destete
1			1	185
2			1	180
3			2	175
4	1	2	1	188
5	1	3	2	182
6	1	3	2	187

El modelo que describe las observaciones es

$$y_{ij} = r_i + a_j + e_{ij}$$

donde:

- y_{ij} = observación (peso al destete);
- r_i = efecto fijo (rodeo);
- a_j = efecto aleatorio (animal);
- e_{ij} = efecto aleatorio (error)

Si utilizamos la **notación matricial**, el modelo que describe las observaciones es

$$\begin{aligned} \mathbf{y} &= \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{G}), \mathbf{G} = \mathbf{A}\sigma_A^2 \\ \mathbf{e} &\sim N(\mathbf{0}, \mathbf{R}), \mathbf{R} = \mathbf{I}\sigma_E^2, cov(\mathbf{u}, \mathbf{e}') = 0 \end{aligned}$$

donde:

- **y** = vector de observaciones (peso al destete) $n \times 1$, siendo n el número de observaciones:

$$\mathbf{y} = \begin{bmatrix} 185 \\ 180 \\ 175 \\ 188 \\ 182 \\ 187 \end{bmatrix}$$

- **b** = vector de efecto fijo (rodeo) $p \times 1$, siendo p el número de niveles de efecto fijo;
- **u** = vector de efecto aleatorio (animal) $q \times 1$, siendo q el número de niveles de efecto aleatorio;
- **e** = vector de efecto aleatorio (error) $n \times 1$, siendo n el número de observaciones;
- **X** = matriz de incidencia de efecto fijo (rodeo) $n \times p$:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

En esta matriz, la primera columna representa la media (ya que vamos a trabajar nuestros resultados como desvío de la media), mientras que la segunda columna representa el rodeo. Como en este tipo de modelos debemos especificar la pertenencia a un efecto fijo u otro con ceros y unos (**cero** quiere decir que **no tiene** ese nivel del efecto y **uno** que **sí lo tiene**), en general vamos a precisar de tantas columnas para cada efecto como niveles menos uno (eso porque uno de los niveles podemos incluirlo en la media). En nuestro caso, como tenemos solamente dos rodeos, alcanza con una sola columna para indicar a que rodeo pertenece. Supongamos que usamos la columna para especificar el nivel «rodeo 2». Por lo tanto, la tercera, quinta y sexta observación pertenecen al rodeo 2 (por eso hay un 1 en dichas observaciones); por lo que la primera, segunda y cuarta observación pertenecen al rodeo 1 (y por lo tanto hay un 0 en la columna «rodeo 2»).

- **Z** = matriz de incidencia de efecto aleatorio (animal) $n \times q$:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

La primera columna indica que la primera observación pertenece al animal 1; la segunda columna indica que la segunda observación pertenece al animal 2; la tercera columna indica que la tercera observación pertenece al animal 3; la cuarta columna indica que la cuarta observación pertenece al animal 4; la quinta columna indica que la quinta observación pertenece al animal 5; la sexta columna indica que la sexta observación pertenece al animal 6.

- $\mathbf{A}$ = matriz de parentesco aditivo entre los animales $n \times n$:

$$\mathbf{A} = \begin{bmatrix} 1{,}0 & 0{,}0 & 0{,}0 & 0{,}50 & 0{,}50 & 0{,}50 \\ 0{,}0 & 1{,}0 & 0{,}0 & 0{,}50 & 0{,}00 & 0{,}00 \\ 0{,}0 & 0{,}0 & 1{,}0 & 0{,}00 & 0{,}50 & 0{,}50 \\ 0{,}5 & 0{,}5 & 0{,}0 & 1{,}00 & 0{,}25 & 0{,}25 \\ 0{,}5 & 0{,}0 & 0{,}5 & 0{,}25 & 1{,}00 & 0{,}50 \\ 0{,}5 & 0{,}0 & 0{,}5 & 0{,}25 & 0{,}50 & 1{,}00 \end{bmatrix}$$

En esta matriz la primera columna indica que el primer animal es padre de los animales 4, 5 y 6 (por eso el valor de $0{,}5 = \frac{1}{2}$, el valor correspondiente al parentesco aditivo entre un progenitor y su progenie si no hay consanguinidad en el progenitor); la segunda columna indica que el segundo animal es padre del animal 4; la tercer columna indica que el tercer animal es padre de los animales 5 y 6; la cuarta columna indica que el cuarto animal es hijo de los animales 1 y 2, y medio hermano de los animales 5 y 6; la quinta columna indica que el quinto animal es hijo de los animales 1 y 3, medio hermano del animal 4, y hermano entero del animal 6; la sexta columna indica que el sexto animal es hijo de los animales 1 y 3, medio hermano del animal 4, y hermano entero del animal 5.

La diagonal indica el parentesco aditivo de cada animal consigo mismo: ninguno de los seis animales es consanguíneo.

Siguiendo el sistema de ecuaciones propuesto por Charles Roy Henderson (1949, 1950, 1959), obtenemos las estimaciones de los efectos fijos y aleatorios del modelo mixto. Primero, con todas la matrices y vectores que especificamos previamente armamos las matrices del lado izquierdo y del lado derecho del sistema de ecuaciones:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \\ = \begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1}$$

Luego, asumiendo que podemos invertir directamente la matriz del lado izquierdo de las ecuaciones BLUP, de acuerdo con las ecuaciones en (13.100), tenemos que la solución está dada por

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix} \therefore \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

lo que en nuestro caso nos deja con la siguiente solución:

$$\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix} 183{,}7098802 \\ -2{,}5482036 \\ 1{,}3952096 \\ -0{,}5149701 \\ -0{,}8802395 \\ 0{,}9901198 \\ 0{,}3404619 \\ 1{,}0547476 \end{bmatrix}$$

Como nuestro modelo tenía dos efectos fijos, la media y el desvío del segundo rodeo respecto al primero, la primera fila indica la media general, mientras que la segunda fila nos indicará la media del rodeo 2 (como desvío del «rodeo 1»). A partir de acá, el resto de las filas representarán los «niveles» de los efectos aleatorios: la tercer fila indica el valor de cría del animal 1; la cuarta fila indica el valor de cría del animal 2; la quinta fila indica el valor de cría del animal 3; la sexta fila indica el valor de cría del animal 4; la séptima fila indica el valor de cría del animal 5; la octava fila indica el valor de cría del animal 6.

PARA RECORDAR

- Un **modelo mixto lineal** es un **modelo mixto que combina efectos fijos y efectos aleatorios:**

1) **Efectos fijos** son aquellos que nos puede interesar contrastar sus niveles, que usualmente tienen pocos niveles y que en otros casos usaremos para compensar diferencias sistemáticas entre niveles. Normalmente, queremos **estimar** su valor.
2) **Efectos aleatorios** son efectos donde asumimos que los datos observados son una muestra aleatoria de una población mayor. Normalmente poseen muchos niveles y en general, el objetivo es **predecir** su valor ya que se trata de un efecto aleatorio.

- El **método BLUP** (Best Linear Unbiased Predictor) implica que si poseemos información completa para n animales (incluyendo su pedigrí completo), que serán los que pretendemos evaluar, los efectos fijos serán codificados en una matriz $\mathbf{X}$, que tendrá tantas filas como individuos con fenotipo, es decir n y tantas columnas como el número total de niveles de los factores considerados, menos el número de factores (para eliminar la dependencia lineal entre columnas). También en la matriz $\mathbf{X}$ se podrá incluir las covariables (variables continuas) que sean necesarias. Por otra parte, se considerarán por simplicidad solamente dos efectos aleatorios, el genético-aditivo de los individuos y el error. Teniendo en cuenta la matriz $\mathbf{A}$ de parentesco aditivo entre los individuos (que si no consideramos a los progenitores de los individuos con registros será de dimensiones $[n \times n]$). De tal manera que podemos describir el modelo mixto de forma matricial como

$$\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}\end{bmatrix} \qquad (13.101)$$

13.7 Conclusión

En este capítulo hemos ahondado nuestros conocimientos sobre la metodología empleada para llevar a cabo el proceso de selección artificial de forma efectiva. En particular se ha tenido en cuenta el problema planteado en el capítulo anterior, la existencia de respuestas correlacionadas entre caracteres fenotípicos. La existencia de dicha correlación plantea cuestiones estratégicas al momento de seleccionar: ¿el orden de selección altera el producto obtenido? ¿Cómo podemos realizar el proceso de la forma más eficiente? Además de algunos métodos intuitivos (por ejemplo, selección en tándem), hemos visto también cómo la aplicación de algunos resultados que pueden parecer más abstractos y matemáticamente complejos (como el teorema de Robertson-Price) pueden ser de suma utilidad para plantear una aproximación cuantitativa a estos problemas. Nuevamente la aproximación a través del modelado y la estadística resultan útiles en nuestro proceso de intentar entender y moldear la diversidad genética que observamos en la naturaleza, un mensaje central que es transversal a lo largo de las tres partes de este libro.

13.8 Actividades

13.8.1 Control de lectura

1. Define objetivos y criterio de selección. ¿Cuál es la diferencia entre ambos conceptos?
2. Agregar información fenotípica de varios individuos genéticamente relacionados permite realizar estimaciones más precisas del valor de cría de un individuo comparado a utilizar solo la información fenotípica de este. ¿A qué se debe esta mejora en la estimación?
3. ¿En qué casos es necesaria la selección por parientes? ¿A qué se debe esta necesidad?
4. ¿Cuándo es mejor la prueba de progenie que la selección fenotípica individual?
5. ¿Qué métodos de selección multicaracter conoces? ¿Qué ventajas presenta cada uno?

13.8.2 ¿Verdadero o falso?

1. La prueba de progenie siempre brinda mejores resultados que la selección individual.
2. Con tamaño finito de progenies es necesario considerar el balance existente entre el número de candidatos a evaluar y el número de registros por candidato.
3. La precisión del valor de cría de una característica siempre es una función de la heredabilidad de la misma.
4. Los valores de cría (EBV) combinan varias características fenotípicas de un individuo a la vez.
5. Los índices de selección ponderan según el valor económico de cada característica contemplada.

Soluciones

1. **Falso**. Si bien las pruebas de progenie brindan la ventaja de que nos permiten obtener una mejor precisión al momento de hacer estimados, conllevan la desventaja de incrementar considerablemente el intervalo generacional. Esto puede implicar una disminución en el progreso genético anual obtenido. Por lo tanto, es necesario calcular los coeficientes de eficiencia a efectos de determinar qué método brinda un mayor progreso genético anual.
2. **Verdadero**. Trabajar con tamaños de progenie finitos implica que al aumentar el número de registros (hijos medidos) por candidato se disminuye el número de candidatos a evaluar. Esto disminuye la capacidad de ejercer presión selectiva, en tanto se seleccionan menos individuos. Por lo tanto, es necesario optimizar estas relaciones para obtener el mayor progreso genético posible.
3. **Verdadero**. Como vimos en casos sencillos, por ejemplo con un solo registro del mismo individuo en el que la precisión es la raíz cuadrada de la heredabilidad, la precisión siempre es una función de la heredabilidad. Desde un punto de vista más intuitivo, en el caso de que la heredabilidad sea cero, la estimación a partir del fenotipo no contiene base genética y por lo tanto es completamente al azar, sin base genética (se basa en efectos ambientales) y tiene precisión de cero; si en cambio, la heredabilidad es de uno la estimación a partir de fenotipo se ve totalmente alineada con el valor de cría «verdadero», y por tanto la precisión es total.

4. **Falso.** El valor de cría de un individuo refiere siempre a una única característica fenotípica de interés. La estimación del mismo se realiza con datos fenotípicos de uno o varios individuos genéticamente relacionados, referentes a este carácter de interés.
5. **Verdadero.** En los índices de selección cada carácter o fuente de información (criterio de selección) considerado es ponderado a través de un coeficiente parcial de regresión que toma en cuenta los parámetros genéticos intervinientes (varianzas fenotípicas, heredabilidades, correlaciones genéticas y fenotípicas, y valor económico de aquellas características incluidas en el objetivo). Este coeficiente parcial de regresión define la importancia relativa de cada fuente de información. Para estimar el índice de selección se reúne la información de distintas fuentes y para distintos caracteres, finalmente logrando un solo valor para cada candidato. Los individuos son posteriormente ordenados según el valor de su índice, para luego elegir a los de mayor puntaje, como si se tratara de una sola característica. (Adaptado del *Manual de Práctica de Zootecnia*, 2014).

13.8.3 Ejercicios

Ejercicio 13.1

Se desea seleccionar 3 toros con el fin de mejorar el contenido de proteína de la leche en ganado jersey y con este objetivo se plantea realizar una prueba de progenie utilizando hembras medias hermanas entre sí. Con este fin se manejan tres alternativas: probar 20, 30 o 50 candidatos. Sabiendo que la capacidad de medición es de 3000 hembras y que la heredabilidad de la característica es de 0,31, ¿con qué alternativa se lograría el mayor progreso genético?

Solución

En primer lugar, debemos recordar que un aumento en el número de candidatos a probar implica una disminución en la proporción de los animales seleccionados lo que significa un aumento en la intensidad de selección. Sin embargo, al mismo tiempo, este aumento en el número de candidatos a probar implica una disminución del número de individuos medidos por candidato (menor n), por lo que no existe una respuesta evidente en cuanto a qué opción es más conveniente. Entonces, dado que ni el IG ni la varianza aditiva varían entre alternativas, deberemos evaluar cuál relación $i \times r_{AC}$ es mayor, lo que implicará un mayor progreso genético.

Candidatos evaluados	p	i	n	r_{AC}	$i \times r_{AC}$
20	$3/20 = 0{,}15$	1,5544	$3000/20 = 150$	0,963	1,497
30	$3/30 = 0{,}10$	1,7550	$3000/30 = 100$	0,945	1,469
50	$3/50 = 0{,}06$	1,9854	$3000/50 = 60$	0,913	1,419

Como podemos notar, la alternativa que nos permite lograr el mayor progreso genético es evaluar 20 candidatos. Es decir, en este caso en particular, el aumento en la intensidad de selección, producto de probar más candidatos, no compensa la pérdida de exactitud por evaluar menos hijas por candidato.

Ejercicio 13.2

Un mago posee una población de conejos y tiene como objetivo mejorar el peso adulto ($h^2 = 0{,}38$) de estos, ya que si son muy pequeños es más difícil sacarlos de la galera. La media poblacional actual es de 4,1 kg, y para mejorarla en la próxima generación el mago ha evaluado dos animales como reproductores:

- Conejo Cartas: el peso promedio de 7 medios hermanos es 4,3 kg
- Conejo Varita: el peso promedio de 3 hermanos enteros es 4,4 kg

¿Qué animal seleccionarías como reproductor?

Solución

Para evaluar a estos animales como reproductores, deberemos calcular el valor de cría con cada uno con su correspondiente precisión.

$$\hat{A}_{cartas} = \frac{r_{AA'} \times n \times h^2}{1 + (n-1) \times t} \times (\bar{x}_i - \mu) = \frac{\frac{1}{4} \times 7 \times 0{,}38}{1 + (7-1) \times \frac{1}{4} \times 0{,}38} \times (4{,}3 - 4{,}1) = 0{,}08\text{kg}$$

$$r_{AC_{cartas}} = r_{AA'} \times h \times \sqrt{\frac{n}{1 + (n-1) \times t}} = \frac{1}{4} \times \sqrt{0{,}38} \times \sqrt{\frac{7}{1 + (7-1) \times \frac{1}{4} \times 0{,}38}} = 0{,}325$$

$$\hat{A}_{varita} = \frac{r_{AA'} \times n \times h^2}{1 + (n-1) \times t} \times (\bar{x}_i - \mu) = \frac{\frac{1}{2} \times 3 \times 0{,}38}{1 + (3-1) \times \frac{1}{2} \times 0{,}38} \times (4{,}4 - 4{,}1) = 0{,}12\text{kg}$$

$$r_{AC_{varita}} = r_{AA'} \times h \times \sqrt{\frac{n}{1 + (n-1) \times t}} = \frac{1}{2} \times \sqrt{0{,}38} \times \sqrt{\frac{3}{1 + (3-1) \times \frac{1}{2} \times 0{,}38}} = 0{,}454$$

El valor de cría del conejo varita es superior, al igual que la precisión del mismo, por lo tanto será el animal seleccionado por el mago para utilizar como reproductor.

Ejercicio 13.3

En una población de bovinos se tiene como objetivo mejorar el tamaño del músculo del lomo ya que el frigorífico estaría dispuesto a pagar un sobreprecio en los animales que muestren un mayor tamaño del mismo. Dado que es una característica cuya medición implica el sacrificio de los animales, se propone medir a 18 medios hermanos (medios hermanos entre sí) de los candidatos a la selección. Se ha determinado a la vez que la $h^2 = 0{,}33$ y que la varianza aditiva de la característica es $\sigma_A^2 = 0{,}8$. Por otra parte, la proporción de machos seleccionados es del 20 % mientras que no se realiza selección en las hembras.

¿Cuál es el progreso genético generacional logrado en estas condiciones?

Solución

Recordemos en primer lugar la fórmula de progreso genético generacional:

$$\text{PGG} = i \times r_{AC} \times \sigma_A$$

En primer lugar, podremos despejar la intensidad de selección a partir de la información de la proporción de animales seleccionados:

$$p = 20\,\% \implies i = 1{,}3998$$

$$i = \frac{i_m + i_h}{2} = \frac{1{,}3998 + 0}{2} = 0{,}6999$$

A continuación deberemos calcular la exactitud r_{AC} y el desvío aditivo:

$$r_{AC} = r_{AA'} \times h \times \sqrt{\frac{n}{1+(n-1)\times t}} = \frac{1}{4} \times \sqrt{0{,}33} \times \sqrt{\frac{18}{1+(18-1)\times \frac{1}{4} \times 0{,}33}} = 0{,}3931$$

$$\sigma_A = \sqrt{\sigma_A^2} = \sqrt{0{,}8} = 0{,}8944$$

Por último podremos aplicar la fórmula de PGG:

$$\text{PGG} = 0{,}6999 \times 0{,}3931 \times 0{,}8944 = 0{,}246$$

Es decir que por generación, se podrá aumentar en 0,246 kg el peso del músculo del lomo.

Ejercicio 13.4

(Adaptado de Miglior, Muir, and Van Doormaal 2005)

En una revisión del año 2005 Miglior y colaboradores (Miglior, Muir, and Van Doormaal 2005) analizaron los índices de selección de utilizados para producción con ganado bovino de la raza holstein de quince países elegidos por posición geográfica, tamaños de las poblaciones productivas consideradas y otros aspectos de relevancia

A continuación se muestra una selección de algunas de estas características en tres países (Dinamarca, Estados Unidos e Israel)

	Índice	Prod. leche	Prod. grasa	Prod. proteína	Longevidad	Tamaño corporal	Fertilidad
Dinamarca	S-index	−8,4	1,2	2,4	6,0	−2,0	9,0
Estados Unidos	TPI/Net Merit	−6,0	18,0	36,0	4,0	−3,0	7,0
Israel	PD01	−11,0	18,0	51,0	11,0	−2,5	9,0

Nota: la fertilidad constituye un índice en sí (índice de fertilidad), el cual pondera variables relacionadas a la capacidad de dejar progenie de una vaca.

a. Sabiendo que un individuo que presenta las siguientes características

	Prod. leche	Prod. grasa	Prod. proteína	Longevidad	Tamaño corporal	Fertilidad
Vaca_1	27,5 L/día	1,17 kg/día	0,98 kg/día	18 años	615 kg	6

Calcula el índice de selección en cada país.

b. ¿A qué atribuye las diferencias observadas? ¿Esperaría que un índice de selección uruguayo sea igual a alguno de los presentados?

Solución

a. Para calcular el índice de selección en cada país realizaremos una combinación lineal de los valores presentados para la Vaca_1 y los pesos dados en cada índice de selección según la primera tabla. Es decir, para calcular el índice de selección operamos según

$$I = w_1 \times x_1 + w_2 \times x_2 + \ldots + w_n \times x_n I$$

donde w_i son los pesos dados y x_i son los valores fenotípicos de la vaca.

Tenemos entonces

$$\begin{aligned} I_{\text{Dinamarca}} =& (-8{,}4 \times 27{,}5) + (1{,}2 \times 1{,}17) + (2{,}4 \times 0{,}98) \\ &+ (6{,}0 \times 18) + (-2{,}0 \times 615) + (9{,}0 \times 6) = -1295{,}244 \end{aligned}$$

$$\begin{aligned} I_{\text{Estados Unidos}} =& (-6{,}0 \times 27{,}5) + (18{,}0 \times 1{,}17) + (36{,}0 \times 0{,}98) \\ &+ (4{,}0 \times 18) + (-3{,}0 \times 615) + (7{,}0 \times 6) = -1839{,}66 \end{aligned}$$

$$\begin{aligned} I_{\text{Israel}} =& (-11{,}0 \times 27{,}5) + (18{,}0 \times 1{,}17) + (51{,}0 \times 0{,}98) \\ &+ (11{,}0 \times 18) + (-2{,}5 \times 615) + (9{,}0 \times 6) = -1516{,}96 \end{aligned}$$

Aclaración: Tener presente que, a efectos de simplificar el ejercicio, no se utilizaron todas las componentes contempladas en cada índice de selección; esto puede explicar el por qué se obtuvieron valores negativos.

b. Notemos que existen diferencias sustanciales en los valores de índice de selección: la diferencia entre el valor obtenido según el índice estadounidense y el danés es de ≈ 600, que es cerca de la mitad del valor estimado con este último. Es importante recordar que cada índice pondera con base en las necesidades que plantea el mercado de cada país, por lo que es de esperar que lugares con culturas y mercados tan disímiles den importancias diferentes a iguales características fenotípicas. En principio es razonable esperar, por lo tanto, que un índice uruguayo también muestre diferencias. Además, debe contemplarse que este estudio es de hace veinte años (al momento de publicación de este libro), por lo que incluso los índices actuales empleados en cada país pueden variar actualmente.

Capítulo 14

Endocría, exocría, consanguinidad y depresión endogámica

Sans être tout à fait un imbécile fini
Je n'ai rien du penseur, du phénix, du génie
Mais je n' suis pas le mauvais bougre et j'ai bon cœur
Et ça compense à la rigueur.[1]

Georges Brassens

Hasta el momento, cuando nos manejamos a nivel de población generalmente supusimos y hemos aceptado como natural que la población se encontraba en equilibrio de Hardy-Weinberg, ignorando los efectos de un apartamiento sistemático de este equilibrio. Sin embargo, en poblaciones de tamaños reducidos, la probabilidad de apareamiento entre parientes se incrementa y por lo tanto aumenta la probabilidad de que dos alelos en un locus sean idénticos por ascendencia (IPA). Este fenómeno está directamente ligado a la disminución relativa de heterocigotos y el aumento de los homocigotos, lo que veremos más adelante se encuentra asociado al fenómeno de la **depresión endogámica**.

Más aún, el fenómeno puede verse en el sentido opuesto, es decir cuando combinamos dos poblaciones que son diferentes y, por lo tanto, como demostraremos en este capítulo, se produce un incremento de los heterocigotos y bajo las misma condiciones de la arquitectura genética de la característica en las que se desarrolla la depresión endogámica, ocurrirá el fenómeno de la **heterosis**. Se trata de un fenómeno que el hombre ha observado por siglos y que se conoce vulgarmente como **vigor híbrido**, es decir, el incremento en una característica de interés al combinar dos poblaciones o razas que han divergido en el tiempo.

The observations as I received them are shown in columns II. and III., where they certainly have no primâ facie *appearance of regularity. But as soon as we arrange them the in order of their magnitudes, as in columns IV. and V., the case is materially altered. We now see, with few exceptions, that the largest plant on the crossed side in each pot exceeds the largest plant on the self-fertilised side, that the second exceeds the second, the third the third, and so on. Out of the fifteen cases in the table, there are only two exceptions to this rule.*

We may therefore confidently affirm that a crossed series will always be found to exceed a self-fertilised series, within the range of the conditions under which the present experiment has been made.

[1] Parte de la canción *Quand les cons sont braves* (que podría traducirse, libremente, como *Cuando los tontos son valientes*). La traducción de este fragmento según ChatGPT, para que mantenga la rima podría ser: «Sin ser del todo un tonto consumado, no tengo nada de sabio, de genio consagrado; pero tampoco soy mal tipo, tengo buen corazón, y eso compensa... más o menos, según la ocasión». En este capítulo que vamos a hablar de endogamia, si ser tonto tuviese una causa genética y fuese recesivo, seguramente los tontos se incrementarían por la endogamia, para gran disgusto de *Tonton Georges*.

Sir Francis Galton a **Charles Darwin**, en referencia a un ensayo con una población de plantas de maíz, entre otras varias especies ensayadas (citado en Darwin 1876)

En el presente capítulo, por lo tanto, vamos a trabajar con los fenómenos que ocurren en poblaciones de tamaño finito y cómo eso afecta las características fenotípicas de interés. En el comienzo estableceremos la relación que existe entre el tamaño poblacional y el incremento en la consanguinidad esperada, algo que ya vimos en el capítulo «Apareamientos no-aleatorios», pero que veremos desde otra óptica, relacionaremos con otros parámetros de interés, como el número de machos y hembras en la población, o como el intervalo generacional. A continuación definiremos el concepto de depresión endogámica, veremos sus causas desde el punto de vista de los mecanismos genéticos que la producen y llegaremos a un modelo matemático muy sencillo para modelarla.

En la segunda parte del capítulo veremos el fenómeno recíproco de la heterosis y su relación con la exocría, el apareamiento de individuos no emparentados (o al menos con menor parentesco que el promedio de la población total). Veremos cómo se conecta todo con la endocría, ya que las fuerzas que juegan son nuevamente la dominancia y abundancia relativa de los heterocigotos. Trabajaremos con un modelo gráfico que nos simplificará la comprensión de las bases genéticas de diferentes cruzamientos y finalmente desarrollaremos un modelo genético formal de cruzamientos, que nos permitirá describir cada tipo de cruzamiento y estimar los parámetros de relevancia.

OBJETIVOS DEL CAPÍTULO

☐ Presentar el concepto de consanguinidad y explorar su relación con el tamaño poblacional.

☐ Estudiar el concepto de depresión endogámica, planteando un modelo matemático sencillo para su estudio.

☐ Investigar el fenómeno de la heterosis y su relación con la exocría.

☐ Estudiar la redistribución en la varianza genética producto de la depresión endogámica.

☐ Analizar la importancia de mantener una proporción adecuada de machos y hembras en la población para preservar un tamaño efectivo poblacional óptimo, minimizando los efectos de la consanguinidad.

14.1 El aumento de la consanguinidad a partir del número de individuos

En una población algo idealizada en la que los reproductores colocan sus gametos en un pool, resulta evidente que a medida que el número de reproductores disminuye, la probabilidad de que sus hijos sean parientes se incrementa, lo que en la siguiente generación llevará ineluctablemente a la aparición de consanguinidad. La pregunta, por lo tanto, es cómo será la relación entre el incremento de la consanguinidad entre generaciones, a partir del número de individuos reproductores. Para eso, supondremos una población algo idealizada (figura 14.1), donde las generaciones son no-solapantes y a partir de la misma vamos a calcular el incremento en la consanguinidad (ΔF) en una generación, que corresponde al número de reproductores usados, pero asumiendo que el apareamiento es al azar. A efecto de realizar los cálculos, vamos a suponer una generación de base (F I), en la que tenemos M machos y H hembras, que suponemos no emparentados y analizaremos lo que ocurre a partir de la primera generación en la que se presenta consanguinidad. El enfoque se basa en la esperanza de consanguinidad en un individuo diploide, formado de la unión al azar de los gametos dentro del pool de sus progenitores.

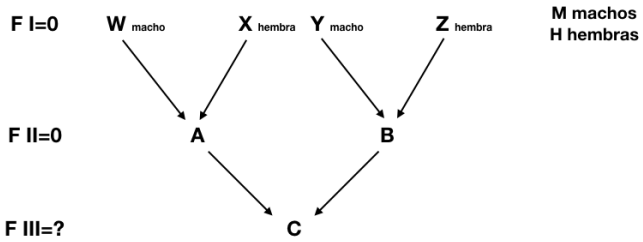


Figura 14.1. Esquema para inferir ΔF en una población a partir del número total de machos y hembras. Para que exista consanguinidad en los individuos de la tercera generación (C), sus padres (A y B) deben ser parientes y eso solo puede ocurrir si alguno de los abuelos (W e Y) o abuelas (X y Z) de C son el mismo individuo.

De acuerdo con la figura 14.1, para que C tenga consanguinidad, sus padres A y B deben de ser parientes, por lo que, si no conocemos más información que la de los padres de estos últimos (W_{macho},X_{hembra},Y_{macho},Z_{hembra}), las dos únicas posibilidades son de que o los abuelos de C son el mismo individuo (W=Y) o las abuelas son el mismo individuo (X=Z). De hecho, pueden ocurrir las dos cosas al mismo tiempo, aunque esta probabilidad es muy pequeña. En cualquiera de los dos casos en que el abuelo es el mismo o la abuela es la misma ($W = Y|X = Z$), el parentesco entre A y B es de medios hermanos, por lo que el parentesco aditivo es de 1/4. Por lo tanto, la consanguinidad en C, $F_C = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$. En el caso en que ocurran los dos eventos al mismo tiempo ($W = Y \ \& \ X = Z$), ahora A y B serán hermanos enteros, por lo que el parentesco aditivo entre ellos será de 1/2 y la consanguinidad en C, $F_C = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Pero ¿cuál es la probabilidad de cada uno de estos eventos? Supongamos que existen M machos distintos y H hembras distintas en la generación de los abuelos. La probabilidad de que por azar, dentro del pool de gametos producidos por los machos, dos tomados al azar

provengan del mismo macho es $\frac{1}{M}$ y por lo tanto la probabilidad de que no provengan del mismo es $\left(1-\frac{1}{M}\right)$. Idéntico razonamiento se aplica a las hembras, pero ahora con H en lugar de M. Por lo tanto, la probabilidad de que los padres (machos) de A y B sean el mismo, pero no las madres, es $\frac{1}{M}\left(1-\frac{1}{H}\right)$. Del mismo modo, la probabilidad de que las madres de A y B, pero no los padres, sean las mismas es $\frac{1}{H}\left(1-\frac{1}{M}\right)$. Finalmente, la probabilidad de que el padre de A y B sea el mismo macho y al mismo tiempo la madre sea la misma hembra es $\frac{1}{M}\frac{1}{H}$. Poniendo todo junto,

$$\begin{aligned}\Delta F_G = \frac{1}{8}\left(\frac{1}{M}\times\left(1-\frac{1}{H}\right)+\frac{1}{H}\times\left(1-\frac{1}{M}\right)\right)+\frac{1}{4}\left(\frac{1}{M}\frac{1}{H}\right) = \\ \frac{1}{8}\left(\frac{1}{M}\times\left(1-\frac{1}{H}\right)+\frac{1}{H}\times\left(1-\frac{1}{M}\right)+2\frac{1}{M}\frac{1}{H}\right) = \\ \frac{1}{8}\left(\frac{1}{M}-\frac{1}{MH}+\frac{1}{H}-\frac{1}{MH}+\frac{2}{MH}\right)=\frac{1}{8}\left(\frac{1}{M}+\frac{1}{H}\right)\end{aligned} \tag{14.1}$$

Es decir, de acuerdo con este modelo algo idealizado, donde las generaciones no se solapan, con M machos y H hembras en total,

$$\Delta F_G = \frac{1}{8}\left(\frac{1}{M}+\frac{1}{H}\right)=\frac{1}{8}\left(\frac{M+H}{MH}\right) \tag{14.2}$$

Más aún, en el caso de una población ideal, podemos asumir que el número de machos es igual al número de hembras, o sea $N/2$, con N igual al número de individuos reproductivos totales. Por lo tanto,

$$\Delta F_G = \frac{1}{8}\left(\frac{2}{N}+\frac{2}{N}\right)=\frac{1}{8}\left(\frac{4}{N}\right)=\frac{1}{2N} \tag{14.3}$$

Sin embargo, en la mayor parte de las poblaciones de interés productivo, el número de machos es claramente diferente al número de hembras, usualmente mucho menor ya que un macho es capaz de aparearse con varias hembras por estación reproductiva. La pregunta entonces que nos podemos formular es, dada una combinación con distinto número de machos que de hembras, a qué tamaño de población ideal le correspondería el mismo incremento en la consanguinidad. Para eso, igualando el ΔF_G, pero ahora con N_e en lugar de N, obtenemos

$$\begin{aligned}\Delta F_G = \frac{1}{8}\left(\frac{M+H}{MH}\right)=\frac{1}{2N_e} \\ \therefore N_e = \frac{4MH}{M+H}\end{aligned} \tag{14.4}$$

Más allá de la utilidad innegable de conocer el aumento en la consanguidad esperable en una generación, desde el punto de vista práctico (al menos en la producción animal) resulta mucho más importante conocer la tasa anual de incremento. La transformación de una en otra resulta elemental, nuevamente dividiendo entre el intervalo generacional de la población,

$$\Delta F_a = \frac{\Delta F_G}{IG} = \frac{\frac{1}{2N_e}}{IG} = \frac{M+H}{8\,M\,H\,IG} \tag{14.5}$$

Pero nuestros desarrollos basados en M y H contados como el número total de machos y hembras funcionan cuando las generaciones son no-solapantes, es decir cuando la generación de los padres deja de reproducirse cuando los hijos entran al pool reproductivo. La formulación alternativa de M y H para la situación en que las generaciones se superponen es $M = m \times IG$ y $H = h \times IG$, con m y h el número de machos y hembras que entran por año, y con IG el intervalo generacional de la población.

Sustituyendo en la ecuación anterior las nuevas definiciones de M y H, tenemos

$$\Delta F_a = \frac{M+H}{8\,M\,H\,IG} = \frac{m\,IG + h\,IG}{8\,(m\,IG)\,(h\,IG)\,IG} = \frac{(m+h)\,IG}{8\,m\,h\,IG^3} = \frac{m+h}{8\,m\,h\,IG^2} \tag{14.6}$$

En contextos como el de la mayoría de las producciones animales, el número de machos suele ser al menos de un orden menor que el de hembras por lo que $\frac{m+h}{h} = \frac{m}{h} + \frac{h}{h} \approx 0+1 = 1$ y una buena aproximación que simplifica las estimaciones de la tasa de incremento anual de la consanguinidad puede obtenerse a partir de

$$\Delta F_a = \frac{m+h}{8\,m\,h\,IG^2} = \frac{1}{8\,m\,IG^2}\left(\frac{m+h}{h}\right) \approx \frac{1}{8\,m\,IG^2} \tag{14.7}$$

Pero veamos qué significan estas ecuaciones en términos prácticos. En la figura 14.2 (a) se aprecian las curvas de nivel correspondientes a un mismo tamaño efectivo poblacional (N_e), dependiendo del número de machos y hembras utilizados. Obviamente, el gráfico es simétrico respecto a la diagonal 1-1, pese a que en la mayor parte de las situaciones reales nos encontraremos situados en la triangular superior izquierda de este gráfico (más hembras que machos). Tal vez el resultado más impactante visualmente es la casi verticalidad de las primeras curvas de nivel, aquellas que corresponden a tamaños efectivos poblacionales de 10, 50 y 100 individuos; una línea vertical como la azul, en 100 machos, por ejemplo, cuantas más curvas de nivel corte al subir, más crecerá el tamaño poblacional al subir el número de hembras. Sin embargo, lo que se aprecia en la región de pocos machos es que una línea vertical apenas cambiará el tamaño efectivo pese a aumentar enormemente el número de

hembras. En el ejemplo de esta figura, la línea azul correspondiente a 100 machos, corta la curva correspondiente a 300 individuos cuando las hembras son 300 (el total de animales es de 400), pero nunca corta la curva de 400 individuos (dentro del gráfico, al menos), pese a que en la parte superior se usan 1000 hembras. De hecho, si re-arreglamos la ecuación (14.4) de forma de obtener el número de hembras necesarias para llegar a un determinado N_e dado un número de machos, tenemos

$$N_e = \frac{4MH}{M+H} \therefore H = \frac{N_e\, M}{4M - N_e}$$
$$\lim_{N_e \to 4M} \frac{N_e\, M}{4M - N_e} = \infty \tag{14.8}$$

O sea, cuando el tamaño efectivo de la población se acerque a 4 veces el número de machos, el número de hembras tenderá a infinito, o dicho de otra forma, aún con un número infinito de hembras, el tamaño efectivo poblacional nunca llegará a 4 veces el número de machos, ¡algo verdaderamente impactante!

Otro resultado interesante consiste en conocer cuál es el número de individuos de un sexo dado que el tamaño efectivo poblacional es igual al número de individuos del otro sexo. Pongamos por caso que queremos conocer el número de machos tal que el tamaño efectivo poblacional es igual al número de hembras. Para eso, igualamos N_e a H y obtenemos

$$M = \frac{N_e\, H}{4H - N_e}$$
$$N_e = H \therefore M = \frac{H^2}{4H - H} = \frac{H}{3} \tag{14.9}$$

Es decir, cuando el número de machos es un tercio del número de hembras, el tamaño efectivo poblacional es igual al número de hembras directamente. Este resultado se puede además verificar visualmente en la figura 7.2(a), trazando una línea horizontal desde el número de hembras deseado (idealmente igual a algunas de las curvas de N_e), hasta que corte el correspondiente valor de N_e y ahí bajando verticalmente hasta encontrar el número de machos.

Ejemplo 14.1

El desequilibrio en el número de machos y hembras que entran cada año a la población es una causa importante de la reducción en su tamaño efectivo, en particular en poblaciones de animales domésticos. Para visualizar, de alguna manera, lo que esto significa en términos de rodeos/majadas típicas de nuestro país, en la siguiente tabla se presentan diferentes situaciones de número de machos y hembras que ingresan por año a la población, y su consecuente tamaño efectivo. En todos los casos se considera un intervalo generacional promedio de la población de 3 años:

Ingresos/año	Ingresos/año (machos) (m)	Ingresos/año (hembras)(h)	Tamaño poblacional	Tamaño efectivo poblacional (N_e)
36	1	35	108	12
30	5	25	90	50
40	5	35	120	53
80	5	75	240	56
45	10	35	135	93
55	20	35	165	153
100	1	99	300	12
100	50	50	300	300

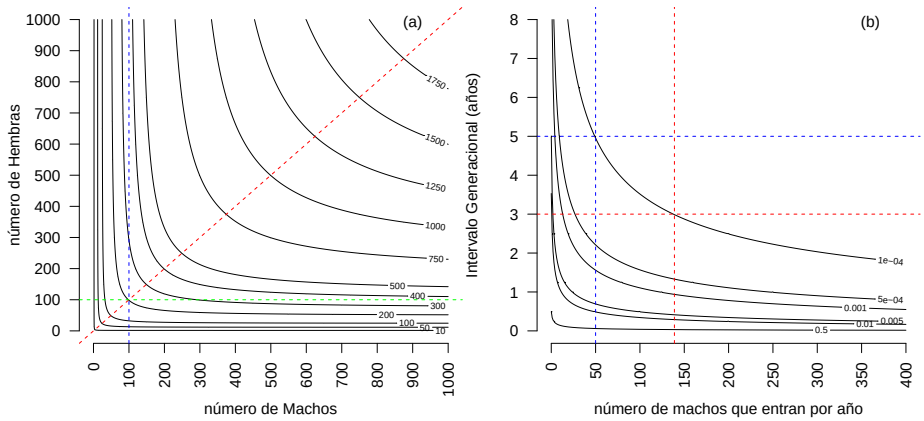


Figura 14.2. (a) Tamaño efectivo de la población de acuerdo con el número de machos y hembras de la misma ($M = m \times IG, H = h \times IG$). (b) Curvas de isoconsanguinidad de acuerdo con la aproximación basada en número de machos por año; pasar de un IG=3 a IG=5 permite reducir el número mínimo de machos por año de 138 a 50 para mantener $\Delta F_a \leq 1\text{x}10^{-04}$

Para calcular el valor de la última columna utilizamos la ecuación $N_e = \frac{4MH}{M+H}$, con $M = m\,\bar{IG}$ (es decir el número de machos que entran por año multiplicado por el intervalo generacional promedio de la población) y $H = h\,\bar{IG}$. El tamaño «real» de la población es igual a $M + H = m\,\bar{IG} + h\,\bar{IG} = (m + h)\,\bar{IG}$

PARA RECORDAR

- Si imaginamos que los reproductores colocan sus gametos en un pool, a **medida que el número de reproductores disminuye, la probabilidad de que sus hijos sean parientes se incrementa**, lo que en la siguiente generación llevará inevitablemente a la aparición de consanguinidad.
- Si imaginamos una **población con generaciones no solapantes** y dónde los apareamientos se realizan al azar, con M machos y H hembras **no emparentados**, la probabilidad de que por azar, dentro del pool de gametos producidos por los machos, dos tomados al azar, provengan del mismo macho es $\frac{1}{M}$ y por lo tanto la probabilidad de que no provengan del mismo es $\left(1 - \frac{1}{M}\right)$. Idéntico razonamiento se aplica a las hembras, pero ahora con H en lugar de M. Por lo tanto, la probabilidad de que los padres (machos) de A y B sean el mismo, pero no las madres, es $\frac{1}{M}\left(1 - \frac{1}{H}\right)$. Del mismo modo, la probabilidad de que las madres de A y B, pero no los padres, sean las mismas es $\frac{1}{H}\left(1 - \frac{1}{M}\right)$. Finalmente, la probabilidad de que el padre de A y B sea el mismo macho y al mismo tiempo la madre sea la misma hembra es $\frac{1}{M}\frac{1}{H}$. Entonces $\Delta F_G = \frac{1}{8}\left(\frac{1}{M} + \frac{1}{H}\right) = \frac{1}{8}\left(\frac{M+H}{MH}\right)$
- Si a su vez asumimos que esta población ideal posee el mismo número de machos que de hembras, o sea $N/2$, con N igual al número de individuos reproductivos totales: $\Delta F_G = \frac{1}{2N}$

- Si salimos de condiciones ideales y asumimos un menor número de machos que de hembras, el tamaño de población ideal que se correspondería el mismo incremento en la consanguinidad será $N_e = \frac{4MH}{M+H}$. Si además queremos calcular la **tasa anual de incremento** podemos: $\Delta F_a = \frac{\Delta F_G}{IG} = \frac{M+H}{8\ M\ H\ IG}$
- Si ahora además suponemos generaciones solapantes donde $M = m \times IG$ y $H = h \times IG$, siendo m yh el número de machos y hembras que entran por año: $\Delta F_a = \frac{m+h}{8\ m\ h\ IG^2}$. Si a su vez reconocemos que el número de machos tiende a ser mucho menor que el de las hembras con $(m+h)/h \approx 1$ entonces $\Delta F_a \approx \frac{1}{8\ m\ IG^2}$

¿Qué proporción de machos es necesaria para mantener un tamaño efectivo poblacional mínimo?

Supongamos ahora que con un número determinado de hembras queremos saber el número mínimo de machos para mantener N_e por encima de un valor. Podemos asumir sin pérdida de generalidad que el número de machos debe ser menor o igual al de hembras. Entonces

$$
\begin{aligned}
N_e &\geq \frac{4MH}{M+H} \\
&\text{Si } 0 \leq a \leq 1 \text{ y } M = a\,H \\
N_e &\geq \frac{4aH^2}{H(a+1)} = \frac{4aH}{(a+1)} \therefore \\
N_e(a+1) \geq 4aH &\iff 0 \leq \frac{N_e}{4H - N_e} \leq a \leq 1, \text{ pero} \\
4H \geq 2N_e &\iff H \geq \frac{N_e}{2}
\end{aligned}
\tag{14.10}
$$

Por ejemplo, si tenemos 300 hembras en el rodeo, pero no queremos que el tamaño efectivo del mismo sea menor a 325, basta con colocar esos números en la ecuación (14.10), para obtener a y luego el número de machos

$$
a \geq \frac{N_e}{4H - N_e} = \frac{325}{4 \times 300 - 325} = 0{,}3714286 \therefore M = 0{,}3714286 \times 300 \approx 112 \tag{14.11}
$$

La alternativa gráfica, aproximada, se encuentra en la figura 14.3.

Dado un N_e mínimo a mantener y un número de hembras, trazamos una línea vertical y otra horizontal en esos valores y los coeficientes permitidos son los que están en el cuadrante inferior derecho, o sea, los mayores o iguales a la primera línea (el menor a posible) que entra en este cuadrante.

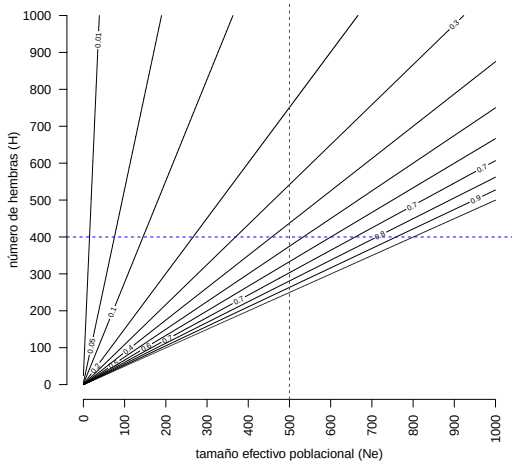


Figura 14.3. Proporción mínima de machos (respecto al número de hembras) necesaria para mantener un tamaño efectivo poblacional mínimo. Dado un N_e y un número de hembras (línea vertical y horizontal, respectivamente), las proporciones permitidas son las curvas que quedan en el cuadrante inferior derecho.

En un plano diferente, pero relacionado, en la figura 14.2 (b) podemos apreciar algunas curvas de idéntica tasa de incremento anual de la consanguinidad en función del número de machos que entran por año y del intervalo generacional de la población, de acuerdo con la ecuación (14.7). La primera observación es que para una consanguinidad dada y valores de intervalo generacional relativamente bajos (la parte derecha de las curvas) es relativamente fácil una reducción importante en el número de machos a usar, a un costo relativamente menor en el incremento del intervalo generacional. Sin embargo, a medida que nos vamos hacia la izquierda por la curva, cada vez se hace más importante el impacto en el intervalo generacional de una reducción en el número de machos. Por ejemplo, es posible mantener $\Delta F_a = 1\text{x}10^{-4}$ usando solo 50 machos en lugar de ≈ 138, pero al precio de incrementar el intervalo generacional de 3 a 5 años. Una reducción equivalente de 226 a 138 machos, hubiera costado solo 0,65 años de incremento en el IG

$$\begin{aligned} IG_{226Machos} &= \sqrt{\frac{1}{8\ \Delta F_a\ m}} = \sqrt{\frac{1}{8\ 1\text{x}10{-}4226}} \approx 2{,}35 \text{ años} \\ IG_{138Machos} &= \sqrt{\frac{1}{8\ 1\text{x}10{-}4\ 138}} \approx 3 \text{ años} \\ IG_{138Machos} - IG_{226Machos} &= 3 - 2{,}35 = 0{,}65 \text{ años} \end{aligned} \tag{14.12}$$

Ejemplo 14.2a

Como vimos, la consanguinidad es el resultado de la endocría un animal es consanguíneo únicamente si sus padres son parientes. El grado de consanguinidad que tenga un animal, depende del nivel de parentesco que tengan sus padres

$$F_Z = \frac{1}{2} \times a_{X,Y} \tag{14.13}$$

Donde: F_Z: coeficiente de consanguinidad del individuo Z; $a_{X,Y}$: parentesco aditivo entre X e Y, ambos padres de Z.

En la siguiente tabla se presentan diferentes situaciones de parentesco entre los padres y su consecuente coeficiente de consanguinidad del individuo:

Parentesco entre los padres	Coeficiente de consanguinidad del individuo
0	0
0,010	0,0050
0,015	0,0075
0,020	0,0100
0,030	0,0150
0,050	0,0250
0,100	0,0500
0,150	0,0750
0,200	0,1000
0,300	0,1500
0,500	0,2500
1	0,5000

Sus valores se encuentran dentro del rango entre 0 (sus padres no están emparentados) y 0,5 (sus padres son gemelos idénticos o clones).

Ejemplo 14.2b

(**Tasa de consanguinidad generacional**)

En la siguiente tabla, a modo de ejemplo, se presentan diferentes situaciones de número de machos y hembras que ingresan por año a la población, y su consecuente tasa de consanguinidad generacional, manteniendo en todas las situaciones, un intervalo generacional promedio de la población de 2 años:

Ingresos/año	Ingresos/año (machos) (m)	Ingresos/año (hembras) (h)	Tamaño efectivo (N_e)	Tasa de consanguinidad/gen.
36	1	35	8	0,0643
30	5	25	33	0,0150
40	5	35	35	0,0143
80	5	75	38	0,0133
45	10	35	62	0,0080
55	20	35	102	0,0049
100	1	99	8	0,0631
100	50	50	200	0,0025

Ejemplo 14.2c

En la siguiente tabla, a modo de ejemplo, se presentan diferentes situaciones de número de machos y hembras que ingresan por año a la población, y su consecuente tasa de consanguinidad anual, manteniendo en todas las situaciones, un intervalo generacional promedio de la población de 2 años:

Ingresos/año	Ingresos/año (machos) (m)	Ingresos/año (hembras) (h)	Tamaño efectivo (N_e)	Tasa consanguinidad/gen.	Tasa consanguinidad/año
36	1	35	8	0,0643	0,0321
30	5	25	33	0,0150	0,0075
40	5	35	35	0,0143	0,0071
80	5	75	38	0,0133	0,0067
45	10	35	62	0,0080	0,0040
55	20	35	102	0,0049	0,0025
100	1	99	8	0,0631	0,0316
100	50	50	200	0,0025	0,0013

Ejemplo 14.2d

En la siguiente tabla se presentan diferentes situaciones de tasa de incremento en la consanguinidad anual, y su consecuente número de machos necesarios que ingresen por año a la población, manteniendo en todas las situaciones, un intervalo generacional promedio de la población de 4,5 años:

Tasa de consanguinidad anual	Número de machos que ingresan por año a la población (m)
0,0050	$1{,}2 \simeq 2$
0,0020	$3{,}1 \simeq 4$
0,0010	$6{,}2 \simeq 7$
0,0005	$12{,}3 \simeq 14$

PARA RECORDAR

- Si quisiéramos determinar, dado un cierto número de hembras, cuál sería el número mínimo de machos necesarios para mantener N_e por encima de un valor podemos: $H \geq \frac{N_e}{2}$
- Para **una consanguinidad dada y valores de intervalo generacional relativamente bajos** es relativamente fácil una **reducción importante en el número de machos a usar**, a un **costo relativamente menor en el incremento del intervalo generacional.**

14.2 El coeficiente de consanguinidad en razas lecheras

Claramente, como vimos antes la relación entre la tasa de incremento de la consanguinidad y el tamaño efectivo de una población nos permite reflexionar sobre la estructura de diferentes poblaciones de animales domésticos y sus implicancias para el mejoramiento o aún la subsistencia a largo plazo de razas y especies. La producción lechera es un excelente caso de estudio ya que existen diferentes razas que han seguido trayectorias diferentes durante el siglo XX y que en algunos casos han soportado presiones de selección extremadamente elevadas. La búsqueda de los animales más productivos puede ser un excelente objetivo a

corto plazo o aún a largo plazo para un establecimiento, pero sin duda es una estrategia peligrosa para el largo plazo en toda una raza o aún en una especie.

El problema es muy claro y lo podemos representar con una situación absurda por lo extrema: si elegimos al mejor toro como padre de toda la siguiente generación (cosa posible y que acontece), todos los individuos de la siguiente generación serán medios hermanos entre sí. Más aún, los **cromosomas Y** de los terneros machos serán todos iguales, desapareciendo toda la diversidad en ese cromosoma dentro de nuestro establecimiento. Pero peor aún es si todos los establecimientos del país utilizan el semen del mismo toro que elegimos nosotros, ya que claramente la situación se extenderá a toda la población nacional. Como toda la generación es hija del mismo toro, tenemos tres opciones: *a)* que los toros elegidos ahora sean de los que no consideramos óptimos en la generación anterior, *b)* utilizar el mismo semen, pero ahora con sus hijas y *c)* elegir toritos entre los medios hermanos de las vaquillonas. Claramente, ninguna de las opciones nos parece buena y esto viene de que el esquema que propusimos para representar el problema es absurdo. Sin embargo, a pesar de lo absurdo, la realidad en muchas razas no dista mucho de esto.

En algunas razas, la introducción de la tecnología de inseminación artificial (**IA**) en la década de 1960, en los países con las mayores poblaciones de la raza, implicó un profundo impacto en la estructura del rodeo nacional y aún mundial. Un ejemplo extremo de esto es la raza **holstein** (conocida como **holando** en nuestro país), que vivió a escala mundial una expansión gigantesca durante el siglo XX. De acuerdo con un estudio realizado por Yue, Dechow y Liu (2015), de los 1821 toros presentes en la década de 1960, que llamaron IA-fundadores, solamente 3 de ellos, HOUSA1427381 (**Pawnee Farm Arlinda Chief**, 5/9/1962), HOUSA1441440 (**Penstate Ivanhoe Star**, 20/01/1963) y HOUSA1491007 (**Round Oak Rag Apple Elevation**, 30/08/1965), estaban presentes con descendientes vivos en la década de 2010. Es decir, más allá de la posible importación de material de otros orígenes (lo autores observan lo mismo para los datos de **INTERBULL**, un consorcio internacional, en la evaluación **MACE**), casi todo el material genético específico de machos (cromosoma Y, donde se encuentran genes de fertilidad en machos) pertenece a estos fundadores. Más aún, dos de estos 3 fundadores se convirtieron en dominantes en la década de 2010, con el 48,78 % de todos los toros en el pedigrí de **Pawnee Farm Arlinda Chief**, y el 51,06 % en el pedigrí de **Round Oak Rag Apple Elevation**. A causa de que se trataba de un toro portador de dos enfermedades autosomales recesivas (CVM y BLAD), la contribución del tercer fundador (**Penstate Ivanhoe Star**) disminuyó gradualmente, pasando del 10 % en los años noventa a solo un 0,16 % en la década de 2010, no sin antes haber provocado un importante daño, como veremos en la sección Los riesgos del uso masivo de reproductores y la endogamia elevada.

Por otra parte, de acuerdo con los mismos autores, la tasa de crecimiento de la consanguinidad entre los años 2010 y 2014 fue de 0,57 % (determinado a partir de datos en 2021). Asumiendo un intervalo generacional de 4 años para **holsteins de América del Norte**, como $\Delta F = \frac{1}{2N_e} \therefore N_e = \frac{1}{2\Delta F}$ y con esa tasa de incremento generacional el tamaño efectivo poblacional será igual a $N_e = \frac{1}{2\times 0{,}0057} \approx 88$ individuos. En otras palabras, toda la población **holstein** de América del Norte, que es de más de 9 millones de individuos, sería equivalente a menos de 100 individuos reproduciéndose en condiciones ideales.

La evolución de la consanguinidad en las poblaciones de las razas **holstein/roja y blanca*** y **milking shorthorn** de América del Norte pueden observarse en la figura 14.4 a partir de datos obtenidos de la base del Council on Dairy Cattle Breeding (**CDCB**).

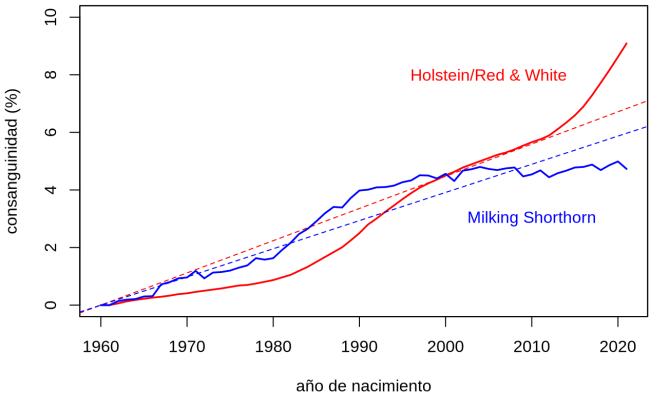


Figura 14.4. Evolución de la consanguinidad promedio en vacas de acuerdo con el año de nacimiento, para la población de las razas **holstein/roja y blanca** (rojo) y **milking shorthorn** (azul) con registros en la base de datos del Council on Dairy Cattle Breeding. Las líneas a trazos marcan las tendencias lineales asumiendo que el valor de $F = 0$ en 1960 es un punto fijo. Realización propia sobre datos en https://queries.uscdcb.com/eval/summary/inbrd.cfm

Claramente, en **holstein** se observa una fuerte tendencia al crecimiento de la consanguinidad en el rodeo, tendencia que se ha visto acelerada en los últimos años, en particular desde el año 2014. Esto podría estar, al menos parcialmente, relacionado con el uso de información genómica en las evaluaciones genéticas. De hecho, en toros jóvenes «genómicos», el coeficiente de consanguinidad a partir del pedigrí es superior al 12 % y superior al 14 % en el caso del coeficiente genómico. Dicho sea de paso, las correlaciones entre los coeficientes de consanguinidad obtenidos a partir del pedigrí y los obtenidos a partir de datos genómicos no necesariamente tienen una alta correlación (Cortes-Hernández et al. (2021)), aunque sí la tienen entre las distintas formas de calcularlo a partir de datos genómicos (la matriz de parentesco, las corridas de homocigosidad **ROH**, etcétera).

Por otra parte, en el caso de la raza **milking shorthorn** la tendencia creciente que se observaba después de los años sesenta mostró una reducción importante y en la actualidad se encuentra bastante «estancada» (lo que sería muy deseable). En otras razas lecheras (**ayrshire**, **brown swiss**, **guernsey** y **jersey**) de la misma base de datos las tendencias han sido bastante similares a la de **holstein**, por lo que la raza **milking shorthorn** parece ser más la excepción que la regla. Mientras que se trata de una raza plenamente productiva y que no se encuentra en peligro a escala mundial, la población de la raza original (**dairy shorthorn**), en el noreste de Inglaterra y que no ha sufrido del cruzamiento indiscriminado en el siglo XX sí se encuentra en el nivel máximo de riesgo de desaparecer.

Ejemplo 14.3

De acuerdo con los datos del **CDCB**, la consanguinidad del rodeo **holstein/roja y blanca** de América del Norte pasó de $F = 7{,}72\,\%$ a $F = 9{,}09\,\%$ en los últimos 4 años (2018-2021), mientras que entre 1960 y 1963 pasó de $F = 0\,\%$ a $F = 0{,}13\,\%$. Considerando que 4 años es el intervalo generacional de la raza, calcular el tamaño efectivo poblacional

que se corresponde a cada uno de estos períodos. Asumiendo que el número de hembras es muchísimo mayor que el de machos, cosa que es cierta en este caso (más de 9 millones de vacas en el rodeo de América del Norte), ¿cuántos machos estarían usando actualmente y cuántos se deberían usar para que la consanguinidad anual no aumente más de 0,1 %?

Utilizando la ecuación que nos relaciona el incremento en la consanguinidad de una generación con el tamaño efectivo poblacional, tenemos que

$$\Delta F = \frac{1}{2N_e} \;\therefore\; N_e = \frac{1}{2\Delta F} \tag{14.14}$$

En el período 2018-2021 $\Delta F = 9{,}09\,\% - 7{,}72\,\% = 1{,}37\,\%$, mientras que en el período 1960-1963 tenemos $\Delta F = 0{,}13\,\% - 0{,}00\,\% = 0{,}13\,\%$. Por lo tanto, sustituyendo esos valores en la ecuación de arriba, los correspondientes tamaños efectivos poblacionales son de $N_e = \frac{1}{2\times 0{,}0137} = 36{,}5$ y $N_e = \frac{1}{2\times 0{,}0013} = 384{,}6$ individuos.

Si el número de hembras es muchísimo mayor que el de machos, entonces

$$\Delta F_a \approx \frac{1}{8\, m\, IG^2} \therefore m \approx \frac{1}{8\, \Delta F_a\, IG^2} \tag{14.15}$$

Sustituyendo por los valores, tenemos que actualmente, con una tasa de consanguinidad anual de $\Delta F_a = \frac{1{,}37\,\%}{4} = 0{,}343\,\%$, se están usando un equivalente a

$$\begin{aligned} m \approx \frac{1}{8\, \Delta F_a\, IG^2} = \frac{1}{8 \times \frac{0{,}0137}{4} \times 4^2} = \\ = \frac{1}{8\, \Delta F_a\, IG^2} = \frac{1}{8 \times 0{,}0137 \times 4} = 2{,}28 \text{ machos} \end{aligned} \tag{14.16}$$

mientras que para mantener un incremento anual de la consanguinidad de 0,1 % se debería usar al menos

$$m \approx \frac{1}{8 \times 0{,}001 \times 4^2} = 7{,}81 \text{ machos} \tag{14.17}$$

PARA RECORDAR

- Para la población de vacas **holstein** de América del Norte, fue determinado que entre los años 2010 y 2014 la tasa de crecimiento de la consanguinidad fue de 0,57.
- A su vez, se ha determinado que toda la población **holstein** de América del Norte (más de 9 millones de individuos) sería equivalente a aproximadamente 88 individuos reproduciéndose en condiciones ideales.

14.3 Depresión endogámica

Una de las consecuencias que ya discutimos del incremento en la consanguinidad es el consiguiente aumento en la proporción de individuos homocigotos, a costa obviamente de la proporción de heterocigotos. Por otra parte, esto lleva a que algunas características donde **la dominancia juega un papel relativamente importante** se vean afectadas en su expresión, disminuyendo en general. En particular, dentro de este grupo de características parecen estar las que tienen directa incidencia en el **fitness** de los individuos (las reproductivas, por ejemplo) y las fisiológicas, que es la definición de Douglas S. Falconer (1983), por ejemplo, aunque se ha observado en otros tipos de características. Es importante mencionar que cuando hablamos de depresión endogámica, generalmente nos referimos a la disminución de una media poblacional y no al efecto de un individuo en particular.

Existen diferentes formas de estimar la depresión endogámica (Keller and Waller 2002), dependiendo de la especie, los registros o recursos disponibles. En algunos casos, alcanza con comparar una subpoblación creada con una determinada estructura de parentesco (un coeficiente de endocría particular) y comparar la media fenotípica a la de la población de referencia (que se asume con $F = 0$). Por ejemplo, en las poblaciones de plantas **autógamas**, la depresión por endogamia puede calcularse comparando la aptitud de la progenie **autofecundada** y la **cruzada**. Si se utiliza un progenitor no consanguíneo ($F = 0$), la progenie cruzada al azar también será no consanguínea ($F = 0$), mientras que la progenie autofecundada tendrá un $F = \frac{1}{2}$

En otros casos, es posible explotar la información del pedigrí de los individuos e incluir una covariable en el modelo que sea el coeficiente de consanguinidad del individuo (calculado a partir del pedigrí como la mitad del parentesco entre sus progenitores). Por ejemplo, de acuerdo con Mc Parland et al. (2007), en Irlanda una vaca **holstein** primípara, con una consanguinidad de $\frac{1}{8} = 12{,}5\,\%$ (es decir, tras el apareamiento de medios hermanos no consanguíneos), tuvo una disminución de $-61{,}8$ kg de leche, $-5{,}3$ kg de grasa y $-1{,}2$ kg de proteína en la lactancia. Además, las concentraciones de grasa y proteína se redujeron en $0{,}05$ y $0{,}01\,\%$, respectivamente, mientras que el score de células somáticas (que se calcula como el logaritmo natural del recuento de células somáticas dividido por 1.000) aumentó en $0{,}03$. Por otra parte, de acuerdo con las frecuencias alélicas de la población irlandesa de holstein, en animales con este nivel de consanguinidad se espera una incidencia de distocia un $2\,\%$ mayor, un $1\,\%$ más de incidencia de mortinatos, un $0{,}7\,\%$ más de incidencia de terneros machos (es decir, una disminución en el número de reemplazos hembras, que son los importantes), un aumento del intervalo entre partos de $8{,}8$ días, un aumento de la edad al primer parto de $2{,}5$ días y una disminución del número de animales que llega a la segunda lactancia de $-4\,\%$. Claramente, se trata de un nivel de consanguinidad muy importante y que debería resultar fácil de evitar, aun cuando la población de holstein a escala mundial muestra signos evidentes de consanguinidad elevada.

Sin embargo, mientras que Bezdíček, Šubrt y Filipčík (2008) confirman los hallazgos de la depresión endogámica para producción de leche en la población de vacas **holstein** de la República Checa, los resultados para la concentración de grasa y proteína van en sentido contrario a los de Mc Parland et al. (2007). Mientras que por cada incremento en la consanguinidad de $F_x = 1\,\%$ se espera **una reducción** en la **producción de leche** de $-59{,}75$ kg, las concentraciones de grasa y proteína se incrementaron en $+0{,}0112\,\%$ y $+0{,}0030\,\%$ respectivamente. A primera vista parecería que los dos estudios difieren significativamente en las conclusiones alcanzadas respecto a grasa y proteína, mientras que coinciden para

la producción lechera, pero si lo miramos con un poco más detenimiento las diferencias no vienen por ese lado. Una advertencia obvia es que se trata de dos estudios diferentes, en países completamente diferentes y con diferente tradición lechera y diferentes animales. Las diferencias en los estimados obtenidos para la depresión endogámica en producción lechera son enormes: asumiendo como cierta la linealidad de la depresión endogámica con el coeficiente de endocría (algo que veremos más adelante), si la reducción en leche es de $-59{,}75$ kg para un coeficiente de endocría de $F_x = 1\,\%$, entonces **aplicando regla de tres**, la reducción será igual $-746{,}9$ para un coeficiente de endocría de $F_x = 12{,}5\,\%$ (el que aparece en Mc Parland et al. (2007) para una reducción equivalente a $-61{,}8$ kg). Claramente, la concentración de sólidos (grasa y proteína) es igual a la masa producida (en kg, por ejemplo) de dichos sólidos, dividida entre el volumen de leche. Como el mismo se redujo de una manera drástica, aunque la producción de sólidos también hubiese disminuido a causa de la depresión endogámica, si la proporción de reducción es menor que la observada para leche, entonces aumentará la concentración con la depresión endogámica, que es lo que se observa.

Otro punto relevante al extrapolar resultados es la posible no-linealidad de la depresión endogámica con el coeficiente de consanguinidad, como es de esperar si la causa de la misma es la interacción epistática entre los genes. Miglior, Szkotnicki y Burnside (1992) analizaron los registros de leche, grasa y porcentaje de grasa de la primera lactación de 53.592 vacas jersey. A partir de esos datos, los coeficientes de regresión de la leche, la grasa y el porcentaje de grasa en función de la consanguinidad fueron de $-9{,}84$ kg, $-0{,}55$ kg y $-0{,}0011\,\%$ por cada $1\,\%$ de aumento de la consanguinidad. Sin embargo, cuando el coeficiente de consanguinidad fue mayor a $12{,}5\,\%$ la depresión endogámica fue mucho mayor a la esperada.

La depresión endogámica es un fenómeno que ha sido observado consistentemente, sin embargo, las causas biológicas del mismo aún no han sido totalmente clarificadas (Charlesworth and Willis 2009). Esencialmente, para las características que tienen impacto directo en el **fitness**, tradicionalmente se han manejado tres hipótesis que podrían explicar la existencia del fenómeno: *a)* la **hipótesis de la dominancia**, *b)* la **hipótesis de la sobredominancia** y *c)* la **hipótesis de la epistasis**.

En la primera hipótesis, se trata de que en presencia de alelos recesivos mayormente deletéreos gobernando la característica, en baja frecuencia, el incremento de los homocigotas llevaría a una depresión en la característica ya que al tratarse de recesivos se expresarán fenotípicamente solo en homocigosis. Por otro lado, en la segunda hipótesis, si el mecanismo de los genes que gobiernan una característica es la **sobredominancia**, es decir, el mayor **fitness** de los heterocigotas respecto a los homocigotas, la reducción en el número de heterocigotas (que son lo que tienen el «mejor» fenotipo) tendrá un impacto directo en la característica. Sin embargo, Charlesworth y Willis (2009) sugieren que la presencia de mutaciones deletéreas en (por ejemplo) dos genes **en fase de repulsión** se manifieste como una ventaja de los heterocigotos respecto a los dos haplotipos homocigotos (uno sería homocigoto para un alelo recesivo deletéreo en un gen y el otro para el alelo del otro gen). En este caso, si bien es difícil de distinguir en la práctica de la sobredominancia, en realidad, desde el punto de vista teórico se trata de una **pseudosobredominancia** causada por el desequilibrio de ligamiento entre los genes. Por último, en la hipótesis de la epistasis, la idea es que los heterocigotos son más libres de encontrar combinaciones óptimas y a diferencia de las otras dos hipótesis esta plantea un comportamiento no-lineal respecto al coeficiente de consanguinidad.

En la sección «Endocría y depresión endogámica» del capítulo «Apareamientos no-aleatorios» ya hicimos una introducción al modelado de la depresión en función del

coeficiente de endocría, las frecuencias génicas y el modo de acción génica. En esta sección vamos a ver el tema desde un punto de vista ligeramente diferente y que nos llevará naturalmente a generalizar las conclusiones a características gobernadas por unos pocos o por muchos loci, sin distinción. Para eso vamos a partir de los cambios en las frecuencias genotípicas que podemos esperar en presencia de endogamia, respecto a la ausencia de la misma, en ambos casos asumiendo Hardy-Weinberg, en el segundo caso la «Generalización de Hardy-Weinberg para apareamientos no-aleatorios».

Supongamos primero que la característica está determinada por un solo locus con dos alelos, como ya vimos previamente. De acuerdo con lo esperado en presencia de endogamia, las frecuencias genotípicas serán $\bar{p}^2+\bar{p}\bar{q}F$ para los A_1A_1, $2\bar{p}\bar{q}-2\bar{p}\bar{q}F$ para los A_1A_2 y $\bar{q}^2+\bar{p}\bar{q}F$ para A_2A_2. Como vimos en el capítulo «El modelo genético básico, los valores de estos genotipos serán $+a$ para el A_1A_1, d para el A_1A_2 y $-a$ para el A_2A_2, por lo que poniendo todo junto en una tabla tenemos

Genotipo	Frecuencia	Valor	Valor ponderado
A_1A_1	$\bar{p}^2+\bar{p}\bar{q}F$	$+a$	$\bar{p}^2a+\bar{p}\bar{q}aF$
A_1A_2	$2\bar{p}\bar{q}-2\bar{p}\bar{q}F$	d	$2\bar{p}\bar{q}d-2\bar{p}\bar{q}dF$
A_2A_2	$\bar{q}^2+\bar{p}\bar{q}F$	$-a$	$-\bar{q}^2a-\bar{p}\bar{q}aF$
SUMA			$a(\bar{p}-\bar{q})+2d\bar{p}\bar{q}(1-F)$

Para calcular la media genotípica de la población en presencia de endogamia alcanza con sumar los genotipos ponderados por su frecuencia, como está en la última columna de la tabla. En particular, recordando que $(p^2-q^2)=(p+q)(p-q)=(p-q)$ ya que $p+q=1$, entonces

$$M_F=[\bar{p}^2a+\bar{p}\bar{q}aF]+[2\bar{p}\bar{q}d-2\bar{p}\bar{q}dF]+[-\bar{q}^2a-\bar{p}\bar{q}aF]= \tag{14.18}$$

$$a(\bar{p}^2-\bar{q}^2)+[\bar{p}\bar{q}aF-\bar{p}\bar{q}aF]+[2\bar{p}\bar{q}d-2\bar{p}\bar{q}dF]= \tag{14.19}$$

$$a(\bar{p}^2-\bar{q}^2)+[2\bar{p}\bar{q}d-2\bar{p}\bar{q}dF]=a(\bar{p}^2-\bar{q}^2)+2\bar{p}\bar{q}d(1-F)\ \therefore \tag{14.20}$$

$$M_F=a(\bar{p}-\bar{q})+2\bar{p}\bar{q}d(1-F) \tag{14.21}$$

o, teniendo en cuenta que $M_0=a(\bar{p}^2-\bar{q}^2)+2\bar{p}\bar{q}d$, entonces

$$M_F=M_0-2d\bar{p}\bar{q}F \tag{14.22}$$

El cambio resultante en la media debido a la endogamia es por lo tanto $2d\bar{p}\bar{q}F$. Dicho de otra manera, la depresión causada por la endogamia (observar el signo negativo que antecede a $2d\bar{p}\bar{q}F$) es proporcional al **coeficiente de endocría**, a la **dominancia** y también proporcional al **producto de las frecuencias alélicas** medias. Cuando $\bar{p}=\frac{1}{2}=\bar{q}$ la depresión será máxima, mientras que a valores de $\bar{p}$ cercanos a uno o a cero la depresión será nula. Esto se explica muy simplemente de manera intuitiva, ya que como manejamos previamente la depresión endogámica es «producto» de la reducción en los heterocigotos (e incremento de los homocigotas), los que son una función directa de $\bar{p}\bar{q}$. Además, tanto F como d afectan en forma directa a la depresión endogámica.

Como en el desarrollo anterior no hicimos ninguna alusión a un modo particular de acción alélica y solo apelamos a los valores genotípicos a y d, en la medida en que la epistasis no sea un mecanismo relevante para la característica, alcanza con sumar el efecto de depresión a través de todos los loci que participan de la determinación genética de la característica. En cada uno de ellos la frecuencia de los alelos será diferente y por lo tanto deberemos expresarlas como p_i y $q_i = 1 - p_i$. Además, en cada locus los valores genotípicos serán diferentes, por lo que también debemos expresarlos como a_i y d_i. El coeficiente de endocría, sin embargo, será el mismo para todos los loci ya que se trata de una propiedad general de la población y que no varía entre loci.

De acuerdo con lo anterior, podemos escribir el efecto de todos los loci en la media de la característica bajo endogamia como

$$M_F = \sum_i a_i(\bar{p}_i^2 - \bar{q}_i^2) + 2\sum_i \bar{p}_i\bar{q}_i d_i(1 - F) \tag{14.23}$$

y el consiguiente cambio en la media, respecto a la población sin endogamia como

$$M_F = M_0 - 2\sum_i \bar{p}_i\bar{q}_i d_i F \tag{14.24}$$

Ejemplo 14.4a

En la siguiente tabla se presentan diferentes situaciones de porcentaje de consanguinidad anual, y su consecuente tiempo en disminuir cierto peso al destete promedio, manteniendo en todas las situaciones, una depresión endogámica para peso al destete de 3,4 kg por cada 1 % de consanguinidad:

Consanguinidad de la población (%/año)	Depresión endogámica (kg peso al destete/año)	Disminución del peso al destete promedio (kg)	Tiempo en disminuir el peso al destete promedio (años)
0,22	0,73	10	$13{,}6 \simeq 14$
0,40	1,36	10	$7{,}4 \simeq 8$
0,80	2,72	10	$3{,}7 \simeq 4$
1,30	4,42	10	$2{,}3 \simeq 3$
1,50	5,10	10	$2{,}0 \simeq 2$
4,00	13,60	10	$0{,}7 \simeq 1$
0,22	0,73	20	$27{,}2 \simeq 28$
0,40	1,36	20	$14{,}7 \simeq 15$
0,80	2,72	20	$7{,}4 \simeq 8$
1,30	4,42	20	$4{,}5 \simeq 5$
1,50	5,10	20	$3{,}9 \simeq 4$
4,00	13,6	20	$1{,}5 \simeq 2$

Ejemplo 14.4b

En la siguiente tabla se presentan diferentes situaciones de porcentaje de consanguinidad generacional y su consecuente tiempo en disminuir cierto porcentaje de destete, manteniendo en todas las situaciones, una depresión endogámica para porcentaje de destete de 1,5 % por cada 1 % de consanguinidad:

Consanguinidad (%/generación)	Depr. endogámica (% de destete/generación)	Disminución del porcentaje de destete (%)	Tiempo en disminuir el porcentaje de destete (gens.)
0,20	0,30	10,4	$34{,}7 \simeq 35$
0,40	0,60	10,4	$17{,}3 \simeq 18$
0,88	1,32	10,4	$7{,}9 \simeq 8$
1,30	1,95	10,4	$5{,}3 \simeq 6$
1,50	2,25	10,4	$4{,}6 \simeq 5$
4,00	6,00	10,4	$1{,}7 \simeq 2$
0,20	0,30	15	$50{,}0 \simeq 50$
0,40	0,60	15	$25{,}0 \simeq 25$
0,88	1,32	15	$11{,}4 \simeq 12$
1,30	1,95	15	$7{,}7 \simeq 8$
1,50	2,25	15	$6{,}7 \simeq 7$
4,00	6,00	15	$2{,}5 \simeq 3$

PARA RECORDAR

- El incremento en la consanguinidad, que implica un aumento en la proporción de homocigotos y esto lleva a que algunas **características donde la dominancia juega un papel relativamente importante** se vean afectadas en su expresión, disminuyendo en general. En particular, dentro de este grupo de características parecen estar las que **tienen directa incidencia en el fitness de los individuos y las fisiológicas.**
- Es posible a través del **pedigrí de los individuos** determinar el coeficiente de consanguinidad del individuo (calculado como la mitad del parentesco entre sus progenitores).
- El cambio resultante en la media debido a la endogamia ($M_F = M_0 - 2d\bar{p}\bar{q}F$) es $2d\bar{p}\bar{q}F$. Tal que, la depresión causada por la endogamia es proporcional al **coeficiente de endocría**, a la **dominancia** y también al **producto de las frecuencias alélicas** medias. Cuando $\bar{p} = \frac{1}{2} = \bar{q}$ la depresión será máxima, mientras que a valores de $\bar{p}$ cercanos a uno o a cero la depresión será nula.
- Si consideramos más de un loci el consiguiente cambio en la media, respecto a la población sin endogamia será $M_F = M_0 - 2\sum_i \bar{p}_i\bar{q}_i d_i F$

Los riesgos del uso masivo de reproductores y la endogamia elevada

La producción lechera es posiblemente el mejor ejemplo de un uso masivo y abusivo de un número pequeño de reproductores. Las claves de este uso indiscriminado de los mismos reproductores se encuentran en el deseo de maximizar las ganancias a corto plazo, considerando solamente aquellas características que rinden en lo inmediato y descuidando el incremento consiguiente de la consanguinidad, la reducción en la fertilidad, la reducción de la variabilidad genética y los riesgos de reproducir animales portadores de enfermedades de base recesiva. Claramente, pese a estas «motivaciones» importantes, al potencial desastre que enfrentamos solo fue posible arribar gracias a las tecnologías de reproducción masivas, en particular la producción de semen congelado y las técnicas de inseminación *in vitro*. Como ejemplo de esto mencionaremos solo algunas enfermedades de carácter autosomal recesivo y su impacto en la producción lechera.

Malformación vertebral compleja: (**CVM**) Es una enfermedad responsable de terneros con malformaciones que abortan espontáneamente o mueren poco después de nacer. Está causada por una mutación de pérdida de sentido en el gen **SLC35A3** (Thomsen et al. 2006) y fue reportada por primera vez en el año 1999. De acuerdo con el genotipado realizado, dicha mutación se extendió en forma significativa a partir del uso masivo del toro **Carlin-M Ivanhoe Bell** (apodado **Bell**), nacido en 1974 y usado durante las décadas de 1980 y 1990, un toro de gran valor por la gran producción lechera de sus hijas (dejando una progenie de más de 80.000 individuos). Curiosamente, este toro es **portador** (es decir, al tratarse de enfermedades autosomales recesivas los portadores solo tienen una copia de la mutación) de una mutación para otra enfermedad autosomal recesiva conocida como **BLAD**, cuyo gen se encuentra en otro cromosoma y que veremos luego. El padre de **Carlin-M Ivanhoe Bell**, el toro **Penstate Ivanhoe Starissa** ya contaba con las dos mutaciones, pero no así su abuelo **Osborndale Ivanhoe** que solo era **portador para BLAD**. La incidencia de esta mutación es muy alta, alcanzando un porcentaje de portadores mayor al 30 % en rodeos de Dinamarca y Japón. Como solamente los homocigotos son afectados, pero en ese caso no hay descendencia (un 88 % abortan antes del día 260 desde la concepción y los que nacen no son viables), suponiendo esta frecuencia en machos y hembras del rodeo, entonces $q^2 = 0{,}30^2 = 0{,}09$, o una **reducción en la fertilidad del rodeo** de 9 % (que puede ser aún mayor en presencia de endogamia).

Deficiencia de adhesión leucocitaria bovina: (**BLAD**) Es una enfermedad congénita autosómica recesiva de la raza **holstein.**
Se caracteriza por infecciones bacterianas recurrentes, retraso en la cicatrización de las heridas y retraso en el crecimiento, fiebre, úlceras graves, falta de apetito, neumonía crónica y diarrea y que también se asocia a una marcada neutrofilia

persistente (Nagahata 2004). Como vimos antes, su diseminación entre la población de ganado **holstein** se puede remontar al toro **Osborndale Ivanhoe**, muy utilizado en las décadas de los cincuenta y sesenta del siglo pasado. La base molecular de la enfermedad es una mutación puntual (de adenina a guanina) en la posición 383 del gen **CD18**, que provoca una **sustitución de ácido aspártico por glicina** en el aminoácido 128 (**D128G**) de la molécula de adhesión CD18. Como resultado de esto, los neutrófilos tienen una expresión alterada de la integrina beta2 (CD11a,b,c/CD18) de la molécula de adhesión leucocitaria. El ganado afectado muere a una edad temprana debido a las complicaciones infecciosas. Los programas de control de **portadores** que se han utilizado para la raza han sido relativamente exitosos para reducir la frecuencia de esta enfermedad.

Síndrome de brachiespina bovina: Se trata de una enfermedad del ganado **holstein** en la que el ternero nace muerto, con una drástica reducción del peso corporal, retraso en el crecimiento y acortamiento de la espina dorsal con malformaciones vertebrales asociados. Es causada por una deleción de 3,3 kb en el gen del grupo I de complementación de la **anemia de Fanconi bovina** (FANCI) en el cromosoma 21 (Fang et al. (2013)). Fue detectada en un principio en norteamérica y Dinamarca, luego en Asia y recientemente en nuestro país. En principio, su diseminación se puede trazar al uso del toro norteamericano **Sweet Haven Tradition**, usado a fines de los setenta y durante los ochenta. Se estima que dentro de la raza **holstein** de 6 % a 7 % son portadores.

Haplotipos de fertilidad: Se trata de regiones del genoma identificadas a partir de estudios genómicos (es decir, a partir de animales que poseen datos de genotipado masivo) y que implican una reducción en la fertilidad de los portadores de dichos haplotipos. Los conocidos son **holstein Haplotype 1-5** (HH1, HH2, HH3, HH4, HH5), **jersey Haplotype 1** (JH1), **Brown Swiss Haplotype 1 y 2** (BH1 y BH2), y **Ayrshire Haplotype 1** (AH1). Por lo general producen abortos a diferentes edades del feto, siendo las más perjudiciales desde el punto de vista económico aquellas que producen abortos tardíos. Afortunadamente, como su identificación ha sido con base en el genotipado, pese a la complejidad de sus mecanismos moleculares (que en la mayor parte no han sido dilucidados) son fáciles de identificar a partir del genotipado, algo que ya es rutina para los machos muy usados en ganado lechero. En los catálogos la forma usual de identificar a los portadores es a través de la letra **C** (*carrier*, portador) luego del correspondiente haplotipo, por ejemplo, **HH3C** o por la letra **F** para los no-portadores, por ejemplo, **HH3F**.

Sindactilia: Es una enfermedad congénita de carácter autosomal recesivo, en principio causada por un solo gen. El resultado es la unión de las pezuñas en una sola. Se ha reportado en las razas holstein, aberdeen-angus, simmental, brown swiss, chianina, nativa japonesa, hariana, sueca roja y blanca y checa negra y blanca. En **holstein** se puede trazar su aparición a un solo reproductor. Desde el punto de vista molecular la mayor parte de los casos en **holstein** se deben a la sustitución de dos nucleótidos en el exón 33 del gen **LRP4**, aunque recientemente se revelaron nuevas **mutaciones puntuales no-sinónimas** de **LRP4** que coexisten en familias **holstein**, **simmental alemán** y **simmental-charolais** (Drögemüller et al. 2007).

Existen muchas otras enfermedades congénitas del ganado y en particular del ganado lechero, también autosomales recesivas, como la **mielopatía degenerativa progresiva bovina** conocida como **Weaver** por el andar errático de los animales, o la recientemente identificada como **haplotipo para la deficiencia de colesterol (HCD)**, que podría trazarse hacia el toro **Maughlin Storm**, un toro nacido en 1991 (Kipp et al. 2016). El punto importante es que el uso masivo de toros portadores de enfermedades de carácter autosómico recesivo hace que sea difícil detectar el desastre que implica a mediano y largo plazo hasta que el mismo se encuentra en estado muy avanzado.

La endogamia elevada tiene otro efecto importante desde el punto de vista cuantitativo. Como vimos antes, bajo endogamia el coeficiente de endocría aumenta y esto lleva a un incremento en los homocigotos respecto a los heterocigotos. De hecho, como vimos en la sección «Generalización de Hardy-Weinberg para apareamientos no-aleatorios», en el caso de dos alelos, las frecuencias de los homocigotos pasan a ser $p^2 + pqF$ y $q^2 + pqF$. Supongamos que, sin pérdida de generalidad, q representa la frecuencia del alelo recesivo. Por lo tanto, en ausencia de endogamia la frecuencia esperada de la enfermedad (autosomal recesiva) será igual a la frecuencia de homocigotos con ese alelo, que es q^2. Sin embargo, en presencia de endogamia, con un coeficiente de endocría de F la frecuencia de homocigotos del alelo recesivo será $q^2 + pqF$, que será la frecuencia de afectados. Como $pqF \geqslant 0$ y solo será 0 cuando $F = 0$ (ausencia de endogamia) o $pq = 0$, esto último implicando la fijación de uno de los dos alelos (en cuyo caso no hay más enfermedad, tanto por ausencia de afectados o porque desaparecerían los sanos), entonces pqF aumenta la frecuencia de afectados.

Una forma de ver la importancia relativa de las dos fuentes es estudiar su relación, es decir, la función $f(q) = \frac{pqF}{q^2} = \frac{(1-q)F}{q}$. En la figura 14.5 podemos observar el comportamiento de esta función para cuatro valores del coeficiente de endocría F.

La línea a trazos en 1 representa el valor en que ambos componentes, numerador y denominador son iguales o, dicho de otra forma, los aportes **debidos a la endogamia** y los correspondientes a la ausencia de la misma son iguales. Claramente, a valores bajos de q (o sea, cuanto más raro es el alelo recesivo) más importante en términos relativos es el aporte debido a la endogamia y el comportamiento es no lineal, más importante cuanto más cerca estemos de $q = 0$. Por otra parte, para valores muy bajos del coeficiente de endocría, la importancia del componente debido a la endogamia solo empieza a notarse a frecuencias muy bajas del alelo recesivo, mientras que a valores un poco más altos de F las frecuencias

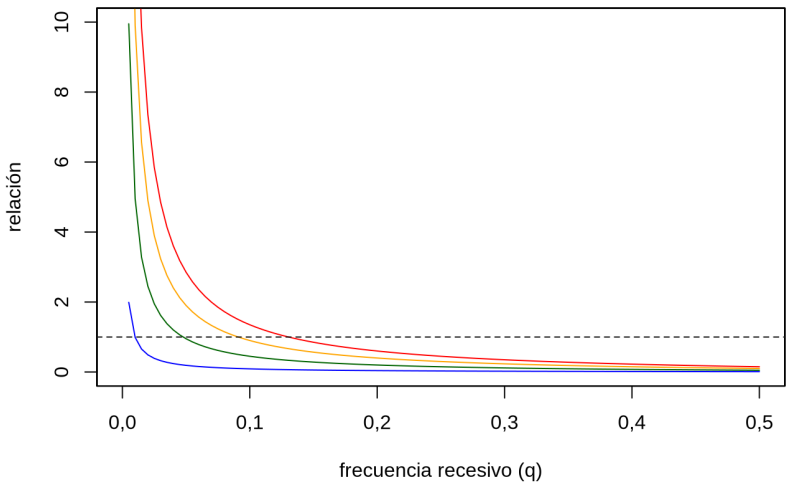


Figura 14.5. Relación entre el aporte de homocigotos **debido a la endogamia** y el aporte esperado en ausencia de endogamia, es decir $\frac{pqF}{q^2}$, en función de la frecuencia del alelo recesivo (q) para cuatro valores del coeficiente de endocría ($F = 0{,}01$ azul, $F = 0{,}05$ verde, $F = 0{,}10$ anaranjado y $F = 0{,}15$ rojo). Cuando la relación vale 1 (línea a trazos) ambos aportes son iguales, mientras que para valores mayores a 1 **solo el aporte debido a la endogamia** es mayor que lo esperado en ausencia de ella. Claramente, cuanto más baja es la frecuencia del alelo recesivo mayor importancia relativa adquiere el componente debido a la endogamia.

de q a las que empieza a observarse el aporte importante de la endogamia son mucho más altas. La igualdad del aporte de los dos componentes se da cuando

$$\frac{(1-q)F}{q} = 1 \Leftrightarrow (1-q)F = q \Leftrightarrow F = q + Fq \Leftrightarrow F = q(1+F) \therefore \quad q = \frac{F}{1+F} \tag{14.25}$$

Por ejemplo, para $F = 0{,}01$ la frecuencia a la que ambos componentes aportan lo mismo (la línea a trazos) es $q = \frac{F}{1+F} = \frac{0{,}01}{1+0{,}01} = 0{,}0099$, mientras que para $F = 0{,}15$ esto se da en $q = \frac{0{,}15}{1+0{,}15} = 0{,}1304$. A partir de estos puntos, si disminuye más la frecuencia del alelo recesivo, mayor será el aporte relativo de los homocigotos debidos a la endogamia.

Ejemplo 14.5

Es conocido que la raza **holstein** presenta un elevado nivel de consanguinidad tanto a nivel de la población mundial como en particular en determinadas poblaciones locales. Asumiendo que la frecuencia del alelo recesivo en el gen **SLC35A3** que causa la **malformación vertebral compleja** tiene una frecuencia de $q = 0{,}15$ en una población determinada, calcular la proporción de afectados cuando no existe consanguinidad ($F = 0$) y cuando el coeficiente de endocría es igual a $F = 0{,}08$.

En el primer caso, la frecuencia esperada de homocigotos (afectados) es $q^2 = 0{,}15^2 = 0{,}0225 = 2{,}225\,\%$. En el segundo caso, que parece describir mejor la situación en la raza **holstein**, la frecuencia esperada de homocigotos será de $q^2 + pqF = q^2 + (1-q)qF = 0{,}15^2 + 0{,}85 \times 0{,}15 \times 0{,}08 = 0{,}0327 = 3{,}27\,\%$. Por lo tanto, $\frac{3{,}27\,\%}{2{,}225\,\%} = 1{,}453$, es decir que tendremos $45{,}3\,\%$ más de terneros perdidos por la enfermedad a causa del incremento en la endogamia del rodeo.

PARA RECORDAR

- La producción lechera es posiblemente el mejor ejemplo de un uso masivo y abusivo de un número pequeño de reproductores. Este uso masivo de toros que a su vez han sido portadores de enfermedades de carácter autosómico recesivo hace que sea difícil detectar el desastre que implica a mediano y largo plazo hasta que el mismo se encuentra en estado muy avanzado.
- En presencia de endogamia, con un coeficiente de endocría de F, la frecuencia de homocigotos del alelo recesivo será $q^2 + pqF$, que será la frecuencia de afectados.

Redistribución en la varianza genética

La presencia de endogamia va a alterar la distribución de frecuencias alélicas entre un conjunto de poblaciones, respecto a esa evolución bajo un proceso de deriva en subpoblaciones de tamaño infinito. Mientras que el cálculo de cómo se distribuirá la varianza genética en presencia de endogamia entre línea y dentro de líneas es relativamente sencillo para genes completamente aditivos, el cálculo en presencia de dominancia es mucho más complejo y sus resultados de escasa utilidad práctica. En el caso de los genes completamente aditivos,

la partición de varianzas (como veremos) no depende de las frecuencias alélicas iniciales, mientras que en el caso de genes con dominancia esto no es así. Lo que sigue está basado en lo expuesto en Douglas S. Falconer (1983), que además da un tratamiento mucho más exhaustivo al tema de las diferentes fuentes de varianza y presenta numerosos ejemplos con datos reales.

Supongamos que tenemos la población total dividida en diferentes subpoblaciones de tamaño N. Vamos primero a considerar el caso de un solo locus, con dos alelos, como de costumbre, aunque la generalización del resultado es inmediata. Supongamos que las frecuencias iniciales globales de los dos alelos son p_0 y q_0.

Recordando que la varianza aditiva era $2pq\alpha^2$, con $\alpha = a + d(q - p)$ y dado que $d = 0 \therefore 2pq\alpha^2 = 2pqa^2$, la varianza genética será entonces

$$V_G = 2p_0q_0\alpha^2 = 2p_0q_0[a + d(q - p)]^2 = 2p_0q_0a^2 \tag{14.26}$$

Por otra parte, la varianza genética **dentro de cada una de las líneas** (subpoblaciones) estará dada por

$$V_{G_i} = 2p_iq_ia^2 \tag{14.27}$$

donde p_i y q_i son las frecuencias correspondientes a cada línea (el efecto a no depende de la línea y es, por lo tanto, igual para todas).

Si llamamos $\overline{pq}$ a la media de los productos p_iq_i, la **varianza media dentro** de las líneas es

$$V_{G_w} = 2(\overline{pq})a^2 \tag{14.28}$$

Por otra parte, $2\overline{pq}$ es la frecuencia de heterocigotas, que en función del proceso de endogamia es igual a $2p_0q_0(1 - F)$, por lo que sustituyendo en la ecuación (14.28), tenemos que

$$V_{G_w} = 2(\overline{pq})a^2 = 2p_0q_0(1 - F)a^2 \tag{14.29}$$

y combinando esto con la ecuación (14.26), nos queda que la varianza genética dentro de líneas en función de la endogamia es igual a

$$V_{G_w} = 2p_0q_0(1 - F)a^2 = (1 - F)\, V_G \tag{14.30}$$

Tenemos que calcular ahora **la varianza entre líneas.** Como $d = 0$, la media de cada una de las líneas estará dada por

$$M_i = a(p_i - q_i) + 2p_i q_i d = a(p_i - q_i) \tag{14.31}$$

Utilizando el hecho de que $p - q = (1 - q) - q = 1 - 2q$, entonces, la ecuación anterior la podemos escribir como $M = a(p_i - q_i) = a(1 - 2q_i) = a - 2aq_i$. Para calcular la varianza de una suma de variables aleatorias no correlacionadas, alcanza con recordar que es la suma de ambas varianzas, y en este caso a es una constante entre líneas, por lo que no va a covariar con $2aq$. Por lo tanto, la varianza de M será

$$\sigma^2_M = Var(a) + Var(2aq) = (2a)^2\sigma^2_q \therefore \\ \sigma^2_M = 4a^2\sigma^2_q \tag{14.32}$$

Ahora, como vimos en «El coeficiente de endocría y los estadísticos F, si re-arreglamos la ecuación (6.51), la varianza en frecuencias alélicas en presencia de endogamia (que es una función del tiempo) será

$$\sigma^2_q = p_0 q_0 F = p_0 q_0 \left[\left(1 - \frac{1}{2N}\right)^t\right] \tag{14.33}$$

Es decir, la dispersión de frecuencias alélicas entre líneas será directamente proporcional al coeficiente de endocría y a las frecuencias iniciales de los alelos en el locus bajo consideración. Por lo tanto, como F es una función del tiempo, la dispersión irá cambiando. Al mismo tiempo, la tasa de incremento en la dispersión de las frecuencias alélicas debido al efecto de la deriva, es decir, en poblaciones de tamaño finito y por lo tanto en presencia de endogamia será igual a ΔF.

Sustituyendo en resultado de la ecuación (14.33) en la ecuación (14.32) tenemos

$$\sigma^2_M = 4a^2\sigma^2_q = 4a^2 p_0 q_0 F \tag{14.34}$$

Finalmente, si recordamos de la ecuación (14.26) que $V_G = 2p_0 q_0 a^2$, entonces

$$\sigma^2_M = 4a^2 p_0 q_0 F = 2[2p_0 q_0 a^2]F \therefore \\ \sigma^2_M = 2\, V_G\, F \tag{14.35}$$

Como hemos trabajado solamente con un locus de efectos aditivos, el resultado se extiende a todos los loci con este tipo de acción y el resultado $\sigma^2_M = 2\, V_G\, F$ se mantendrá.

De acuerdo con esto, considerando solo el caso de características determinadas por genes de acción aditiva y en ausencia de epistasis, de acuerdo con las ecuaciones (14.35) y (14.30), la partición de la varianza genética estará dada por

Fuente de varianza	Varianza
Entre líneas	$2FV_G$
Dentro de líneas	$(1 - F)V_G$
Total	$(1 + F)V_G$

Como lo muestra la última línea de la tabla, la varianza total estará dada por $2\,F\,V_G + (1 - F)V_G = V_G[2F + 1 - F] = (1 + F)\,V_G$. Como el coeficiente de endocría tenderá a crecer con el paso del tiempo y eventualmente alcanzará el valor de $F \approx 1$, la varianza total será

$$V_{G_F} = (1 + F)\,V_G \approx (1 + 1)\,V_G = 2\,V_G \tag{14.36}$$

Dicho de otra forma, **con endogamia completa, la varianza genética de la población será el doble que en la población de base.** Además, como la varianza dentro de líneas es igual a $(1 - F)\,V_G$, para ese entonces la misma será $(1 - 1)\,V_G = 0$, o lo que es lo mismo, las líneas serán completamente homogéneas en lo interno, por lo que la varianza total será igual a la varianza entre líneas.

PARA RECORDAR

- Si tenemos la **población total dividida en diferentes subpoblaciones de tamaño** N, a la vez que consideramos el caso de un solo locus con dos alelos y las frecuencias iniciales globales de los dos alelos son p_0 y q_0, la **varianza genética** será $V_G = 2p_0q_0a^2$, siendo la **varianza dentro de cada una de las líneas** o subpoblaciones $V_{G_i} = 2p_iq_ia^2$ con p_i y q_i las frecuencias dentro de cada línea.
- Por su parte la varianza media **dentro de las líneas** será $V_{G_w} = 2(\overline{pq})a^2$, si la representamos en **función de la endogamia** sería $V_{G_w} = 2p_0q_0(1 - F)a^2 = (1 - F)\,V_G$
- Por su parte, la **dispersión de frecuencias alélicas entre líneas** será **directamente proporcional al coeficiente de endocría y a las frecuencias iniciales de los alelos en el locus bajo consideración**, tal que $\sigma_q^2 = p_0q_0F = p_0q_0\left[\left(1 - \frac{1}{2N}\right)^t\right]$
- La **tasa de incremento en la dispersión de las frecuencias alélicas** debido al efecto de la deriva, es decir, en poblaciones de tamaño finito y por lo tanto en presencia de endogamia será igual a ΔF.
- Con endogamia completa, **la varianza genética de la población será el doble que en la población de base** tal que $V_{G_F} = (1 + F)\,V_G \approx (1 + 1)\,V_G = 2\,V_G$

14.4 Exocría y heterosis

Hasta ahora, en diferentes capítulos, pero en particular en este, hemos trabajado el tema de cómo la separación de una población en varias produce diferentes fenómenos evolutivos, entre ellos la divergencia en frecuencias alélicas debido al tamaño finito de las subpoblaciones, que entre otras cosas lleva a eventos como la especiación. En particular, en este capítulo hemos trabajado sobre el fenómeno de la endogamia, un fenómeno inevitable cuando las poblaciones son pequeñas y como en algunos casos eso lleva a la depresión endogámica. Pero si se piensa este fenómeno desde el lado opuesto, la pregunta sería ¿qué ocurre cuando combinamos diferentes poblaciones?

Exocría es el proceso reproductivo (o una forma de reproducción) donde el **apareamiento** se produce entre individuos **menos emparentados que la media de la población total.**

Dicho de otra forma, la pregunta anterior se refiere a que, si al separar poblaciones se dan determinados procesos, qué ocurrirá si juntamos poblaciones que han sido separadas un determinado tiempo atrás. Algo de esto ya intuimos a partir del capítulo «Variación y equilibrio de Hardy-Weinberg» y vimos en mayor detalle en el capítulo «Apareamientos no-aleatorios» cuando discutimos el efecto Wahlund. Por un lado, poblaciones separadas suelen acumular diferencias a nivel genómico y a nivel fenotípico a medida que pasa el tiempo. La causa de los procesos de diferenciación a nivel genómico los hemos venido discutiendo y pueden resumirse, en selección natural y artificial, deriva genética, mutación y migración (además de otros como los procesos de recombinación, etcétera). Los principales procesos que nosotros hemos seguido hasta acá tienen que ver con el cambio de las frecuencias alélicas en los genes, ya que no hemos trabajado otros fenómenos como los re-arreglos cromosómicos, las duplicaciones génicas y una pléyade de otros eventos que ocurren en los genomas.

En el capítulo «Deriva genética» realizamos una analogía sobre los procesos de muestreos con jarras conteniendo bolitas de colores. Para pensar en lo que sigue del presente capítulo podemos manejarnos con la misma analogía tenemos dos jarras que difieren en las proporciones de bolitas rojas y azules. El proceso de exocría será similar a combinar las dos jarras y a partir de ello muestrear, como hacíamos en el capítulo de «Deriva genética».

Cuando previamente discutimos la depresión endogámica, vimos que las posibles causas tenían que ver con la existencia del fenómeno de dominancia, de sobredominancia o de epistasis. Dejando de lado la última, ya que el tratamiento es mucho más elaborado de los objetivos que nos hemos trazado y su importancia es debatible, las otras dos causas involucran la dominancia. Más adelante, en la sección «Un modelo sencillo de heterosis bajo exocría» desarrollaremos un modelo matemático muy sencillo que nos permitirá entender el rol de la dominancia y del cambio de las frecuencias genotípicas bajo exocría, pero por ahora nos alcanza con suponer (usando un argumento de simetría con la endocría) que los heterocigotos jugarán también un papel muy relevante acá.

Cuando se produce un cruzamiento entre dos poblaciones de la misma especie, por ejemplo, entre dos razas, si todos los genes que determinan desde el punto de vista genético la característica tienen un modo de acción aditivo, entonces lo esperable para el promedio fenotípico de los **individuos cruza** sería estar en el punto medio de los fenotipos de las poblaciones «parentales puras» (que dan origen a la cruza) de la misma generación. Esto, en particular, si los ambientes en los que se expresa la característica fenotípica son idénticos para las tres poblaciones (las dos originales y la cruza). Es importante aclarar aquí que **siempre trataremos de comparaciones entre poblaciones en la misma generación**, ya que así eliminamos otros efectos del ambiente o de selección.

A pesar de nuestras expectativas, en muchos casos se puede observar que el fenotipo de los individuos cruza es mejor (tiene un **fitness** más alto) que el promedio de las poblaciones o razas que originaron la cruza. Más aún, en muchos casos no resulta idéntico el fenotipo de los cruzamientos en un sentido, es decir machos de la primera población con hembras de la segunda, que el cruzamiento recíproco (hembras de la primera con machos de la segunda).

Pronto analizaremos las causas de esto en particular, pero por ahora nos manejaremos con el promedio de los cruzamientos recíprocos para definir la heterosis.

Definimos **heterosis** como la diferencia entre el fenotipo promedio de las cruzas (recíprocas) y el promedio de las poblaciones «puras» que las originaron.

Supongamos que tenemos dos poblaciones (o razas) diferentes, la **A** y la **B** y que luego del cruzamiento de machos **A** con hembras **B** y de hembras **A** con machos **B**, así como de la reproducción libre **dentro** de cada una de las poblaciones **A** y **B** arribamos a cuatro grupos (poblaciones), todas en la misma generación. Como los cruzamientos tienen dirección, desde el punto de vista del sexo, vamos a usar dos letras, la primera para referirnos a la raza de los padres (machos) y la segunda de las madres (hembras). Es decir, **AB** se refiere a una cruza de machos **AA** con hembras **BB** (notemos que ambos son «puros»).

Utilizando la definición de heterosis, desde el punto de vista matemático la heterosis entre las razas **A** y **B**, que llamaremos h_{AB} es igual a

$$h_{AB} = \frac{\bar{X}_{AB} + \bar{X}_{BA}}{2} - \frac{\bar{X}_{AA} + \bar{X}_{BB}}{2} \tag{14.37}$$

Notemos que $h_{AB} = h_{BA}$ ya que la definición de heterosis involucra el promedio de las cruzas recíprocas y también el promedio de ambas «puras», pero en forma simétrica respecto al sexo de padres y madres.

Como todas las cantidades que aparecen en la ecuación (14.37) tienen las unidades de la característica fenotípica de interés, también la heterosis tendrá dicha unidad. Esto nos deja sin la posibilidad de comparar características que se expresan en diferentes unidades, obviamente, pero también características que se expresan en las mismas unidades pero que no tienen nada que ver. Claramente, no tiene ningún sentido compara kilogramos de heterosis en producción de leche con kilogramos de heterosis en peso de vellón limpio.

Una forma de eliminar este problema es estandarizando la heterosis de cada característica, refiriéndola, por ejemplo, al promedio de la característica en las «puras». En términos matemáticos, la «heterosis porcentual» la podemos calcular de la siguiente manera

$$h_{AB\,\%} = \frac{h_{AB}}{\frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}} \times 100 = \frac{\frac{\bar{X}_{AB}+\bar{X}_{BA}}{2} - \frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}}{\frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}} \times 100 \tag{14.38}$$

Ejemplo 14.5

A fin de aumentar la producción de lana coloreada y mejorar la adaptación a nuestras condiciones productivas se realizó un experimento de cruzamiento entre individuos de las **ovejas verdes** e individuos de la raza corriedale.

La siguiente tabla representa los promedios obtenidos en peso de vellón limpio (**PVL**), tanto para los animales puros como para las cruzas de la misma generación:

Padres/Madres	Verdes	Corriedale
Verdes	3,275 kg	4,225 kg
Corriedale	4,055 kg	3,925 kg

Determinar la heterosis individual en este cruzamiento, tanto en las unidades de la característica como en porcentaje.

De acuerdo con la ecuación (14.37) la heterosis individual para dos razas **V** (verdes) y **C** (corriedale) la calculamos como

$$h_{VC} = \frac{\bar{X}_{VC} + \bar{X}_{CV}}{2} - \frac{\bar{X}_{VV} + \bar{X}_{CC}}{2} \tag{14.39}$$

por lo que, en nuestro caso, utilizando los valores de la tabla tenemos que

$$h_{VC} = \frac{4{,}225 + 4{,}055}{2} - \frac{3{,}275 + 3{,}925}{2}\ \text{kg} = 0{,}540\ \text{kg} \tag{14.40}$$

es decir, 540 gramos. Expresado como porcentaje respecto a las puras, utilizando la ecuación (14.38), la heterosis individual queda como

$$h_{VC\,\%} = \frac{h_{VC}}{\frac{\bar{X}_{VV}+\bar{X}_{CC}}{2}} \times 100 = \frac{\frac{\bar{X}_{VC}+\bar{X}_{CV}}{2} - \frac{\bar{X}_{VV}+\bar{X}_{CC}}{2}}{\frac{\bar{X}_{VV}+\bar{X}_{CC}}{2}} \times 100 \tag{14.41}$$

y sustituyendo valores, como el numerador equivale a la heterosis individual en unidades de la característica

$$h_{VC\,\%} = \frac{0{,}540\ \text{kg}}{\frac{3{,}275+3{,}925}{2}\ \text{kg}} = \frac{0{,}540\ \text{kg}}{3{,}600\ \text{kg}} = 0{,}15 = 15\,\% \tag{14.42}$$

Heterosis individual es la diferencia en producción de los individuos cruza en relación con los individuos puros contemporáneos. La vamos a representar, más adelante en nuestros modelos con h^I para diferenciarla de otras formas de heterosis que vamos a incluir en los modelos.
La forma de calcularla es la misma que utilizamos en la ecuación (14.37), es decir

$$h_{AB} = \frac{\bar{X}_{AB} + \bar{X}_{BA}}{2} - \frac{\bar{X}_{AA} + \bar{X}_{BB}}{2} \tag{14.43}$$

Heterosis materna se refiere a la diferencia **en la producción de los individuos** que es **atribuible al uso de madres cruza** (AB) en lugar de madres puras (AA o BB). Este tipo de heterosis **se manifiesta en los hijos de madres cruzas**, a través de un aumento en la producción de leche, un mejor ambiente prenatal, mayor habilidad materna, etcétera. La vamos a notar en muestro modelos como h^M y la calcularemos como

$$h^M_{AB} = \frac{\bar{X}_{C(AB)} + \bar{X}_{C(BA)}}{2} - \frac{\bar{X}_{C(AA)} + \bar{X}_{C(BB)}}{2} \qquad (14.44)$$

En la ecuación anterior se puede observar que para calcularla necesitamos de una tercera raza, **C** en este caso, que servirá de raza del padre en todos los casos. Entenderemos mejor cómo funcionan estas comparaciones cuando veamos el modelo genético de cruzamientos. Un ejemplo claro de heterosis materna es cuando a causa de la heterosis las madres de terneros son más vigorosas, producen más leche o son más resistentes a parásitos, lo que se traduce en un mejor «ambiente de crecimiento» para los terneros.

Heterosis paterna es la ventaja que se obtiene al usar padres cruzas en lugar de padres de raza pura, medido como desempeño de la progenie. En general no se incluye en los modelos ya que no resultan tan obvias las causas por detrás de estas diferencias. Las mismas podrían deberse a una mayor fertilidad, calidad de semen, libido, etcétera. lo que implicaría una mayor eficiencia reproductiva del sistema de cría que utiliza padres cruzas.

PARA RECORDAR

- La **exocría** es el **proceso reproductivo** donde el **apareamiento se produce entre individuos menos emparentados que la media de la población total.** Siempre trataremos de comparaciones entre poblaciones en la misma generación, ya que así eliminamos otros efectos del ambiente o de selección.
- Definimos **heterosis** como la **diferencia entre el fenotipo promedio de las cruzas (recíprocas) y el promedio de las poblaciones «puras» que las originaron.** Entonces, siendo **A** y **B** dos razas diferentes la heterosis ($h_{AB} = h_{BA}$) entre las mismas será $h_{AB} = \frac{\bar{X}_{AB}+\bar{X}_{BA}}{2} - \frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}$
- Podemos **estandarizar la heterosis** para hacerla comparable entre características expresándola en referencia al promedio de las razas puras tal que $h_{AB\,\%} = \frac{h_{AB}}{\frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}} \times 100 = \frac{\frac{\bar{X}_{AB}+\bar{X}_{BA}}{2} - \frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}}{\frac{\bar{X}_{AA}+\bar{X}_{BB}}{2}} \times 100$

- La **heterosis individual** (h^I) es la **diferencia en producción de los individuos cruza en relación con los individuos puros contemporáneos** y se calcula con las mismas ecuaciones de los puntos previos.
- La **heterosis materna** (h^M) se refiere a la **diferencia en la producción de los individuos que es atribuible al uso de madres cruza (AB) en lugar de madres puras (AA o BB)**. Este tipo de heterosis se manifiesta en los hijos de madres cruzas y se calcula como $h^M_{AB} = \frac{\bar{X}_{C(AB)}+\bar{X}_{C(BA)}}{2} - \frac{\bar{X}_{C(AA)}+\bar{X}_{C(BB)}}{2}$ donde **C** es una tercera raza que utilizaremos como padre en todos los casos.
- La **heterosis paterna** en general no se incluye en los modelos ya que no resultan tan obvias las causas por detrás de estas diferencias.
- Para tener una representación gráfica de la heterosis es importante revisar las figuras de la siguiente sección, «**El modelo del templo griego**».

El modelo del templo griego

Como vimos en el capítulo «El modelo genético básico», para el modelo de un locus con dos alelos y apareamientos que cumplen **Hardy-Weinberg**, la media de la población en la escala transformada era igual a $M = a(p-q) + 2pqd$. A un nivel determinado de dominancia la media es una función de la proporción de heterocigotos, el $2pq$ del segundo término del lado derecho de la igualdad. Más aún, si consideramos fijas las frecuencias de los alelos, p y q, la media es una función lineal del grado de dominancia en el locus en cuestión, cosa que puede extenderse fácilmente a muchos loci en ausencia de epistasis.

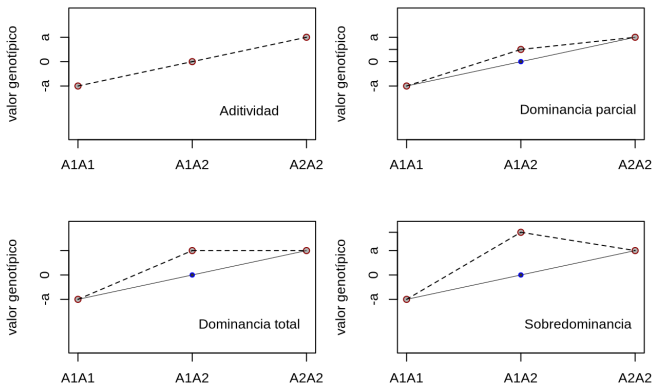


Figura 14.6. Representación de los valores genotípicos correspondiente a los diferentes tipos de dominancia en un locus con dos alelos.

La figura 14.6 representa los diferentes niveles de dominancia que cambian la denominación del tipo de acción génica y que como vimos en la sección «Depresión endogámica», pueden estar detrás de los mecanismos que explican la depresión endogámica.

Cuando realizamos un cruzamiento entre razas, las frecuencias de los alelos en los distintos locus serán el promedio de las frecuencias en cada una de las razas, ya que los individuos cruza tendrán exactamente la mitad del genoma nuclear proveniente de una raza y la otra mitad de la otra raza. Cunningham (1987) plantea la posibilidad de representar la heterosis a partir de un modelo gráfico sencillo, que llamó de **«templo griego»** ya que en algunos casos parece la vista frontal de un templo griego (aunque con menos pretensiones también podría haber sido llamado «modelo de la casita», entre otros muchos nombres posibles). La idea aparece representada en la figura 14.7, donde el cruzamiento de dos razas **A** y **B**, produce individuos cruza **AB**. Los fenotipos de cada genotipo se representan en el eje de las ordenadas. La «proporción de cada raza» (en el genoma) de los individuos cruza es $\frac{1}{2}$. Si todos los genes que participan en la definición del fenotipo tuviesen un modo de acción completamente aditivo, de acuerdo lo que vimos en la figura 14.6, el fenotipo esperable para los individuos cruza estaría en la mitad exacta en $P1$ y $P2$, es decir, en $\frac{P1+P2}{2}$ («prom.» en la figura). Sin embargo, el fenotipo observado de las cruzas se encuentra por encima de ese valor, en $F1$. Si volvemos a mirar la figura 14.6, esto nos recuerda de alguna manera a la «dominancia parcial». Claramente, los efectos aditivos no explican totalmente el fenotipo promedio de los individuos cruza. La diferencia entre el fenotipo promedio observado para los individuos cruza y el promedio de las «razas parentales», es decir $F1 - \frac{P1+P2}{2}$, es la **heterosis observada**.

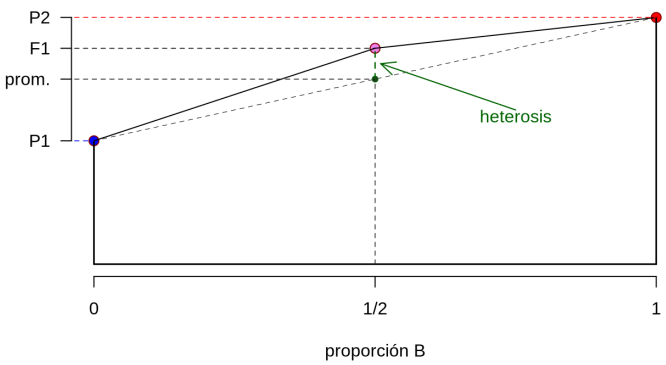


Figura 14.7. Modelo del templo griego para el cruzamiento de dos razas, **A** y **B**. En el eje de las abscisas se representa la proporción de la raza **B** en cada uno de los genotipos (0 para los individuos de la raza **A**, $\frac{1}{2}$ para los **AB** y 1 para los **B**). En el eje de las ordenadas se representan los valores fenotípicos del promedio de los individuos de cada genotipo. Si todos los loci que participan en la característica fuesen aditivos, la media de los individuos cruza debería estar en el promedio (punto de color verde oscuro). En realidad, el promedio de los individuos cruza se encuentra en $F1$ y **la diferencia** $F1 - \frac{P1+P2}{2}$ es **la heterosis observada** (segmento en verde oscuro).

El modelo del templo griego nos permite representar de forma gráfica el comportamiento de diferentes cruzas, así como también ayuda a entender lo que ocurre en determinadas situaciones. La figura 14.8 representa un ejemplo tomado de Cunningham (1987) de interacción entre el ambiente y los efectos aditivos, pero también con los de heterosis. La figura

representa el comportamiento de dos razas (por ejemplo, en producción lechera), una local, la raza **A** y una raza exótica **B**, en dos ambientes contrastantes: a la izquierda en el ambiente favorable y a la derecha en el ambiente desfavorable (o pobre). En el ambiente favorable la raza exótica supera ampliamente a la raza local, esencialmente debido a diferencia aditivas, que en este caso alcanzan las 50 unidades (de lo que sea). En ese ambiente, la cruza presenta una heterosis de $H = 10$, bastante menor en comparación a lo que aportan las diferencias entre razas. Esto apunta a que posiblemente la mejor estrategia sea ir sustituyendo a la raza local por la raza exótica (en la medida en que el resto de los factores, incluyendo resistencia a parásitos, a la temperatura, fertilidad, etcétera. lo hagan posible). Por otro lado, en el ambiente pobre las diferencias entre razas son bastante menores, $A = 10$, marcando una interacción **aditivo-ambiental**. A su vez, en este ambiente la heterosis aparece como muy importante, con $H = 50$, marcando una interacción **heterosis-ambiental**. En este segundo caso resulta mucho más claro que las **F1** son una excelente apuesta. Sin embargo, en muchos casos resulta impráctico trabajar con las **F1** ya que se debe mantener una estructura de al menos dos rodeos, uno reproductivo (por ejemplo, de la raza **A**, inseminado con la raza **B**) y otro productivo de animales **AB**. En este caso, un cruce rotacional podría ser una alternativa y posiblemente mucho más práctico, el uso de una raza sintética derivada a partir de la libre reproducción de los individuos **F1** entre sí, que se estabiliza en la **F2**, como veremos más adelante.

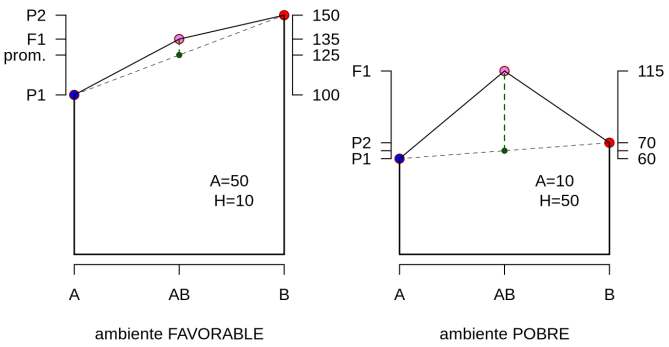


Figura 14.8. Interacción entre ambiente y los efectos aditivos y de heterosis. La figura de la izquierda representa lo observado en un **ambiente favorable**, mientras que la figura de la derecha representa un **ambiente pobre**. Claramente, cambian los valores aditivos relativos entre genotipos, pero también cambian los efectos de la heterosis. Los valores de A y H reportados son, respectivamente, las diferencias en valores aditivos entre razas y los efectos de heterosis. Realización propia sobre idea en Cunningham (1987).

Un modelo sencillo de heterosis bajo exocría

Antes manejamos que la heterosis es un fenómeno que podría explicarse en buena medida por el efecto de la dominancia actuando en un locus o más, unido al incremento en la frecuencia de los heterocigotos, pero no hicimos ningún esfuerzo por cuantificar esto a partir de un modelo. Para eso vamos a recurrir a un modelo muy simple y ya conocido por nosotros y luego veremos que la extensión a múltiples loci es trivial bajo determinados supuestos.

Supongamos que tenemos dos poblaciones, una **P1** con frecuencias p y q para los alelos A_1 y A_2, respectivamente y la otra **P2** con frecuencias p' y q' para los mismos alelos. Además, como siempre, al haber solo dos alelos se cumple $p+q=1$ (población **P1**) y $p'+q'=1$ (población **P2**). Supongamos que la diferencia de frecuencias para el alelo A_1 entre las dos poblaciones es igual a y, es decir $p-p'=y$. Por lo tanto, de acuerdo con todo lo anterior

$$\begin{aligned} p-p'=y \Leftrightarrow p'=p-y \\ p'+q'=1 \Leftrightarrow q'=1-(p-y) \Leftrightarrow q'=(1-p)+y \Leftrightarrow q'=q+y \end{aligned} \tag{14.45}$$

Supongamos ahora que **cruzamos** las dos poblaciones, machos de una con hembras de la otra y viceversa. Por lo tanto, usando los resultados de la ecuación (14.45) podemos representar el cruzamiento de la siguiente forma

Población P2/Población P1	Alelo A_1, frecuencia p	Alelo A_2, frecuencia q
alelo A_1, frec. $p-y$	A_1A_1 frecuencia $p(p-y)$	A_1A_2 frecuencia $q(p-y)$
alelo A_2, frec. $q+y$	A_2A_1 frecuencia $p(q+y)$	A_2A_2 frecuencia $q(q+y)$

De acuerdo con esto, las frecuencias de los tres genotipos en la descendencia del cruzamiento (que es una **F1**) serán

$$fr(A_1A_1)=p(p-y)=p^2-py \tag{14.46}$$

$$fr(A_1A_2)=q(p-y)+p(q+y)=pq-qy+pq+py=2pq+(p-q)y \tag{14.47}$$

$$fr(A_2A_2)=q(q+y)=q^2+qy \tag{14.48}$$

Mientras tanto, el promedio de las frecuencias de $P1$ y $P2$ serán

$$fr(A_1A_1)=\frac{p^2+(p')^2}{2}=\frac{p^2+(p-y)^2}{2}=\frac{p^2+p^2-2py+y^2}{2} \quad \therefore \tag{14.49}$$

$$fr(A_1A_1)=p^2-py+\frac{y^2}{2} \tag{14.50}$$

$$fr(A_1A_2)=\frac{2pq+2p'q'}{2}=pq+p'q'=pq+(p-y)(q+y)=pq+pq+py-qy-y^2 \quad \therefore \tag{14.51}$$

$$fr(A_1A_2)=2pq+(p-q)y-y^2 \tag{14.52}$$

$$fr(A_2A_2)=\frac{q^2+(q')^2}{2}=\frac{q^2+(q+y)^2}{2}=\frac{q^2+q^2+2qy+y^2}{2} \quad \therefore \tag{14.53}$$

$$fr(A_2A_2) = q^2 + qy + \frac{y^2}{2} \tag{14.54}$$

Por lo tanto, si comparamos las frecuencias de los 3 genotipos, antes (promediando **P1** y **P2**) y después del cruzamiento tenemos la siguiente tabla

Genotipo	Frecuencia F1	Frecuencia $(\mathbf{P1}+\mathbf{P2})/2$	Diferencia
A_1A_1	$p^2 - py$	$p^2 - py + \frac{y^2}{2}$	$-\frac{y^2}{2}$
A_1A_2	$2pq + (p-q)y$	$2pq + (p-q)y - y^2$	$+y^2$
A_2A_2	$q^2 + qy$	$q^2 + qy + \frac{y^2}{2}$	$-\frac{y^2}{2}$

De acuerdo con esto, como y^2 es siempre positivo, es decir, no importa si $p > p'$ $(y > 0)$ o si por el contrario $p' > p$ $(y < 0)$, entonces en la **F1** los homocigotos pierden $-\frac{y^2}{2}$ cada uno y por lo tanto los heterocigotos ganan y^2 (todo respecto al promedio de $P1$ y $P2$). Claramente, lo observado tiene una relación directa con el efecto Wahlund que discutimos previamente en el capítulo «Apareamientos no-aleatorios». Como en cierto sentido la heterosis es el fenómeno inverso de la depresión endogámica, este incremento en los heterocigotos estará directamente asociado a un incremento de la media en la **F1**.

Más formalmente, supongamos que en ese locus los valores de los genotipos transformados son a, d y $-a$. Además, de la definición de heterosis en la ecuación (14.37), tenemos que $h_{F1} = F1 - \frac{P1+P2}{2}$, donde $F1$ es la multiplicación de cada frecuencia por su correspondiente valor genotípico y lo mismo para $\frac{P1+P2}{2}$. Como para cada genotipo multiplicamos la correspondiente frecuencia por el mismo valor genotípico (por ejemplo, a para A_1A_1, tanto en $F1$ como en $\frac{P1+P2}{2}$), entonces el valor genotípico de la diferencia de valores genotípicos será igual a la diferencia de frecuencias multiplicada por el valor del genotipo correspondiente.

De acuerdo con esto, usando el resultado de última columna de la tabla anterior, la heterosis será igual a

$$h_{F1} = a\,(-\frac{y^2}{2}) + d\,(+y^2) - a(-\frac{y^2}{2}) = -a\frac{y^2}{2} + d\,y^2 + a\frac{y^2}{2} \;\therefore \tag{14.55}$$

$$h_{F1} = y^2 d \tag{14.56}$$

Por otra parte, si recordamos la definición de $y = (p - p')$, entonces resulta trivial generalizar el resultado de la heterosis a la situación en que varios loci controlan la característica en ausencia de epistasis, ya que en ese caso

$$h_{F1} = \sum_i y_i^2\, d_i = \sum_i (p_i - p_i')^2\, d_i \tag{14.57}$$

Claramente, cuanto mayor sea la diferencia $(p-p')$ mayor será el término $(p-p')^2$ y mayor la heterosis, a un mismo nivel de d (es decir, el mismo valor de dominancia). Por lo tanto, la heterosis depende en este modelo tanto de los efectos de dominancia como de las diferencias en las frecuencias alélicas entre las poblaciones que se cruzan. Si las mismas tienen las mismas frecuencias, entonces $p_i = p'_i \ \forall i$ y $(p_i - p'_i)^2 = 0 \ \forall i$, es decir que $h_{F1} = 0$ y no existirá heterosis. En general, como hemos visto en el capítulo Deriva Genética, las frecuencias alélicas derivan en el tiempo y tenderán a separarse entre poblaciones a medida que transcurra el tiempo (evolutivo). Esto sin descartar la existencia de otros fenómenos evolutivos, como la selección (natural o artificial), las migraciones o como las mutaciones, que también pueden separar las frecuencias alélicas. Por lo tanto, solo como una guía general, **cuanto más distantes entre sí** sean las razas que participan en un cruzamiento **mayor será la heterosis observada**.

La pregunta que nos podemos hacer en este momento es si de aquí en más, el efecto observado de la heterosis se conservará indefinidamente cuando la nueva población se reproduzca entre sí. Analicemos el caso para un locus, ya que luego la extensión es trivial a muchos loci. Para esto, como ahora se trata de un simple apareamiento siguiendo la ley de Hardy-Weinberg, alcanza con hacer el producto de las correspondientes frecuencias alélicas para obtener las frecuencias de los genotipos. Las frecuencias alélicas en la **F2** serán

$$\begin{aligned} fr(A_1) &= \frac{p+p'}{2} = \frac{p+(p-y)}{2} = p - \frac{y}{2} \\ fr(A_2) &= \frac{q+q'}{2} = \frac{q+(q+y)}{2} = q + \frac{y}{2} \end{aligned} \tag{14.58}$$

De acuerdo con esto, las frecuencias genotípicas en la **F2** serán

$$fr(A_1A_1) = \left(p - \frac{y}{2}\right)^2 = p^2 - py + \frac{y^2}{4} \tag{14.59}$$

$$fr(A_1A_2) = 2\left(p - \frac{y}{2}\right)\left(q + \frac{y}{2}\right) = 2pq + py - qy - \frac{y^2}{4} \tag{14.60}$$

$$fr(A_2A_2) = \left(q + \frac{y}{2}\right)^2 = q^2 + py + \frac{y^2}{4} \tag{14.61}$$

y ahora la tabla de diferencias de frecuencias respecto al promedio de las poblaciones **P1** y **P2** será

Genotipo	Frecuencia F2	Frecuencia $(\mathbf{P1}+\mathbf{P2})/2$	Diferencia
A_1A_1	$p^2 - py + \frac{y^2}{4}$	$p^2 - py + \frac{y^2}{2}$	$-\frac{y^2}{4}$
A_1A_2	$2pq + py - qy - \frac{y^2}{4}$	$2pq + (p-q)y - y^2$	$+\frac{y^2}{2}$
A_2A_2	$q^2 + py + \frac{y^2}{4}$	$q^2 + qy + \frac{y^2}{2}$	$-\frac{y^2}{4}$

Multiplicando estas frecuencias por los valores a, d y $-a$, tenemos ahora que la heterosis de la **F2** será

$$h_{F2} = a\,(-\frac{y^2}{4}) + d\,(\frac{y^2}{2}) - a(-\frac{y^2}{4}) = -a\frac{y^2}{4} + d\,\frac{y^2}{2} + a\frac{y^2}{4} \therefore \quad h_{F2} = \frac{y^2 d}{2} \tag{14.62}$$

y como $h_{F1} = y^2 d$, o sea, el numerador de la ecuación (14.62), entonces $h_{F2} = \frac{h_{F1}}{2}$. La generalización a múltiples loci es inmediata, al menos en ausencia de epistasis. En palabras, la heterosis retenida en la **F2** es la mitad de la heterosis observada en la **F1**. A partir de este momento, como la población se encuentra en equilibrio de Hardy-Weinberg, las frecuencias genotípicas observadas ya no diferirán de las descritas en la ecuación (14.61) y, por lo tanto, las diferencias respecto al promedio de **P1** y **P2** se mantendrán, es decir, se retendrá esa proporción de la heterosis inicial (la mitad de la heterosis de la **F1**). En este caso, si **P1** y **P2** son dos razas diferentes, si continuamos reproduciendo la misma población cruza entre **P1** y **P2**, a partir de la **F2** podríamos considerar que se trataría de una raza sintética, con un 50 % de los efectos aditivos de cada una de las dos razas y la mitad de la heterosis máxima retenida.

Una observación general que se deriva del origen mecanístico de la depresión endogámica y de la heterosis es la que hace a la relación entre estas y la heredabilidad. Como el mecanismo subyacente a la depresión endogámica y a la heterosis es la presencia de dominancia (y eventualmente epistasis), cuando estas juegan un rol relativamente importante en la varianza genética, a una proporción fija de varianza genética habrá un descenso concomitante en la importancia de la varianza aditiva. Como la heredabilidad (en sentido estricto) la definimos como la relación entre la varianza aditiva y la varianza fenotípica (total), entonces **las características con alta y muy alta heredabilidad tienden a sufrir poca depresión endogámica y a mostrar poca heterosis**, mientras que **en características que muestran depresión endogámica importante y heterosis importante**, por lo general **la heredabilidad será baja.** Se trata de una regla grosera, ya que en casos de baja varianza ambiental teóricamente podríamos tener alta heredabilidad y también importante depresión endogámica (en presencia de consanguinidad).

Ejemplo 14.6

Una característica se encuentra determinada por dos loci (**A** y **B**), en los que ambos tienen dos alelos. En una población, las frecuencias en los dos loci son $p_{A_1} = 0{,}75$ para el alelo A_1 y $p_{B_1} = 0{,}25$ para el alelo B_1. En una segunda población, las frecuencias son $p'_{A_1} = 0{,}45$ para el alelo A_1 y $p'_{B_1} = 0{,}85$ para el alelo B_1. Si los valores de dominancia para los dos loci son respectivamente $d_A = 12$ unidades y $d_B = 8$ unidades, determinar la heterosis en la generación **F1**.

De acuerdo con la ecuación (14.57), la heterosis para esta característica estaría dada por

$$h_{F1} = \sum_i y_i^2 \, d_i = \sum_i (p_i - p'_i)^2 \, d_i \tag{14.63}$$

Si ponemos los valores de frecuencias y dominancias en una tabla y realizamos los cálculos, llegamos a que

locus	Frec. Pob. 1 (p)	Frec. Pob. 2 (p')	$y = p - p'$	y^2	d_i	$y^2 d_i$
B	0,25	0,85	−0,60	0,36	$8extun-$	2,88
TOTAL	–	–	–	–	–	3,96

Por lo tanto, de acuerdo con estos valores de frecuencias, la heterosis en la **F1** será de 3,96 unidades de la característica.

PARA RECORDAR

- Si partimos de dos poblaciones, una **P1** con frecuencias p y q para los alelos A_1 y A_2, respectivamente y la otra **P2** con frecuencias $p\text{'}$ y $q\text{'}$ para los mismos alelos y denominando a la **diferencia de frecuencias para el alelo A_1 entre las dos poblaciones** es igual a y, es decir $p - p\text{'} = y$. Las frecuencias de los tres genotipos de la descendencia del cruzamiento (**F1**) serán $fr(A_1A_1) = p(p-y) = p^2 - py$
$fr(A_1A_2) = q(p-y) + p(q+y) = pq - qy + pq + py = 2pq + (p-q)y$
$fr(A_2A_2) = q(q+y) = q^2 + qy$

- Como en cierto sentido la **heterosis es el fenómeno inverso de la depresión endogámica**, este incremento en los heterocigotos estará directamente asociado a un incremento de la media en la **F1**.
- Podemos expresar la **heterosis de la F1** (h_{F1}) tal que $h_{F1} = y^2 d$ y podemos generalizar el resultado de la heterosis a la situación en que varios loci controlan la característica en ausencia de epistasis tal que $h_{F1} = \sum_i y_i^2 \; d_i = \sum_i (p_i - p_i')^2 \; d_i$
- Como vemos, la heterosis depende en este modelo tanto de los efectos de dominancia como de las diferencias en las frecuencias alélicas entre las poblaciones que se cruzan lo que nos lleva a pensar que cuanto más distantes entre sí sean las razas que participan en un cruzamiento mayor será la heterosis observada.
- Como la **heredabilidad (en sentido estricto) la definimos como la relación entre la varianza aditiva y la varianza fenotípica (total)**, entonces las **características con alta y muy alta heredabilidad tienden a sufrir poca depresión endogámica y a mostrar poca heterosis**, mientras que en características que muestran **depresión endogámica importante y heterosis importante, por lo general la heredabilidad será baja.**

Complementariedad

Cada una de las razas para una especie dada presenta algunas características para la que es ventajosa y otras características para las que no lo es tanto, o es simplemente deficiente. Una forma adicional de aprovechar los cruzamientos es usando la complementariedad que existe entre las razas a cruzar.

Por ejemplo, existen determinadas razas con fuertes habilidades maternas, es decir, capacidad de hacer crecer mejor a sus (posiblemente más) crías. En esto influyen diferentes factores genéticos, como la fertilidad, la capacidad de producir leche, la capacidad de buscar refugio adecuado o de protegerlas de depredadores. Posiblemente estas razas no posean los mejores atributos desde el punto de vista de la producción terminal de carne y sean superadas por otras razas que poseen, por ejemplo, mayor velocidad de crecimiento. En este sentido, se puede utilizar la complementariedad entre razas en un cruzamiento, eligiendo madres de la primera raza y padres de otra raza de buena calidad de carne y alta velocidad de crecimiento. Dicho de otra forma, la complementariedad entre razas se explotará cuando se puedan combinar estos recursos genéticos (cada uno con aptitudes diferentes) dentro de un mismo sistema de producción manteniendo o aumentando la eficiencia de dicho sistema.

Tipos de cruzamientos

Existen diferentes tipos de cruzamiento, cada uno de ellos explotando diferentes efectos en diferentes proporciones. Además, cada tipo de cruzamiento presenta dificultades diferentes a la hora de implementarlo, principalmente debido al mantenimiento de diferentes poblaciones que coexisten, tanto en lo requerido de instalaciones como en los requerimientos de manejo. En muchos casos implica además una separación entre rodeo reproductivo y productivo, lo que lleva a la posibilidad de considerar las razas sintéticas como alternativa. Existen una infinidad de sistemas de cruzamiento propuestos y nosotros solo mencionaremos unos pocos para visualizar las ventajas y complejidades involucradas.

Cruzamiento simple: Es el cruzamiento más simple posible y consiste en aparear hembras de una raza con machos de otra raza. Permite utilizar además de los efectos raciales la complementariedad y la heterosis del cruzamiento. Un ejemplo típico de esto es cuando se elige como raza de las hembras una raza con habilidad materna, por ejemplo, cerdas **pampa** y machos de una raza terminal como la **duroc**.

Cruzamiento rotacional: Es un cruzamiento relativamente sencillo de implementar, donde al principio hembras de una raza se aparean con machos de otra. La descendencia hembras de este cruzamiento se aparean ahora con machos de la otra raza respecto a los primeros machos. Luego, la descendencia de esta cruza se apareará con machos del mismo tipo que la primera vez. Es decir, vamos alternando la raza de machos sobre la población de hembras cruza. Este tipo de cruza permite explotar 2/3 de la heterosis individual y materna. La principal desventaja de este sistema es la complejidad de mantener dos rodeos y además mantener identificada la paternidad de cada animal. Como principal ventaja la posibilidad de utilizar la heterosis individual y maternal en todo el rodeo, aunque no en su máximo potencial. Además, como es un sistema en el que siempre alternan los efectos raciales es necesario trabajar con razas relativamente compatibles en tamaño y otras características.

Cruzamiento triple: En este sistema se trabaja con tres razas, una terminal (T) y dos razas para producir características favorables y complementarias a las de la raza terminal. La idea es que el apareamiento final sea $T \times (AB) = T(AB)$. Para que funcione es necesario mantener un rodeo de la raza A, para lo que se apareará con toros A y también con toros B para producir los híbridos (AB) que finalmente se aparearán con T. Permite vender diferentes tipos de productos, por ejemplo, terneros machos A y (AB) y todos lo generado de T(AB). Presenta una importante complejidad y necesidad de buenos parámetros reproductivos para poder mantener la producción de reemplazos A e híbridos (AB).

Razas sintéticas: Se trata de una alternativa muy práctica cuando se consigue un producto razonable y adecuado a los sistemas de producción. Se trata de crear una combinación de razas que a partir de determinado momento se empiezan a reproducir dentro de la misma población. No utilizan al máximo la heterosis y se pierde el efecto de la complementariedad, pero pueden constituir una excelente alternativa cuando no se cuenta con instalaciones ni personal para manejos más complicados. Presentan el riesgo de un incremento importante en la consanguinidad si el rodeo se cierra muy rápido y no se permiten ingresar «sangres» diferentes. Entre otros aspectos interesantes de este tipo de estrategia está la posibilidad de incorporar características de resistencia a parásitos y climáticas en razas de alta producción. Ejemplo de esto son la raza barzona (africander, hereford, beef shorthorn y angus), la santa gertrudis (la idea fue de crearla con 37 % cebuina y 62 %; entre 1923 y 1935 fue usado en forma extensiva un toro llamado **Monkey** y todo el ganado actual desciende de él) o la lechera girolando (5/8 holstein y 3/8 gyr, una raza cebuina).

14.5 Modelo genético de cruzamientos

Hasta acá, en lo que hace a exocría y cruzamientos, hemos visto los principios generales y que ocurre en particular cuando cruzamos dos poblaciones que presentan diferencias en las frecuencias alélicas. Sin embargo, aún no hemos desarrollado un modelo que nos permita entender diferentes tipos de cruzamientos e interpretar los parámetros estimados a partir del modelo. Claramente, esta es la forma de poder tomar decisiones importantes acerca de la mejor estrategia para el uso de la genética como herramienta para mejorar la producción, por lo que ahora supliremos esta carencia desarrollando un **modelo genético básico para los cruzamientos.**

Su desarrollo se debe a Dickerson (1973) y es rico en extremo, por lo que en general seguiremos su notación, de forma que el artículo mencionado quede como material de consulta adicional. Si llamamos $\bar{G}$ a la media de la población o cruza considerada, G_I a los efectos directos del individuo (también llamados individuales) y G_M a los efectos maternales, entonces podemos establecer un modelo básico para cruzamientos que sería

$$\bar{G} = \mu + G_I + G_M \tag{14.64}$$

También se podrían incorporar otros efectos en el modelo, por ejemplo G_P para los efectos paternales, efectos G_{AM} del abuelo materno o aún efectos de herencia citoplasmática (por ejemplo, la herencia mitocondrial, que es por vía materna generalmente). Dado que en general se trata de efectos relativamente menores desde el punto de vista conceptual y mecanístico y que la estimación de parámetros adicionales (con datos limitados) implica una pérdida de robustez de los modelos, generalmente no se incluyen en los modelos.

Un tema que no hemos tratado hasta aquí y que apenas mencionaremos por completitud es el de las pérdidas por recombinación que aparecen en los cruzamientos. En muchos casos, las razas se encuentran adaptadas y eso implica combinaciones de alelos en loci no ligados que mejoran la característica por medio de la epistasis. Cuando cruzamos individuos de diferentes razas, se pueden producir en estos recombinaciones a nivel de gametos que rompan esa asociación entre alelos favorables que existía en las razas puras. Esto será lo que llamaremos **pérdidas por recombinación.**

En lo que sigue veremos un gran número de ecuaciones que involucran términos con subíndices y superíndices, por lo que vamos a aclarar la notación que usaremos. En la siguiente tabla se presentan los componentes del modelo genético que podemos utilizar, aunque en general los correspondientes a la última columna no los vamos a incluir:

Efectos	Individual	Maternal	Paternal
Aditivos	g^I	g^M	g^P
Heterosis	h^I	h^M	h^P
Pérdidas recombinación	r^I	r^M	r^P

Observando la tabla anterior queda claro entonces que usaremos los **superíndices** para indicar el **tipo de efecto, individual, maternal, paternal**, al que nos referimos.

En cualquier tipo de cruzamiento hay dos o más razas en consideración. Cuando aparecen más de dos razas esto plantea un tema de posibles combinaciones de efectos, por ejemplo, si tenemos tres razas involucradas **A**, **B** y **C**, entonces podemos tener heterosis **AB**, heterosis **AC** y heterosis del tipo **BC**. Como las mismas no tienen por qué ser iguales, por ejemplo **A** y **B** pueden ser dos razas locales muy próximas y **C** una exótica, necesitamos poder distinguir a qué efecto particular, **raza o combinación de razas**, nos estamos refiriendo y para eso usaremos los **subíndices**. Es importante notar que en el caso de k razas, los tipos de heterosis a considerar están dados por las combinaciones **de a par** entre ellas, que son por lo tanto $k(k-1)$. ¿Por qué no consideramos combinaciones de orden mayor, por ejemplo, heterosis **ABC**? No las consideramos porque en principio, en cada posición (locus) del genoma solo pueden existir como máximo dos alelos diferentes, en organismos diploides que son sobre los que desarrollaremos nuestro modelo.

Por otra parte, además de la presencia o ausencia de los efectos que mencionamos antes, debemos saber en qué proporción se encuentran presentes los mismos en un determinado cruzamiento. Por lo tanto, para las proporciones de cada tipo de efecto usaremos coeficientes, que denominaremos de acuerdo con la siguiente tabla:

Coeficientes de efectos	Individual	Maternal	Paternal
Aditivo raza i	k_i^I	k_i^M	k_i^P
Heterosis razas i,j	$k_{i,j}^I$	$k_{i,j}^M$	$k_{i,j}^P$
Pérdidas Recombinación razas i,j	$w_{i,j}^I$	$w_{i,j}^M$	$w_{i,j}^P$

La idea es sencilla y se puede representar con una analogía alcohólica. Disponemos de una serie de «bebidas base», por ejemplo, whisky, vodka, ron, gin, jugos de naranja, jugo de frutilla, agua tónica, así como hielo, etcétera. Las bebidas base son los efectos genéticos del

modelo. Si nos piden un gin tonic, debemos mezclar gin y agua tónica. ¿En qué proporción debemos mezclarlos? Esos son los coeficientes en nuestro modelo. Claramente, podemos tomar whisky solo o podemos también tomar agua mineral sola, pero el twice up es una mezcla de iguales proporciones de ambos (50 % y 50 %) y el mizuwari japonés es una mezcla de 1/3 whisky con 2/3 de agua.

Es fundamental entender las **diferencias** entre los **efectos causales** que participan en el modelo de un cruzamiento dado **y los correspondientes coeficientes.** Cuáles efectos participan en determinado cruzamiento y sus correspondientes coeficientes se puede determinar del análisis teórico del cruzamiento. La **estimación de los efectos** correspondientes es una tarea posterior a establecer el modelo y **se requiere de datos reales** (observaciones) para poder realizar las estimaciones.

La pregunta siguiente es ¿cómo se calculan estos coeficientes? Comencemos por el más sencillo, el coeficiente que hace al aporte de una raza i a los efectos directos de los individuos cruza. Claramente, en cualquier cruzamiento, los padres aportan la mitad del genoma de las cruzas y las madres hacen otro tanto. Por esta razón, la proporción de la raza i en las cruzas será la suma de la mitad de proporción de la raza i en el padre y la mitad de la proporción de la raza i en la madre. Esto no es otra cosa que el promedio de la proporción en que la raza i aparece en las «razas» (poblaciones) paterna y materna. En términos matemáticos pues, definimos el coeficiente del efecto genético directo de la raza i en las cruzas como

$$k_i^I = \frac{1}{2}k_i^P + \frac{1}{2}k_i^M = \frac{1}{2}(k_i^P + k_i^M) \tag{14.65}$$

sujeto a la restricción obvia de que, para n razas

$$\sum_i^n k_i^I = 1 \tag{14.66}$$

o lo que es lo mismo, la suma de proporciones de las distintas razas en los individuos cruza debe ser 1.

Por otra parte, la heterosis es el resultado de la combinación de pares de razas en cada locus del genoma, por lo que para dos razas i y j en particular, las combinaciones que pueden dar esta heterosis son recibir un alelo correspondiente a la raza i del padre y uno de la raza j de la madre, o lo opuesto, recibir uno de la raza i de la madre y uno de la raza j del padre.

Las probabilidades de estas dos combinaciones son en cada caso el producto de las proporciones raciales correspondientes en las razas (poblaciones) paternas y maternas, por lo que en términos matemáticos definimos el coeficiente de heterosis entre las razas i y j (siempre con $i \neq j$) como

$$k_{ij}^{I} = k_i^P \; k_j^M + k_i^M \; k_j^P \tag{14.67}$$

Claramente, todos los conceptos anteriores y sus formulaciones matemáticas se puede extender en forma trivial a los efectos maternales, donde ahora para mayor claridad usaremos cuando corresponde los superíndices ^{Pm} y ^{Mm} para referirnos al padre de la madre y a la madre de la madre en consideración, respectivamente. Por lo tanto, para el coeficiente del efecto maternal de la raza i, su expresión matemática sería

$$k_i^M = \frac{1}{2} k_i^{Pm} + \frac{1}{2} k_i^{Mm} = \frac{1}{2}(k_i^{Pm} + k_i^{Mm}) \tag{14.68}$$

nuevamente, con la restricción obvia de que, para n razas

$$\sum_{i}^{n} k_i^M = 1 \tag{14.69}$$

o lo que es lo mismo, la suma de proporciones de las distintas razas en la madre debe ser 1.

En el caso del coeficiente de heterosis materna, ahora en principio necesitamos conocer las razas que participaron en la creación de la madre, cosa que muchas veces se puede deducir simplemente de conocer la composición racial de la madre. En particular, hay que tener en cuenta que el orden de las letras, arbitrario, en combinaciones raciales es primero padre y luego madre, usando paréntesis cuando sea necesario para describir cruzamiento que involucran otras cruzas previas. Con la notación que definimos previamente, el coeficiente de heterosis materna se calcula como

$$k_{ij}^{M} = k_i^{Pm} \; k_j^{Mm} + k_i^{Mm} \; k_j^{Pm} \tag{14.70}$$

Finalmente, nos queda el tema del cálculo de los coeficientes de pérdidas por recombinación. Los parámetros r pretenden medir las desviaciones de la asociación lineal de la heterosis con el grado de heterocigosidad. Estos coeficientes describen la fracción media de pares de loci que segregan independientemente en gametos de ambos padres que se espera que sean combinaciones no parentales.

Mencionaremos aquí solo el cálculo del coeficiente correspondiente al individuo, ya que los otros se calculan en forma análoga. La probabilidad de que se rompan las combinaciones favorables será la suma de los productos de la recombinación en padres y madres, es decir

$$w_{ij}^I = k_i^P \; k_j^P + k_i^M \; k_j^M \tag{14.71}$$

Con un poco de álgebra, la ecuación anterior se puede escribir como

$$w_{ij}^I = 4 \; k_i^I \; k_j^I - k_{ij}^I \tag{14.72}$$

ya que

$$4k_i^I k_j^I - k_{ij}^I = 4 \left[\left(\frac{1}{2} k_i^P + \frac{1}{2} k_i^M \right) \left(\frac{1}{2} k_j^P + \frac{1}{2} k_j^M \right) \right] - \left[k_i^P k_j^M + k_i^M k_j^P \right] = \tag{14.73}$$

$$k_i^P k_j^P + k_i^P k_j^M + k_i^M k_j^P + k_i^M k_j^M - k_i^P k_j^M - k_i^M k_j^P = k_i^P k_j^P + k_i^M k_j^M = w_{ij}^I \tag{14.74}$$

lo que la deja en función de los coeficientes de los efectos directos de las razas i y j, así como del coeficiente de heterosis entre las razas i y j.

Veamos ahora cómo expresar el modelo genético de cruzamientos en función de los efectos y de sus coeficientes respectivos. Para los efectos individuales tenemos que, sumando todos los componentes («aditivos», de heterosis y de pérdidas por recombinación), podemos expresarlos como

$$G_I = \sum_i k_i^I \; g_i^I + \sum_{i<j} k_{ij}^I \; h_{ij}^I + \sum_{i<j} [4 \; k_i^I \; k_j^I - k_{ij}^I] \; r_{ij}^I \tag{14.75}$$

De la misma manera, para los efectos maternales, sumando todos los aportes tenemos

$$G_M = \sum_i k_i^M \; g_i^M + \sum_{i<j} k_{ij}^M \; h_{ij}^M + \sum_{i<j} [4 \; k_i^M \; k_j^M - k_{ij}^M] \; r_{ij}^M \tag{14.76}$$

Claramente, recordando la ecuación inicial $\bar{G} = \mu + G_I + G_M$, si solo consideramos los efectos directos y los maternales alcanza con sumar ambos aportes a la media de población para describir el producto de una cruza.

Ejemplo 14.7

Veamos un ejemplo sencillo, que servirá para entender cómo especificamos un modelo, obtenemos los coeficientes correspondientes a los efectos que se incluyen y obtenemos los estimados. A fin de intentar mejorar el peso en animales adultos, se realizó un cruzamiento entre nuestras ovejas verdes (raza **V**) y las scottish blackface de las Hébridas Exteriores (**S**). El cruzamiento lo podemos representar por $V \times S$ y el cruzamiento recíproco por $S \times V$. Esto nos dejó con la siguiente tabla de pesos promedios

Machos/Hembras	S	V
S	$SS = 90$ kg	$SV = 84$ kg
V	$VS = 76$ kg	$VV = 60$ kg

Lo primero es determinar el modelo del cruzamiento desde el punto de vista de los efectos causales.

Un animal puro **S** será hijo de puros **S** y por lo tanto $k_S^I = 1$ (o sea, será totalmente **S** desde el punto de vista racial). Lo mismo, para los puros **V**, $k_V^I = 1$ ya que no hay ninguna otra raza en su constitución. Además, como ambos son «puros» de su respectiva raza, no tendrán heterosis y $k_{SS}^I = k_{VV}^I = 0$. Para los efectos maternales ocurre exactamente lo mismo ya que ambas madres de los puros deben ser puras de la misma raza, por lo que $k_S^M = k_V^M = 1$ y la heterosis materna debe ser cero.

En tanto, en el cruzamiento $S \times V = SV$ los coeficientes de efectos raciales individuales, de acuerdo con la ecuación (14.65), serán

$$\begin{aligned} k_S^I &= \frac{1}{2}k_S^P + \frac{1}{2}k_S^M = \frac{1}{2}\,1 + \frac{1}{2}\,0 = \frac{1}{2} \\ k_V^I &= \frac{1}{2}k_V^P + \frac{1}{2}k_V^M = \frac{1}{2}\,0 + \frac{1}{2}\,1 = \frac{1}{2} \end{aligned} \tag{14.77}$$

con $k_S^I + k_V^I = \frac{1}{2} + \frac{1}{2} = 1$, como debería ser. Ahora, para la heterosis individual entre **S** y **V** tenemos, de acuerdo con la ecuación (14.67) que

$$k_{SV}^I = k_S^P\, k_V^M + k_S^M\, k_V^P = (1 \times 1) + (0 \times 0) = 1 \tag{14.78}$$

Los efectos genéticos maternos, como la madre es pura **V** deberán ser $k_V^M = 1$ y al ser pura no tiene heterosis materna.

Para el cruzamiento recíproco, es decir $V \times S = VS$, tenemos que

$$\begin{aligned} k_V^I &= \frac{1}{2}k_V^P + \frac{1}{2}k_V^M = \frac{1}{2}\,1 + \frac{1}{2}\,0 = \frac{1}{2} \\ k_S^I &= \frac{1}{2}k_S^P + \frac{1}{2}k_S^M = \frac{1}{2}\,0 + \frac{1}{2}\,1 = \frac{1}{2} \end{aligned} \tag{14.79}$$

con $k_V^I + k_S^I = \frac{1}{2} + \frac{1}{2} = 1$, como debería ser. Ahora, para la heterosis individual entre **S** y **V** tenemos

$$k_{SV}^I = k_S^P\, k_V^M + k_S^M\, k_V^P = (1 \times 1) + (0 \times 0) = 1 \tag{14.80}$$

que es el mismo coeficiente que antes. Los efectos genéticos maternos, como la madre es pura **S** deberán ser $k_S^M = 1$ y al ser pura no tiene heterosis materna.

Poniendo todo esto junto tenemos la siguiente tabla de coeficientes del modelo

Constitución	g_S^I	g_V^I	h_{SV}^I	g_S^M	g_V^M	h_{SV}^M
SS	1	0	0	1	0	0
VV	0	1	0	0	1	0
SV	$\frac{1}{2}$	$\frac{1}{2}$	1	0	1	0
VS	$\frac{1}{2}$	$\frac{1}{2}$	1	1	0	0

Tenemos estimados los coeficientes del modelo genético de cruzamiento para las 4 constituciones genéticas de muestro experimento. Ahora debemos combinar esto con los datos de las observaciones.

De acuerdo con la definición de heterosis dada en la ecuación (14.37), la heterosis entre **S** y **V** sería

$$h_{SV} = \frac{\bar{X}_{SV} + \bar{X}_{VS}}{2} - \frac{\bar{X}_{SS} + \bar{X}_{VV}}{2} = \\ = \frac{84+76}{2} - \frac{90+60}{2} = \frac{160}{2} - \frac{150}{2} = 80 - 75 = 5 \text{ kg} \tag{14.81}$$

La relación con el modelo genético de cruzamiento es obvia si tratamos la tabla de coeficientes como coeficientes de ecuaciones. El promedio de las dos últimas líneas de la tabla es

Constitución	g_S^I	g_V^I	h_{SV}^I	g_S^M	g_V^M	h_{SV}^M
SV	$\frac{1}{2}$	$\frac{1}{2}$	1	0	1	0
VS	$\frac{1}{2}$	$\frac{1}{2}$	1	1	0	0
Promedio	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0

que podemos representar como

$$\frac{SV+VS}{2} = \frac{1}{2}g_S^I + \frac{1}{2}g_V^I + h_{SV}^I + \frac{1}{2}g_S^M + \frac{1}{2}g_V^M = 80 \text{ kg} \tag{14.82}$$

De la misma manera, el promedio de las puras (las dos primeras líneas de la tabla de coeficientes) será

Constitución	g_S^I	g_V^I	h_{SV}^I	g_S^M	g_V^M	h_{SV}^M
SS	1	0	0	1	0	0
VV	0	1	0	0	1	0
Promedio	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0

lo que nos deja con un modelo del promedio

$$\frac{SS+VV}{2} = \frac{1}{2}g_S^I + \frac{1}{2}g_V^I + \frac{1}{2}g_S^M + \frac{1}{2}g_V^M = 75 \text{ kg} \tag{14.83}$$

Ahora, si le restamos al promedio de las cruzas que aparece en la ecuación (14.82) recíprocas el promedio de las puras que aparece en la ecuación (14.83), tenemos la siguiente ecuación

$$\frac{SV+VS}{2} = \frac{1}{2}g_S^I + \frac{1}{2}g_V^I + h_{SV}^I + \frac{1}{2}g_S^M + \frac{1}{2}g_V^M = 80 \text{ kg} \tag{14.84}$$

$$-\left[\frac{SS+VV}{2} = \frac{1}{2}g_S^I + \frac{1}{2}g_V^I + \ \ 0 \ + \frac{1}{2}g_S^M + \frac{1}{2}g_V^M = 75 \text{ kg}\right] \tag{14.85}$$

$$h_{SV}^I = 80 - 75 = 5 \text{ kg} \tag{14.86}$$

que es el mismo resultado obtenido antes a partir de la ecuación (14.37).

Vamos a utilizar esta misma aproximación de operación sobre las ecuaciones que describen a cada constitución genética para estimar ahora los efectos genéticos maternos. Si observamos con cuidado la constitución genética de las cruzas recíprocas enseguida nos percatamos que

la única diferencia aparece en los coeficientes de los efectos maternales, por lo que si las restamos tenemos

Constitución	g_S^I	g_V^I	h_{SV}^I	g_S^M	g_V^M	h_{SV}^M
SV	$\frac{1}{2}$	$\frac{1}{2}$	1	0	1	0
VS	$\frac{1}{2}$	$\frac{1}{2}$	1	1	0	0
Diferencia	0	0	0	-1	1	0

que se traduce en

$$SV - VS = -g_S^M + g_V^M = g_V^M - g_S^M = 84 - 76 = 8 \text{ kg} \tag{14.87}$$

Ahora, de acuerdo con la ecuación (14.87) tenemos dos parámetros a determinar y un solo valor numérico, por lo que la única opción que nos queda para levantar esta indeterminación es tomar una resolución arbitraria, por ejemplo, asumir que $g_V^M + g_S^M = 0 \Leftrightarrow g_V^M = -g_S^M$. Sustituyendo en la ecuación (14.87), tenemos ahora que

$$\begin{aligned} g_V^M - g_S^M = g_V^M - (-g_V^M) = g_V^M + g_V^M = 2\, g_V^M = 8 \text{ kg} \therefore \\ g_V^M = \frac{8}{2} = 4 \text{ kg} \end{aligned} \tag{14.88}$$

y por lo tanto $g_S^M = -g_V^M = -4$ kg

Para estimar los efectos directos de raza alcanza con observar los coeficientes de las razas puras. Por ejemplo, para **S** tenemos

$$SS = 1\, g_S^I + 1\, g_S^M \Leftrightarrow g_S^I = SS - g_S^M = 90 - (-4) = 94 \text{ kg} \tag{14.89}$$

y para **V**

$$VV = 1\, g_V^I + 1\, g_V^M \Leftrightarrow g_V^I = VV - g_V^M = 60 - 4 = 56 \text{ kg} \tag{14.90}$$

PARA RECORDAR

- Una forma adicional de aprovechar los cruzamientos es usando la **complementariedad que existe entre las razas a cruzar**, ella se explotará cuando se puedan combinar estos recursos genéticos (cada uno con aptitudes diferentes) dentro de un mismo sistema de producción manteniendo o aumentando la eficiencia de dicho sistema.
- El **cruzamiento simple** consiste en aparear hembras de una raza con machos de otra raza. Este sistema permite utilizar además de los efectos raciales la complementariedad y la heterosis del cruzamiento.
- En el **cruzamiento rotacional** al principio hembras de una raza se aparean con machos de otra y las hijas producto de esta cruza se aparearán siempre con machos de la raza opuesta a la de su padre. La principal desventaja de este sistema es la complejidad de mantener dos rodeos y además mantener identificada la paternidad de cada animal. Como principal ventaja la posibilidad de utilizar la heterosis individual y maternal en todo el rodeo, aunque no en su máximo potencial. Además, como es un sistema en el que siempre alternan los efectos raciales es necesario trabajar con razas relativamente compatibles en tamaño y otras características.

- El **cruzamiento triple** implica el uso de tres razas, una terminal (T) y dos razas para producir características favorables y complementarias a las de la raza terminal. La idea es que el apareamiento final sea T × (AB) = T(AB). Este sistema presenta una importante complejidad y necesidad de buenos parámetros reproductivos para poder mantener la producción de reemplazos A e híbridos (AB).
- Cuando se crea una **raza sintética** se trata de crear una combinación de razas que a partir de determinado momento se empiezan a reproducir dentro de la misma población, estos sistemas no utilizan al máximo la heterosis y se pierde el efecto de la complementariedad a la vez que presentan el riesgo de un incremento importante en la consanguinidad si el rodeo se cierra muy rápido y no se permiten ingresar «sangres» diferentes. Sin embargo, son una excelente alternativa cuando no se cuenta con instalaciones ni personal para manejos más complicados.
- Podemos desarrollar un **modelo genético básico para los cruzamientos**, donde llamaremos $\bar{G}$ a la media de la población o cruza considerada, G_I a los efectos directos del individuo (también llamados individuales) y G_M a los efectos maternales, entonces el modelo básico para cruzamientos sería $\bar{G} = \mu + G_I + G_M$
- Cuando cruzamos individuos de diferentes razas, se pueden producir en estos recombinaciones a nivel de gametos que rompan esa asociación entre alelos favorables que existía en las razas puras. Esto será lo que llamaremos **pérdidas por recombinación.**
- Para trabajar con los diferentes efectos y coeficientes, utilizaremos **superíndices para indicar el tipo de efecto, individual, maternal, paternal, al que nos referimos** y **subíndices para indicar a qué raza o razas corresponde cada efecto o coeficiente.**
- Es muy importante **no confundir las diferencias entre los efectos causales que participan en el modelo de un cruzamiento dado y los correspondientes coeficientes.** La estimación de los efectos correspondientes es una tarea posterior a establecer el modelo y se requiere de datos reales (observaciones) para poder realizar las estimaciones.
- El **coeficiente del efecto genético directo de la raza** i **y en las cruzas** será $k_i^I = \frac{1}{2}k_i^P + \frac{1}{2}k_i^M = \frac{1}{2}(k_i^P + k_i^M)$, donde para n razas, **la suma de proporciones de las distintas razas en los individuos cruza deben ser 1.**
- Podemos definir el **coeficiente de heterosis** entre las razas i y j como $k_{ij}^I = k_i^P\ k_j^M + k_i^M\ k_j^P$
- Para el **coeficiente del efecto maternal de la raza** i, su expresión matemática sería $k_i^M = \frac{1}{2}k_i^{Pm} + \frac{1}{2}k_i^{Mm} = \frac{1}{2}(k_i^{Pm} + k_i^{Mm})$ donde nuevamente para n razas, **la suma de proporciones de las distintas razas en los individuos cruza deben ser 1.**
- El **coeficiente de heterosis materna** se calcula como $k_{ij}^M = k_i^{Pm}\ k_j^{Mm} + k_i^{Mm}\ k_j^{Pm}$
- La probabilidad de que se rompan las combinaciones favorables será la suma de los productos de la recombinación en padres y madres y se puede expresar como $w_{ij}^I = 4\ k_i^I\ k_j^I - k_{ij}^I$

14.6 Conclusión

En el capítulo «Apareamientos no-aleatorios» vimos cómo la existencia de estructuración poblacional lleva a la divergencia de las frecuencias génicas en un ensemble de subpoblaciones meramente por el efecto del azar en el muestreo de gametos. En este capítulo retomamos esta noción, pero haciendo énfasis en el estudio de cómo este fenómeno afecta a caracteres cuantitativos, noción importante si tenemos en cuenta que los apareamientos no se dan al azar en ningún establecimiento productivo. Además de introducir la noción de consanguinidad y su uso para el análisis de la proporción de material genético compartido por individuos donde conocemos el parentesco, se introdujo una visión conceptual al asunto desde la estadística: a medida que en un ensemble poblacional se diferencian diferentes subpoblaciones, se da una conversión de la varianza genética. Así, la varianza intrapoblacional disminuye (conforme se homogenizan las frecuencias génicas en distintos loci), aumentando a expensas de esta la varianza intrapoblacional (justamente reflejando la mayor diferencia entre subpoblaciones). No es de extrañar que hayamos dejado este tipo de análisis para la sección de genética cuantitativa de este libro. Es una noción común que la existencia de apareamientos no-azarosos lleva al fenómeno de *depresión endogámica*, el cual puede ser sumamente perjudicial en un contexto productivo. Lo que es más, introducimos en este capítulo también la noción de *heterosis* (o *vigor híbrido*), donde se observa el fenómeno opuesto. Utilizando el marco conceptual y analítico de los últimos capítulos analizamos estos fenómenos empleando el «modelo del templo griego». Nuevamente vemos cómo las nociones presentadas por la genética de poblaciones resultan de importancia tanto a nivel teórico como práctico, dando así un significado renovado a las nociones adquiridas en el primer tramo de este libro.

14.7 Actividades

14.7.1 Control de lectura

1. En establecimientos productivos es común que exista un desbalance entre el número de machos y el número de hembras. ¿Cómo afecta esto la dinámica evolutiva de estas poblaciones? Desarrolla teniendo en cuenta los conceptos de tamaño poblacional efectivo y consanguinidad.
2. ¿Qué es la depresión endogámica? ¿Qué causas mecanísticas se han planteado para explicar este fenómeno?
3. Define el concepto de heterosis, así como su vínculo con el concepto de exocría. Contrasta con el concepto de depresión endogámica.
4. ¿Qué relación existe entre el concepto de heredabilidad y los fenómenos de depresión endogámica y heterosis?
5. ¿Con qué fin se realiza un cruzamiento? Menciona algunos tipos de cruzamiento, mencionando ventajas y desventajas en cada caso.

14.7.2 ¿Verdadero o falso?

1. El parámetro M es el número total de machos de una población, mientras que H corresponde al número total de hembras de una población.
2. Por definición un individuo en el cual se observa el fenómeno de heterosis presenta un valor fenotípico mayor al de sus progenitores.
3. El incremento anual en la consanguinidad está dado por $\Delta F = \frac{m+h}{8mh\mathrm{IG}^2}$. Podemos obtener un resultado aproximado considerando solo el número de machos (m), tal que $\Delta F = \frac{1}{8m\mathrm{IG}^2}$. Esta última aproximación se vuelve más exacta a medida que $N_e \to 0$.
4. El efecto de la depresión endogámica tiene su máximo cuando $p = q$ (si tratamos un locus bialélico).
5. Matemáticamente, la heterosis se define como $h_{AB} = \frac{\bar{X}_{AB} - \bar{X}_{BA}}{2} - \frac{\bar{X}_{AA} - \bar{X}_{BB}}{2}$, donde $\bar{X}_{..}$ son las medias fenotípicas de cada combinación de razas.

Soluciones

1. **Falso.** Recordemos que $M = m \times \mathrm{IG}$, donde m es el número de machos que ingresan a la población por año e IG corresponde al intervalo generacional. Lo mismo acontece con H, por razonamiento análogo.
2. **Falso.** Es importante recordar que el fenómeno de heterosis refiere al despliegue de un fenotipo mayor al promedio de los progenitores (es decir, alguno de los progenitores puede tener de todas formas mayor valor para el carácter estudiado que su progenie).
3. **Falso.** Se puede constatar tomando el límite de la expresión original que la aproximación es válida en la medida que $h >> m$, lo cual no necesariamente implica $N_e \to 0$ (de hecho, se puede dar esta última condición y que $m >> h$, caso cual la aseveración es falsa).

4. **Verdadero.** Esto se puede ver teniendo en cuenta que la suma ponderada de heterocigosis en una población equivale a $a(\bar{p} - \bar{q}) + 2d\bar{p}\bar{q}(1 - F)$. Esta relación también deja en evidencia la dependencia de la heterocigosis respecto a tres factores: tamaño poblacional (ya que la consanguinidad depende de esta), frecuencias alélicas y modo de dominancia.

5. **Verdadero.** Es importante recalcar que las cruzas entre razas en estos casos se hacen con individuos pertenecientes a una misma generación (a diferencia de cruzas contempladas en otros capítulos donde se combinan distintas generaciones emparentadas).

14.7.3 Ejercicios

Ejercicio 14.1

Tres loci bialélicos (que llamaremos A, B y C) se encuentran segregando en una población de *D. melanogaster* de $N_e = 243$. Los tres alelos influyen en el largo de alas (carácter que llamaremos l). En un momento inicial el valor medio en la población para l es de $M_0 = 2{,}6$ mm. Cien generaciones después las frecuencias alélicas cambian, y por lo tanto también varía el valor medio en la población. Los cambios en las frecuencias y los valores de d estimados para cada locus son las mostradas en la tabla a continuación.

	$f(A_1)$	$f(B_1)$	$f(C_1)$
Gen. 0	0,310	0,560	0,500
Gen. 100	0,776	0,278	0,716
d	0,240	0,500	0,340

¿Cuál es el valor medio para largo de ala en la generación final?

Solución

La media poblacional luego del cambio de frecuencias alélicas se calcula según

$$M_F = M_0 - 2F \sum_i p_i q_i d \tag{14.91}$$

donde p_i, q_i y d_i son las frecuencias alélicas y el valor del parámetro d para el respectivo locus i.

Comencemos por calcular el valor del coeficiente de fijación F luego de cien generaciones. Tenemos que

$$F_t = 1 - (1 - \frac{1}{2N_e})^t \Rightarrow \tag{14.92}$$

$$F_{100} = 1 - (1 - \frac{1}{2 \times 243})^{100} \Rightarrow \tag{14.93}$$

$$F_{100} \approx 0{,}186 \tag{14.94}$$

Tenemos entonces

$$M_F = 2{,}6 - 2 \times 0{,}186 \times [(0{,}776 \times 0{,}224 \times 0{,}24) + (0{,}278 \times 0{,}722 \times 0{,}5) + (0{,}716 \times 0{,}284 \times 0{,}34)] \tag{14.95}$$

$$M_F \approx 2{,}52 \text{ mm} \tag{14.96}$$

Ejercicio 14.2

(Inspirado en Wang et al. 2015)

Un estudio analiza el fenómeno de heterosis en *Bombyx mori*, un gusano de seda. Para ello los investigadores analizaron diferentes rasgos asociados al valor productivo de los gusanos, entre ellos «peso total del capullo» y «diámetro de la fibra de seda», observando estas características fenotípicas en progenitores y en descendencia híbrida.

Se asume que

1. El coeficiente de fijación en la población y momento estudiados es de $F = 0{,}12$.
2. El carácter fenotípico «peso total del capullo» se vincula al locus bialélico A ($f(A_1) = 0{,}22$), con media poblacional en progenitores de 2,6 g y $d = -0{,}18$.
3. El carácter fenotípico «diámetro de la fibra de seda» se vincula al locus bialélico B ($f(B_1) = 0{,}47$), con media poblacional en progenitores de $11 \mu M$ y $d = -0{,}22$.

¿Cuál de los dos caracteres fenotípicos presenta mayor heterosis?

Solución

Calcularemos las medias fenotípicas en la descendencia híbrida y a partir de este dato obtendremos la desviación obtenida respecto a la media de los progenitores.

Recordemos que

$$M_F = M_0 - 2F \sum_i p_i q_i d \Rightarrow \tag{14.97}$$

$$M_F - M_0 = -2F \sum_i p_i q_i d \tag{14.98}$$

Tenemos para el carácter fenotípico «peso total del capullo»

$$M_F - M_0 = 2 \times 0{,}12 \times 0{,}22 \times 0{,}78 \times 0{,}18 \tag{14.99}$$

La desviación es entonces de

$$\frac{2 \times 0{,}12 \times 0{,}22 \times 0{,}78 \times 0{,}18}{2{,}6} \times 100\,\% \approx 0{,}29\,\% \tag{14.100}$$

Para el caso de «diámetro de la fibra de seda», por razonamiento análogo, se tiene una desviación de

$$\frac{2 \times 0{,}12 \times 0{,}47 \times 0{,}53 \times 0{,}22}{11} \times 100\,\% \approx 0{,}12\,\% \tag{14.101}$$

Vemos por lo tanto que el peso total del capullo es la característica fenotípica que presenta mayor efecto de heterosis de los dos.

Ejercicio 14.3

(Adaptado de D. S. Falconer and MacKay 1996)

La varianza genética del número de setas abdominales en *Drosophila* es predominantemente aditiva. Las varianzas fenotípicas y aditivas en dicha población son estimadas en $V_A = 1{,}98$ y $V_P = 3{,}79$, respectivamente. Supongamos que una población grande se subdivide en un ensemble, donde cada subpoblación está compuesta por 20 individuos, y que se estima que estas subpoblaciones continúan reproduciéndose de forma independiente por 25 generaciones. Se estima que las varianzas fenotípicas y aditivo genética luego de dicho proceso son de $V_P = 4{,}0$; $V_A = 2{,}08$ ¿Cuál es el cambio en la heredabilidad de la característica en la población total? ¿Cómo se explica de manera conceptual este cambio en la heredabilidad del carácter fenotípico estudiado?

Solución

Antes de que la población sea subdividida en un ensemble la heredabilidad total corresponde a $h^2 = \frac{V_A}{V_P} = \frac{1{,}98}{3{,}79} \approx 0{,}522$.

Luego de 25 generaciones el coeficiente de fijación estará dado por

$$F = (1 - \frac{1}{2 \times 20}) \tag{14.102}$$

$$F = (1 - \frac{1}{40})^{25} \tag{14.103}$$

Recordemos que la varianza genética entre líneas está dada por $(1 - F) \times V_G = (1 - (1 - \frac{1}{40})^{25}) \times 2{,}08 \approx 0{,}975$, mientras que la varianza total está dada por $(1 + F) \times V_G = (1 + (1 - \frac{1}{40})^{25}) \times 2{,}08 \approx 3{,}185$. La heredabilidad será entonces, luego del proceso de evolución del ensemble de $h^2 \approx \frac{3{,}185}{4{,}0} = 0{,}796$; es decir, la heredabilidad total aumentó $\approx \frac{0{,}796 - 0{,}522}{0{,}522} \times 100\,\% = 52{,}5\,\%$.

¿Cómo podemos explicar esto? Tengamos en cuenta que a medida que las generaciones pasan en cada subpoblación del ensemble se van fijando alternativamente distintos alelos de los loci que controlan la expresión de este carácter fenotípico. Esto se traduce en

una disminución progresiva de la varianza genético-aditiva dentro de cada subpoblación (es decir, dentro de cada población los individuos serán cada vez más homogéneos). Visto de otra forma, la variabilidad fenotípica dentro de cada subpoblación es cada vez menos explicable por una variabilidad genético-aditiva. Al mismo tiempo, a medida que cada subpoblación marca un perfil de alelos para cada locus, la varianza existente en el componente genético aditivo en la población total comienza a ser cada vez mayor: existirá mayor heterogeneidad entre los individuos. Esto implica que las diferencias observadas a nivel fenotípico son cada vez más atribuibles a diferencias en el componente genético aditivo; esto no es más que otra forma de decir que la heredabilidad aumentará para el carácter fenotípico.

Ejercicio 14.4

Un productor criador utiliza cuatro genotipos de bovinos: AA, ZZ y sus cruzas recíprocas AZ y ZA. Al momento de realizar el destete de una generación de terneros, encontró que en promedio, los animales AA pesaban 140 kg, los ZZ 160 kg, los AZ 155 kg y los ZA 165 kg. Con esta información estima:

a. La heterosis individual en valor absoluto entre las razas A y Z. **b.** Los efectos maternales para cada raza. **c.** Los efectos individuales para cada raza.

Solución

a. La heterosis individual será

$$h^I_{AZ} = \frac{\overline{X}_{AZ} + \overline{X}_{ZA}}{2} - \frac{\overline{X}_{AA} + \overline{X}_{ZZ}}{2} = \frac{155 + 165}{2} - \frac{140 + 160}{2} = 10 \text{ kg} \quad (14.104)$$

b. Para calcular los efectos maternales utilizaremos la información de los animales cruza

$$ZA = \frac{1}{2}g^I_Z + \frac{1}{2}g^I_A + 1h^I_{ZA} + 1g^M_A = 165 \quad (14.105)$$

$$AZ = \frac{1}{2}g^I_Z + \frac{1}{2}g^I_A + 1h^I_{ZA} + 1g^M_Z = 155 \quad (14.106)$$

$$ZA - AZ = g^M_A - g^M_Z = 165 - 55 = 10 \quad (14.107)$$

$$g^M_A + g^M_Z = 0 \implies g^M_A = -g^M_Z \quad (14.108)$$

$$10 = -g^M_Z - g^M_Z = 2 \times -g^M_Z \implies -g^M_Z = \frac{10}{2} = 5 \quad (14.109)$$

$$g^M_A = 5 \text{ kg}; g^M_Z = -5 \text{ kg} \quad (14.110)$$

c. Por último, para calcular los efectos individuales deberemos utilizar la información de los animales puros, así como el valor de los efectos maternales tal que

$$AA = g^I_A + g^M_A \implies g^I_A = AA - g^M_A = 140 - 5 = 135 \text{ kg} \quad (14.111)$$

$$ZZ = g^I_Z + g^M_Z \implies g^I_Z = ZZ - g^M_Z = 160 - (-5) = 165 \text{ kg} \quad (14.112)$$

Ejercicio 14.5

Una manada de equinos posee la siguiente estructura:

Edad parto	Hembras	Machos
4	41	8
5	38	8
6	35	7
7	34	4
8	26	4
9	21	4

¿Cuál es la tasa anual de consanguinidad de esta manada?

Solución

En primera instancia, la tasa de consanguinidad anual será

$$\Delta F_A = \frac{\Delta F_G}{IG} \tag{14.113}$$

A la vez que la tasa de consanguinidad generacional (ΔF_G) puede ser calculada como

$$\Delta F_G = \frac{1}{2 \times N_e} \tag{14.114}$$

Por lo que comenzaremos calculando el tamaño efectivo (Ne) de la población

$$N_e = \frac{4 \times M \times H}{M + H} \tag{14.115}$$

$$M = m \times IG \tag{14.116}$$

$$H = h \times IG \tag{14.117}$$

$$IGh = \frac{41 \times 4 + 38 \times 5 + 35 \times 6 + 34 \times 7 + 26 \times 8 + 21 \times 9}{195} = 6{,}15 \tag{14.118}$$

$$IGm = \frac{8 \times 4 + 8 \times 5 + 7 \times 6 + 4 \times 7 + 4 \times 8 + 4 \times 9}{35} = 6{,}00 \tag{14.119}$$

$$IGp = \frac{6{,}15 + 6{,}00}{2} = 6{,}075 \tag{14.120}$$

Entonces

$$M = m \times IG = 8 \times 6{,}076 = 48{,}6 \tag{14.121}$$

$$H = h \times IG = 41 \times 6{,}075 = 249{,}075 \tag{14.122}$$

$$N_e = \frac{4 \times M \times H}{M + H} = \frac{4 \times 48{,}6 \times 249{,}075}{48{,}6 + 249{,}075} = 162{,}66 \tag{14.123}$$

Por lo que la tasa de consanguinidad generacional será

$$\Delta F_G = \frac{1}{2 \times N_e} = \frac{1}{2 \times 162{,}66} = 0{,}003 \tag{14.124}$$

Por último, la tasa de consanguinidad anual será

$$\Delta F_A = \frac{\Delta F_G}{IG} = \frac{0{,}003}{6{,}075} = 0{,}0005 \tag{14.125}$$

Capítulo 15

Normas de reacción e interacción genotipo-ambiente

No efforts of mine could avail to make the book easy reading.[1]

Ronald A. Fisher

En capítulos previos hemos desarrollado los principales conceptos de la genética cuantitativa, así como obtenido diferentes resultados cuantitativos a partir del modelo genético básico. Cuando comenzamos a trabajar con dicho modelo, mencionamos la existencia de la interacción entre genotipo y ambiente, pero decidimos dejarla fuera de la mayor parte de nuestros desarrollos. El argumento usado es que su importancia es relativamente baja en las condiciones normales asociadas a la producción animal (no así en la vegetal), pero en realidad lo que nos motivó a no considerarla son las complicaciones asociadas a su tratamiento. El presente capítulo vamos a dedicarlo entonces a entender el significado biológico de esta interacción, comprender su significado desde el punto de vista de nuestro modelo genético básico, así como delinear algunas implicancias de esta interacción sobre parámetros genéticos como la varianza aditiva.

Para esto vamos a comenzar por mostrar cuán extendido es el alcance de este efecto y que natural nos resulta su existencia cuando pensamos un poco sobre ello. Vamos a definir conceptos claves como el de **plasticidad fenotípica** o el de **norma de reacción** y entender que nuestro concepto de genotipo como algo incambiado y que no reacciona al ambiente es poco natural. Luego pasaremos a analizar formas posibles de estimar su impacto y para eso recurriremos primero al análisis de una característica en dos ambientes. Eso nos remitirá al concepto de respuesta correlacionada, que vimos en el capítulo «Correlaciones y respuesta correlacionada» y la utilidad de la correlación genética como forma de detectar interacciones significativas. Por último, pasaremos a vincular las normas de reacción lineales con los conceptos de genética cuantitativa que vimos en el caso del modelo de un locus con dos alelos y veremos el impacto en la varianza aditiva que puede tener el ambiente.

[1] «The Genetical Theory of Natural Selection», Oxford at the Clarendon Press, 1930. Esta frase aparece en el prefacio, firmado por Fisher, en Rothamsted en junio de 1929. Se trata de una obra monumental que sentó las bases de la genética de poblaciones y de la genética cuantitativa, pero de muy difícil lectura (como gran parte de la obra de Fisher). La traducción nuestra sería algo como «Por más que lo intenté, no conseguí que el libro fuera fácil de leer». Y seguro que, a nuestra infinitesimal escala, desprovistos además de cualquier originalidad, tampoco lo logramos.

OBJETIVOS DEL CAPÍTULO

☐ Analizar la interacción genotipo-ambiente, con perspectiva desde el modelo genético básico.

☐ Introducir conceptos como la plasticidad fenotípica y la norma de reacción.

☐ Presentar métodos para estimar el impacto de la interacción genotipo-ambiente.

☐ Analizar ejemplos particulares de interacción genotipo-ambiente, para tomar dimensión de la diversidad de situaciones en las que el fenómeno es de relevancia.

15.1 Plasticidad fenotípica y normas de reacción

Hasta ahora, en lo que hace a la genética cuantitativa, hemos venido trabajando esencialmente con una versión algo simplificada del modelo genético básico. Si bien establecimos desde un principio que el modelo, además de los efectos genéticos y ambientales, podía incorporar los efectos de la interacción entre genotipo y ambiente, hasta ahora logramos evitar razonablemente tener que lidiar con este último factor, que sin duda complica sobremanera muchas de las derivaciones. De hecho, en gran parte de las situaciones que modelamos no parece tan descabellado el ignorar esta interacción. Por ejemplo, cuando el ambiente es relativamente constante o uniforme para todos los individuos, el término de interacción no tendrá significación o importancia práctica. Por ejemplo, si los tambos de una determinada región se encuentran sometidos a las mismas condiciones geográficas, climáticas, etcétera y además presentan un manejo muy similar de los animales (algo que parece razonable, ya que los mismos precios de insumos y egresos afectan las decisiones de manejo en forma similar), entonces podemos considerar el «ambiente» como relativamente homogéneo; en estas condiciones, no tiene mucho sentido considerar la interacción entre el genotipo y el ambiente ya que el ambiente es casi el mismo para todos los animales.

Sin embargo, en muchas situaciones los ambientes son claramente contrastantes y ya no es posible evitar la consideración de la interacción entre genotipo y ambiente. Siguiendo con el mismo ejemplo, el uso de la raza **holando** (holstein) se encuentra extendido por todo el orbe, debido a su enorme producción. De hecho, el «tipo» de la raza ha ido cambiando a lo largo del siglo XX, principalmente guiado por los criadores de América del Norte hacia animales más grandes y pesados, adaptados a las condiciones productivas de dichos criadores. Mirada como unidad productiva, cuanto más grande la vaca (dentro de la raza holando) mayor la expectativa promedio de producción, lo que ha llevado a animales de gran porte (vacas de más de 650 kg). Estos «genotipos» (en un sentido liberal del término) de animales se encuentran perfectamente adaptados a un ambiente controlado y agradable (en general estabulados buena parte del año), con comida *ad libitum* y un manejo sanitario excelente, lo que permite que una partición de la energía hacia la producción lechera sea favorecida (es decir, seleccionada favorablemente por los productores).

Por otra parte, cuando estos mismos animales deben producir en otros ambientes, por ejemplo, en los trópicos, las condiciones ambientales cambian en forma dramática y las mismas características que eran tan deseables antes ahora parecen claramente perjudiciales. Vacas grandes precisan mucho más alimento que vacas pequeñas y difícilmente puedan mantener una condición corporal razonable ya que además estaban seleccionadas para sacrificar su condición a cambio de la producción de leche. Mucho peor, como mencionamos antes, el calor se genera proporcionalmente al volumen del animal, mientras que se disipa en proporción al área, por lo que animales grandes sufrirán en extremo el calor. Más aún, el manejo sanitario tan cuidadoso que recibieron seguramente evitó el desarrollo de resistencia frente a parásitos, en particular los del nuevo ambiente, que seguramente tiene mayor cantidad y diversidad de los mismos, por lo que apenas podrá mantenerse en condición sanitarias aceptables. Frente a esto, animales de la misma raza, de menor tamaño, criados desde hace varias generaciones en las condiciones de los trópicos seguramente se sentirán más «aclimatados» y adaptados al ambiente exigente, por lo que su producción será razonable, así como su condición corporal y su estado sanitario.

Si por alguna extraña razón llevamos animales holando adaptados al trópico a producir en América del Norte, seguramente su producción se verá algo incrementada, aunque con seguridad no sea ni cerca de la producción de las vacas holando en uso allí ya que es un animal más pequeño. Ahora, cuando comparamos la producción de los dos «genotipos» (el del trópico y el de América del Norte) en los dos ambientes, claramente la diferencia observada en el trópico será mucho menor que la observada en el ambiente favorable. Dicho de otra forma, desde el punto de vista del modelo no nos alcanza con la suma de los efectos genéticos y los ambientales para describir el comportamiento de un animal, ya que el mismo dependerá de la combinación específica de genotipo y ambiente; esto es lo que llamamos **interacción genotipo-ambiente**.

Obviamente que se trata de una situación ejemplo de muchas otras posibles en diferentes tipos de producciones. En general, es más fácil observar el efecto de la interacción entre genotipo y ambiente en plantas que en animales, ya que las primeras tienen muy poca capacidad de escoger individualmente el ambiente. Un animal puede, en cierta medida, elegir un área con sombra en un día de calor, o buscar una aguada cuando siente sed, un refugio cuando siente frío, por lo que intenta «controlar» el ambiente externo de tal forma que se encuentre lo más próximo posible a su zona de confort. En cambio, la planta tendrá mucho menos opciones de ir cambiando el ambiente externo a sus necesidades puntuales; si viene seca tendrá que intentar minimizar los daños, pero lo mismo le ocurrirá en la inundación, sol calcinante o ausencia del mismo, invasión de plagas diversas o lo que sea le depare su existencia. Por lo tanto, en general las plantas se encuentran sometidas a condiciones ambientales externas más extremas que los animales y por lo tanto las diferencias en habilidades en cada uno de estos ambientes serán más marcadas.

Existen dos fenómenos distintos que pueden interferir en la simple adición de los efectos genéticos y ambientales: a) la interacción genotipo-ambiente propiamente dicha y que generalmente notamos $G \times E$ y b) la correlación entre genotipo y ambiente, que notamos como r_{GE}.

Como vimos en la sección «La varianza en el modelo genético básico», desde el punto de vista de la varianza fenotípica, el modelo se puede escribir como

$$V_P = \mathbb{V}(G + E + I_{G»E}) = V_G + V_E + V_{G»E} + 2\text{Cov}_{GE} = V_A + V_D + V_I + V_E + V_{G»E} + 2\text{Cov}_{GE} \quad (15.1)$$

Además de los componentes «tradicionales» de la varianza, es decir V_A, V_D, V_I y V_E, tenemos dos términos adicionales: $V_{G»E}$ y dos veces Cov_{GE}. Este último término, Cov_{GE}, es el que se asocia directamente con la correlación r_{GE} que mencionamos antes. Podemos verlo como un término de «inflación» de la varianza fenotípica observada (cuando es positiva), producto de la covariación en el mismo sentido entre el genotipo y el ambiente. De los cursos de estadística sabemos que $\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y) + 2\ \text{Cov}(x,y)$, por lo que siempre que las dos variables aleatorias covaríen (es decir, sea distinto de cero el valor de la covarianza), en nuestro caso **genotipo** y **ambiente**, la varianza de la suma será diferentes que la suma de las varianzas.

Ejemplos claros de este tipo de efectos es cuando a genotipos más productivos les asignamos ambientes más favorables en forma sistemática, o en el sentido contrario, a los menos productivos les asignamos mejores ambientes. En el primer caso, buscamos explotar al máximo el potencial de cada genotipo, por ejemplo, asignando más alimento a las vacas con «mejor genética» (simplemente porque creemos que deben ser más productivas); esto hará que crezca la brecha entre animales y por lo tanto se incremente la varianza fenotípica. Por otro lado, en el segundo caso buscamos una producción más homogénea, por lo que a plantas con peor genética les asignamos más fertilizantes y riego, resultando en mayor homogeneidad y por lo tanto menor varianza fenotípica.

Desde el punto de vista más agronómico, este tipo de covariación o correlación entre genotipo y ambiente puede considerarse como un «ruido» o «contaminación de la información» que muchas veces se puede eliminar haciendo que todos los genotipos se expresen en todos los ambientes. Por ejemplo, cuando comparamos razas de animales (como genotipos diferentes), en general cada raza se encuentra originalmente en determinados ambientes, pero difícilmente todas las razas estén en todos los ambientes de interés. Alcanzaría en principio con tener registros de todas las razas en los ambientes de interés para romper esa asociación específica de genotipos con ambientes.

Por otra parte, el segundo efecto mencionado, la varianza de la interacción entre genotipo y ambiente es de una base mucho más biológica, ya que lo es la interacción misma y generalmente imposible de eliminar. Cuando pensamos en una característica de interés productivo, por ejemplo, la producción de leche, tendemos a pensar que los genes que la gobiernan son aquellos que tienen que ver con el funcionamiento de la glándula mamaria, cosa que en principio suena razonable. Por lo tanto, estos genes se expresarán no importa en qué ambiente, mientras que la vaca se encuentre en lactancia. Sin embargo, la vaca es la entidad biológica relevante para la producción de leche, no la ubre; la vaca sometida a estrés térmico tendrá disparados un conjunto de mecanismos para la homeostasis que perturbarán enormemente la capacidad de la ubre de utilizar los recursos del animal.

En gran parte de lo que hemos trabajado en la parte de genética cuantitativa nos hemos mantenido con la idea de que el genotipo y el ambiente no presentan interacción, al menos

significativa. Es decir, un genotipo siempre aporta lo mismo al fenotipo de una característica. Sin embargo, como acabamos de ver, esto no siempre es así. De hecho, un genotipo puede reaccionar de forma diferentes a diferentes ambientes o de forma más general a un gradiente ambiental (una variación continua del ambiente, como puede ser la humedad del suelo, el nivel de luz o de radiación solar, etcétera). En general un fenotipo es la respuesta visible a través de una característica particular (producción de algo, largo de otra cosa, color de otra diferente, etcétera) de la expresión de un conjunto de proteínas (estructurales, enzimáticas, etcétera) todas reguladas por un conjunto de genes, que a su vez reciben información del medio externo a través de un conjunto de receptores y de vías de transmisión de la información. Por lo tanto, no debería sorprendernos para nada que un determinado «genotipo» responda en forma diferente, para la misma característica, a diferentes condiciones ambientales. Precisamente, la **plasticidad fenotípica** es la capacidad de un organismo de producir **fenotipos diferentes** en **respuesta a cambios en el ambiente**.

> Se llama **plasticidad fenotípica** a la capacidad de realizar cambios en el comportamiento, la morfología y la fisiología de un organismo en respuesta a cambios en el ambiente. La plasticidad fenotípica abarca todos los tipos de cambios inducidos por el medio ambiente (por ejemplo, morfológicos, fisiológicos, de comportamiento, fenológicos) y que pueden o no ser permanentes a lo largo de la vida de un individuo.

Usualmente es más fácil de observar plasticidad fenotípica en las plantas como mecanismo de adaptación a la variabilidad ambiental a las que se ven sometidas. Ejemplos claros incluyen la asignación de más recursos a las raíces en suelos que contienen bajas concentraciones de nutrientes, el tiempo de floración, así como el área y espesor de las hojas. En algunas plantas, por ejemplo, del género *Convolvulus* González and Gianoli (2004), mientras que las hojas que crecen en ambiente sombreado muestran una gran superficie y son de poco espesor para maximizar la captura de la luz, en ambientes soleados muestran hojas de menor superficie y más gruesas. También el **diente de león** (del género *Taraxacum*) es bien conocido por mostrar una considerable plasticidad en su forma cuando crece en ambientes soleados frente a los sombreados.

El concepto de **plasticidad fenotípica** se encuentra estrechamente relacionado al concepto de **norma de reacción**, que sería la expresión de este concepto para una característica fenotípica determinada, en un genotipo y en gradiente ambiental particular.

> Se conoce como **norma de reacción** a la función que relaciona la expresión fenotípica media de un genotipo en diferentes ambientes. Para cada característica fenotípica, genotipo y variable ambiental puede existir una **norma de reacción** específica.

La norma de reacción de un genotipo dado es su rango de respuestas fenotípicas a lo largo de un gradiente ambiental. El concepto fue introducido por Richard Woltereck en 1909 (Woltereck 1909) en un trabajo con **dafnias** (un género de crustáceos planctónicos del orden Cladocera) y se ilustra en la figura 15.1. En un **gradiente ambiental**, marcado por el eje de

las abscisas, los dos genotipos **A** y **B** reaccionan de manera opuesta en la expresión fenotípica de una característica. Por ejemplo, la característica fenotípica podría ser el rendimiento y el gradiente ambiental podría ser la humedad del suelo. Mientras que para el **genotipo A** el rendimiento aumenta en forma aproximadamente lineal con la humedad, para el **genotipo B** el rendimiento decrece de la misma forma con la humedad. Los conceptos de plasticidad

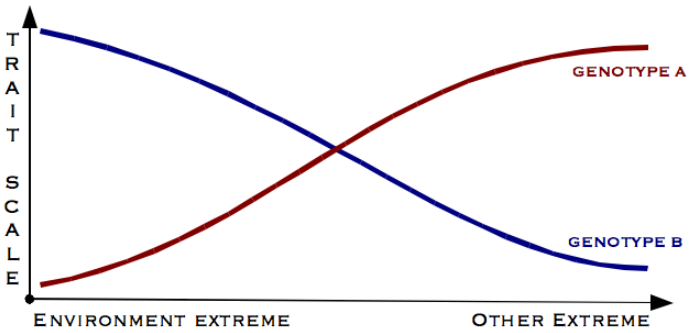


Figura 15.1. Ilustración del concepto de **norma de reacción.** En un **gradiente ambiental** representado en las abscisas, ambos genotipos reaccionan de forma opuesta (y aproximadamente lineal) ya que la respuesta fenotípica varía en sentido opuesto (Dominio público, https://en.wikipedia.org/w/index.php?curid=3925138).

fenotípica y de norma de reacción están estrechamente vinculados al concepto de **canalización**, un término acuñado en 1942 por Conrad Hal Waddington (1942) con la idea de conceptualizar la robustez que demuestran ciertos procesos y reacciones involucradas en el desarrollo de los organismos y que pese a ser apartadas (ligeramente) de las condiciones normales igual llevan al mismo resultado.

> Se llama **canalización** a una medida de la capacidad de una población **para producir el mismo fenotipo independientemente de la variabilidad de su ambiente o genotipo**. Es una forma de **robustez evolutiva**.

De alguna manera, la canalización juega un papel antagónico con la plasticidad fenotípica, amortiguando cambios «ambientales» para que el fenotipo continúe en una región óptima. Waddignton prefirió la utilización de este nuevo término para distinguir la robustez biológica (más ligada al concepto actual de resiliencia) del concepto tradicional en ingeniería. El término proviene de una analogía donde se puede pensar el paisaje epigenético del desarrollo como un valle rodeado por elevaciones y en el que las características canalizadas seguirán el curso a través del valle hacia su «destino» (el fenotipo objetivo). El concepto puede ilustrarse como lo muestra la figura 15.2. Mientras que el **genotipo A** muestra cambios importantes en el fenotipo (ordenadas) de acuerdo con el gradiente ambiental (en las abscisas), el **genotipo B** muestra una distribución bimodal fenotípica, de acuerdo con la región del gradiente ambiental. En cada una de estas regiones, sin embargo, los cambios fenotípicos son mínimos a pesar del cambio amplio en el ambiente. Desde el punto de vista molecular la canalización podría estar actuando como un **capacitor evolutivo**, manteniendo diversos cambios a nivel genómico sin expresión en el fenotipo, hasta que por algún motivo el mismo se descarga (la analogía es con el funcionamiento de un capacitor en los circuitos

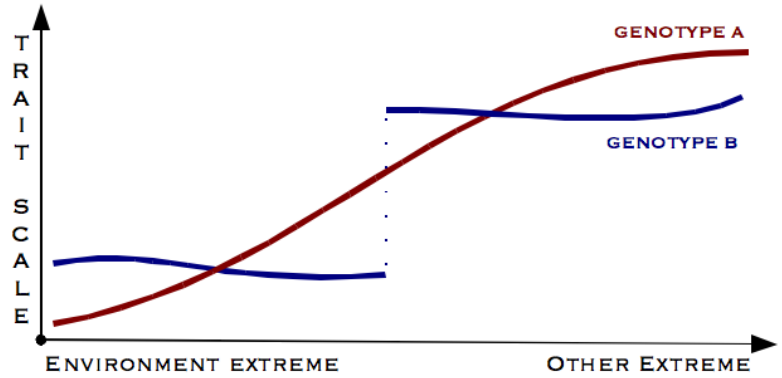


Figura 15.2. Norma de reacción para ilustrar el concepto de canalización. Mientras que el **genotipo A** muestra una amplia respuesta fenotípica al gradiente ambiental, el **genotipo B** muestra dos regiones disjuntas del gradiente en las que, en ambas, el fenotipo no cambia mayormente pese a los cambios ambientales; es decir, el **genotipo B** se encuentra **canalizado.** (Dominio Público, https://commons.wikimedia.org/w/index.php?curid=1223392).

eléctricos, donde el mismo es capaz de almacenar energía mientras se carga). Un ejemplo de un gen con esta función podría ser la proteína **HSP90** (Heat Shock Protein 90, las HSP son un conjunto de proteínas que originalmente se asociaron a la expresión como respuesta al estrés), que cumple una función de «chaperona» (ayudan al plegado de otras proteínas). Se ha demostrado en diversas especies que la mutación de esta proteína está asociada al incremento en la variabilidad observada en diferentes características fenotípicas, que estarían canalizadas por la forma normal de **HSP90**.

PARA RECORDAR

- Existen **dos fenómenos distintos que pueden interferir en la simple adición de los efectos genéticos y ambientales:** *a)* la interacción genotipo-ambiente propiamente dicha y que generalmente notamos *G*x*E* y *b)* la correlación entre genotipo y ambiente, que notamos como r_{GE}.
- En términos de varianzas, podemos escribir ahora el modelo genético básico como $V_P = V_A + V_D + V_I + V_E + V_{G»E} + 2\text{Cov}_{GE}$ dónde el término Cov_{GE} se asocia directamente con la correlación r_{GE}. Podemos verlo como un término de «inflación» de la varianza fenotípica observada (cuando es positiva), producto de la covariación en el mismo sentido entre el genotipo y el ambiente.
- Llamaremos **plasticidad fenotípica** a la capacidad de realizar cambios en el comportamiento, la morfología y la fisiología de un organismo en respuesta a cambios en el ambiente. La plasticidad fenotípica abarca todos los tipos de cambios inducidos por el medio ambiente y que pueden o no ser permanentes a lo largo de la vida de un individuo.
- Se conoce como **norma de reacción** a la función que relaciona la expresión fenotípica media de un genotipo en diferentes ambientes. Para cada característica fenotípica, genotipo y variable ambiental puede existir una norma de reacción específica.

- Llamaremos **canalización** a una medida de la capacidad de una población para producir el mismo fenotipo independientemente de la variabilidad de su ambiente o genotipo. Es una forma de robustez evolutiva. Desde el punto de vista molecular la **canalización podría estar actuando como un capacitor evolutivo**, manteniendo diversos cambios a nivel genómico sin expresión en el fenotipo, hasta que por algún motivo el mismo se descarga y el cambio es apreciable.

15.2 Interacción G×E en dos ambientes

Cuando en lugar de considerar un gradiente ambiental, es decir una variación continua del ambiente, consideramos una serie discreta de ambientes contrastantes, la situación puede simplificarse bastante, particularmente en el caso de dos ambientes. En muchos casos, si bien existe un gradiente ambiental subyacente, nos interesa contrastar dos ambientes en particular, por ejemplo, la expresión de una característica en clima templado versus su expresión en los trópicos. Más específico aún, podemos analizar el uso de una raza cebuina en nuestro país, comparando los resultados de varias razas británicas y de la raza cebuina en nuestro país y en la región de origen de la raza cebuina. Otro ejemplo podría ser analizar la producción de diferentes variedades de maíz en el Medio Oeste estadounidense y en nuestro país.

Un rango de comportamientos posibles de estos experimentos puede apreciarse en la figura 15.3, donde la expresión fenotípica de tres «genotipos» diferentes, identificados por colores diferentes, se muestran en dos ambientes diferentes, uno **pobre** y el otro **favorable**. En la situación representada en (a), los fenotipos correspondientes a los tres genotipos se desplazan la misma cantidad en el ambiente favorable respecto al ambiente pobre, por lo que genotipo y ambiente son aditivos entre ellos (su suma es el fenotipo) y por lo tanto no existe interacción entre genotipo y ambiente. En la situación mostrada en (b), se mantiene el ordenamiento relativo entre los genotipos (el rojo superior al verde y este superior al azul) en los dos ambientes, pero las distancias entre ellos no son idénticas en los dos ambientes, sino que resultan de multiplicar la distancia original por un factor. Por lo tanto, en este caso estamos tratando con un cambio de escala entre los dos ambientes. Por otra parte, la situación representada en (c) implica un cambio en el ordenamiento de los genotipos (las líneas se cruzan), aunque no hay un cambio de escala ya que la varianza fenotípica es igual en ambos ambientes. Por último, la situación representada en (d) implica tanto un reordenamiento de los genotipos como un cambio de escala ya que existe mayor variabilidad fenotípica en el ambiente favorable.

Una simplificación notable en el tratamiento de la interacción entre genotipo y ambiente proviene de la idea de Douglas S. Falconer (1952) de considerar la expresión de la característica en dos ambientes como si se tratase de dos características diferentes y que presentan, por lo tanto, una correlación genética (aditiva). La idea nos remite directamente al *path analysis* de Sewall Wright y a la sección «La relación entre las correlaciones fenotípica, genética y ambiental». La idea se ilustra en la figura 15.4, donde la misma característica se presenta en dos tratamientos diferentes, $T1$ y $T2$. Podemos pensar en estos tratamientos como dos condiciones nutricionales diferentes, o de exposición diferente a desafíos sanitarios,

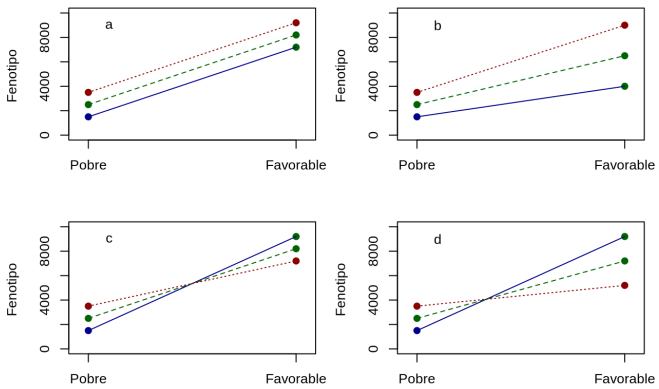


Figura 15.3. Representación esquemática de la interacción genotipo ambiente. Tres genotipos expresan una característica fenotípica cuantitativa (ordenadas) en dos ambientes diferentes (abscisas), un ambiente **pobre** y uno **favorable**. (a) Existe un desplazamiento de la característica entre los dos ambientes que es constante para todos los genotipos (alcanza con sumar la diferencia entre ambientes), por lo que no existe interacción entre genotipo y ambiente. (b) La interacción entre genotipo y ambiente se debe a un cambio de escala. (c) La interacción genotipo-ambiente está asociada a un cambio en el ordenamiento de los genotipos (y a un desplazamiento constante de los ambientes). (d) La interacción genotipo-ambiente está asociada a un reordenamiento y a un cambio de escala.

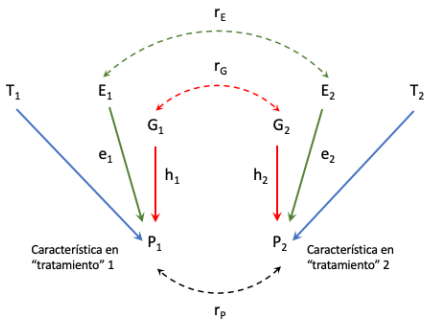


Figura 15.4. Diagrama para el estudio de la interacción genotipo-ambiente en dos ambientes, considerando la misma como una correlación genética entre dos características diferentes. En este diagrama T_1 y T_2 son los dos tratamientos diferentes, que pueden ser considerados como «macroambientes», mientras que E_1 y E_2 son ambientes con una correlación entre ellos establecida en el modelo.

o lo que se nos ocurra como los dos ambientes contrastantes donde queremos explorar el tema de interacción genotipo ambiente.

Si observamos la figura, hay tres fuentes de variación que afectan a cada uno de los dos fenotipos: los dos genotipos, G_1 y G_2; las diferencias ambientales dentro de cada tratamiento, T_1 y T_2, y las diferencias ambientales no asociadas a los tratamientos, E_1 y E_2. Los dos fenotipos, que en realidad es la **misma característica medida en dos ambientes diferentes**, están **correlacionados por una vía genética** y, posiblemente, por una vía ambiental. Por otra parte, los dos genotipos, G_1 y G_2, están conectados por una correlación, r_G.

Como vimos en el capítulo «Correlaciones y respuesta correlacionada», esta correlación genética, junto con los coeficientes de paso, h_1 y h_2 (las raíces cuadradas de las dos heredabilidades), que relacionan el genotipo con el fenotipo son todo lo que se necesita para la solución del problema de determinar el coeficiente de correlación genética, ya que de acuerdo con la ecuación (12.64), si las intensidades de selección son las mismas en las dos características, entonces

$$\frac{\mathrm{RC}_2}{\mathrm{R}_2} = \frac{\mathrm{h}_1}{\mathrm{h}_2}\,\mathrm{r}_{\mathrm{A}_{12}} \tag{15.2}$$

y un estimador razonable del valor absoluto del coeficiente de correlación genética era el dado por la ecuación (15.3)

$$\mathrm{r}^2_{\mathrm{A}_{12}} = \frac{\mathrm{RC}_1}{\mathrm{R}_1}\frac{\mathrm{RC}_2}{\mathrm{R}_2} \;\therefore\; \|\mathrm{r}_{\mathrm{A}_{12}}\| = \sqrt{\frac{\mathrm{RC}_1}{\mathrm{R}_1}\frac{\mathrm{RC}_2}{\mathrm{R}_2}} \tag{15.3}$$

Al tratarse como dos características diferentes, podemos considerar que los ambientes actúan de diferente forma en cada una y lo mismo la genética. Si desde el punto de vista genético el ordenamiento es exactamente igual en los dos ambientes y las distancias proporcionales, entonces **la correlación genética debería ser igual a uno.** Por el contrario, si de alguna manera el orden de los fenotipos se altera o aún las distancias entre ellos dejan de ser proporcionales, entonces la correlación será inferior a uno.

Una de las dificultades mayores con esta idea es la de determinar el fenotipo en dos tratamientos diferentes al mismo tiempo o al menos en tiempos suficientemente similares como que no sean importantes los factores de confusión. Por ejemplo, es imposible por este método determinar la existencia de interacción genotipo-ambiente en peso a los 18 meses para animales criados en confinamiento vs criados a campo, ya que un animal no puede estar en los dos estados (tratamientos) al mismo tiempo (excepto para el **gato de Schrödinger**, pero en este caso es bastante difícil medir los fenotipos). Para otras características, que de alguna manera se repiten en el animal, por ejemplo, las lactancias, podemos ensayar los animales en las dos condiciones, pero resulta un factor importante de confusión la historia previa del animal antes de cada lactancia y además existe el factor de confusión que es el número de lactancia (cuando nos referimos a producción del animal).

Una alternativa razonable es la experimentación con **grupos de individuos** con algún tipo de vinculación de parentesco entre ellos, como la unidad básica, en lugar del individuo. Es decir, podemos dividir a la mitad cada grupo (familia) y dejar una mitad en un tratamiento y la otra mitad en el otro tratamiento. Por comodidad, a veces vamos a llamar macroambiente al tratamiento, ya que se trata de un «ambiente» que afecta a todo un grupo de individuos (el que está sometido a ese tratamiento, que obviamente no es genético). En lo que sigue, en esta y en la próxima sección, vamos a seguir muy estrechamente el tratamiento del tema que le dan Lynch y Walsh (1998), por lo que nuevamente apelaremos a su notación.

Supongamos un experimento en el que pretendemos evaluar la interacción genotipo-ambiente entre dos macroambientes. Para eso, vamos a dividir todos los **grupos genotípicos** de interés (que pueden estar representados, por ejemplo, por familias) en dos, una mitad expresando la característica en un macroambiente, la otra mitad en el otro. En el presente contexto, podemos describir el fenotipo de un individuo que participa en nuestro experimento como

$$z_{ijk} = \mu + G_j + I_{ij} + E_{g,i} + E_{s,ijk} \tag{15.4}$$

donde μ es la media general, promediada sobre todos los genotipos en todos los ambientes, G_j es el efecto del j-ésimo grupo genotípico, promediado sobre todos los ambientes, $E_{g,i}$ es el efecto promedio del i-ésimo ambiente sobre todos los genotipos, $I = ij$ es el efecto de interacción entre el genotipo j y el ambiente i y $E_{s,ijk}$ son los efectos ambientales particulares (error) de la réplica k del genotipo j en el ambiente i. La notación de la letra g en $E_{g,i}$ se refiere a que se trata de un efecto ambiental general (macroambiente, tratamiento T en Douglas S. Falconer 1952), mientras que la letra s en $E_{s,ijk}$ se refiere al ambiente específico (E en Douglas S. Falconer 1952).

De acuerdo con este modelo, el efecto del macroambiente i podemos escribirlo como el desvío de la media de los todos los grupos en dicho macroambiente respecto a la media general, es decir

$$E_{g,i} = \mu_i - \mu \tag{15.5}$$

Del mismo modo, de acuerdo con la definición dada previamente sobre el modelo, el efecto particular de un grupo genético estará dada por el desvío de la media de ese grupo en los dos macroambientes respecto a la media general, es decir

$$G_j = \mu_j - \mu \tag{15.6}$$

Por otra parte, el desvío correspondiente a la interacción entre el genotipo j y el ambiente i es igual a la media de la combinación de dichos genotipos en ese ambiente, respecto a la media general, a la media del genotipo y a la media del ambiente, es decir

$$\begin{aligned} I_{ij} = \mu_{ij} - \mu - E_{g,i} - G_j = \mu_{ij} - \mu - [\mu_i - \mu] - [\mu_j - \mu] = \\ = \mu_{ij} + \mu - \mu_i - \mu_j \end{aligned} \tag{15.7}$$

Cuando no existe interacción entre genotipo y ambiente $I_{ij} = 0$, por lo que

$$I_{ij} = \mu_{ij} + \mu - \mu_i - \mu_j = 0 \Leftrightarrow \mu_{ij} = -\mu + \mu_i + \mu_j \Leftrightarrow \tag{15.8}$$

$$\mu_{ij} = -\mu + \mu_i + \mu_j + \mu - \mu = \mu + (\mu_i - \mu) + (\mu_j - \mu) \Leftrightarrow \tag{15.9}$$

$$\mu_{ij} = \mu + G_j + E_{g,i} \tag{15.10}$$

o lo que es lo mismo, la media del genotipo j en el ambiente i será igual a la suma de la media general, el efecto general del genotipo j y el efecto general del ambiente i. Desde el punto de vista gráfico esto significa que, con dos macroambientes, las líneas que unen a los genotipos en los dos ambientes serán paralelas.

Ejemplo 15.1

En un experimento se determinó que la producción de dos variedades determinadas de maíz en (kg/ha) son funciones lineales de la cantidad de agua disponible en el suelo en un momento específico del crecimiento, al menos en un rango muy estrecho, entre $18\,\% - 22\,\%$. Las funciones que describen las producciones de las dos variedades (**A** y **B**) en función de la humedad del suelo en el momento estudiado son

- **A:** $P_A = [7800 + h\,\% \times 250]$ kg
- **B:** $P_B = [5300 + h\,\% \times 375]$ kg

Graficar la producción de los dos cultivos en el gradiente de humedad establecido. ¿Existe interacción genotipo-ambiente? Determinar el punto en que ambas producciones son iguales. ¿Qué variedad nos conviene más si suponemos un ambiente relativamente seco en esa etapa del cultivo?

Se trata de dos funciones lineales, por lo que alcanza con determinar un par de puntos para cada función y luego unirlos. Por ejemplo, en forma completamente arbitraria podemos elegir los puntos de $19\,\%$ y $21\,\%$, lo que nos deja con los siguientes pares de puntos para **A** y **B**:

Variedad	Humedad	Rendimiento (kg/ha)
A	$19\,\%$	$[7800 + 19 \times 250]$ kg $= 12.550$ kg
A	$21\,\%$	$[7800 + 21 \times 250]$ kg $= 13.050$ kg
B	$19\,\%$	$[5300 + 19 \times 375]$ kg $= 12.425$ kg
B	$21\,\%$	$[5300 + 21 \times 375]$ kg $= 13.175$ kg

La gráfica resultante se puede apreciar en la figura 15.5. Claramente, las dos rectas no son paralelas, para lo que alcanzaba observar que los dos coeficientes de regresión son diferentes ($b_A = 250$, $b_B = 375$, ambas en kg/ %), por lo que se observa interacción genotipo-ambiente en el rango considerado. Pero además, las rectas se cruzan, por lo que a ambos lados del punto de corte el ordenamiento por rendimiento cambia: mientras que a la izquierda del punto de corte la variedad **A** es más productiva, a la derecha del mismo la variedad **B** es la de mayor rendimiento. El punto de corte es fácil de calcular: alcanza igualar ambas rectas y determinar a qué porcentaje de humedad corresponde la solución, es decir

$$P_A = [7800 + h\,\% \times 250]\text{ kg} = [5300 + h\,\% \times 375]\text{ kg} = P_B \Leftrightarrow \tag{15.11}$$

$$(7800 - 5300) = h\,\%(375 - 250) \Leftrightarrow h\,\% = \frac{7800 - 5300}{375 - 250} = \frac{2500}{125} = 20 \tag{15.12}$$

Para verificar alcanza con sustituir $h\,\% = 20\,\%$ en las ecuaciones de rendimiento, lo que nos deja con

$$P_A = [7800 + h\,\% \times 250]\text{ kg} = 12.800\text{ kg} \tag{15.13}$$

$$P_B = [5300 + h\,\% \times 375]\text{ kg} = 12.800\text{ kg} \tag{15.14}$$

La variedad más conveniente en un ambiente relativamente seco parece ser la variedad **A**, ya que es la que muestra mayor rendimiento por debajo de $20\,\%$ de humedad en el suelo.

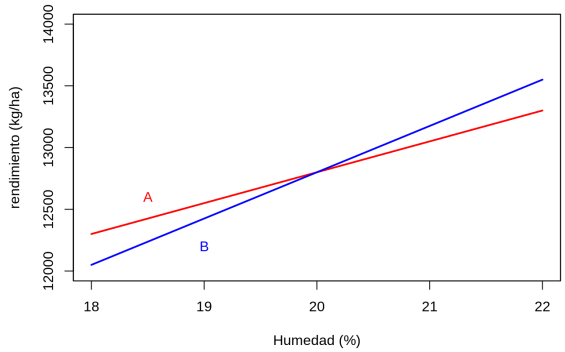


Figura 15.5. Rendimiento en función de la humedad del suelo para dos variedades de maíz, la variedad **A** en rojo y la variedad **B** en azul. Se aprecia claramente que las dos líneas no son paralelas (lo que alcanza para determinar la existencia de interacción genotipo-ambiente), pero, además, estas se cortan en el intervalo considerado, lo que indica un cambio de *ranking* en el mismo.

PARA RECORDAR

- Si bien existe un gradiente ambiental subyacente, usualmente nos interesa contrastar dos (o más) ambientes en particular, considerándolos como una serie discreta.
- Una simplificación plausible en el tratamiento de la interacción entre genotipo y ambiente proviene de la idea de considerar la expresión de la característica en dos ambientes como si se tratase de dos características diferentes y que presentan, por lo tanto, una correlación genética (aditiva).
- Un estimador razonable del valor absoluto del coeficiente de correlación genética será $\|\mathrm{r}_{\mathrm{A}_{12}}\| = \sqrt{\frac{\mathrm{RC}_1}{\mathrm{R}_1}\frac{\mathrm{RC}_2}{\mathrm{R}_2}}$. Al **tratarse como dos características diferentes**, se puede considerar que los ambientes actúan de diferente forma en cada una y lo mismo la genética. **Si desde el punto de vista genético el ordenamiento es exactamente igual en los dos ambientes y las distancias proporcionales, entonces la correlación genética debería ser igual a uno.**
- En ocasiones donde el uso de **un solo individuo** podría dar lugar a errores en el resultado, parece razonable la experimentación con **grupos de individuos con algún tipo de vinculación de parentesco entre ellos como la unidad básica**, en lugar del individuo.
- Si pretendiéramos evaluar la interacción genotipo-ambiente entre dos macroambientes y para ello dividiéramos todos los grupos genotípicos de interés, podríamos describir el fenotipo de un individuo que participa en nuestro experimento como $z_{ijk} = \mu + G_j + I_{ij} + E_{g,i} + E_{s,ijk}$
- De acuerdo con ese modelo, el efecto del macroambiente i podemos escribirlo como el desvío de la media de los todos los grupos en dicho macroambiente respecto a la media general, es decir, $E_{g,i} = \mu_i - \mu$ y consecuentemente, el efecto particular de un grupo genético estará dada por el desvío de la media de ese grupo en los dos macroambientes respecto a la media general tal que $G_j = \mu_j - \mu$

- Además, el desvío correspondiente a la interacción entre el genotipo j y el ambiente i es igual a la media de la combinación de dichos genotipos en ese ambiente, respecto a la media general, a la media del genotipo y a la media del ambiente, es decir, $I_{ij} = \mu_{ij} + \mu - \mu_i - \mu_j$

15.3 Correlación genética a través de dos ambientes

En esta sección seguiremos los desarrollos del libro de Lynch y Walsh (1998). Para simplificar la notación y teniendo en cuenta que en el «error» pueden participar alguna otra fuente genética cuando el fenotipo es de un grupo de individuos, vamos a usar la notación e_{ijk} en lugar de $E_{s,ijk}$. Con esto en mente, los fenotipos de dos individuos k y l, ambos pertenecientes al mismo grupo genotípico j, para la característica de interés en los dos ambientes, 1 y 2 podemos escribirlos como

$$\begin{aligned} z_{1jk} &= \mu + G_j + I_{1j} + E_1 + e_{1jk} \\ z_{2jl} &= \mu + G_j + I_{2j} + E_2 + e_{2jl} \end{aligned} \tag{15.15}$$

Sin embargo, teniendo en cuenta de que se trata de solamente dos ambientes los que estamos comparando y como $E_1 = \mu_1 - \mu$ y $E_2 = \mu_2 - \mu$, al mismo tiempo que $\mu = \frac{\mu_1 + \mu_2}{2}$, entonces

$$\begin{aligned} \mu_1 &= E_1 + \mu \\ \mu_2 &= E_2 + \mu \end{aligned}$$

$$\begin{aligned} \mu = \frac{E_1 + \mu + E_2 + \mu}{2} \Leftrightarrow \mu = \mu + \frac{E_1 + E_2}{2} \Leftrightarrow \\ 0 = \frac{E_1 + E_2}{2} \Leftrightarrow \frac{E_1}{2} = -\frac{E_2}{2} \Leftrightarrow E_1 = -E_2 \end{aligned} \tag{15.16}$$

por lo que vamos a reducir un parámetros en las ecuaciones del modelo (15.15) ya que $E = E_1 = -E_2 \ \therefore \ E_2 = -E$. Por otra parte, $I_{1j} = \mu_{1j} - \mu - G_j - E$ y $I_{2j} = \mu_{2j} - \mu - G_j + E$, además de que $G_j = \frac{(\mu_{1j} + \mu_{2j})}{2} - \mu$. Usando este último resultado, tenemos que

$$\begin{aligned} G_j = \frac{(\mu_{1j} + \mu_{2j})}{2} - \mu \Leftrightarrow 2\,(G_j + \mu) = \mu_{1j} + \mu_{2j} \Leftrightarrow \\ \mu_{1j} = 2G_j + 2\mu - \mu_{2j} \end{aligned} \tag{15.17}$$

y sustituyendo en $I_{1j} = \mu_{1j} - \mu - G_j - E$ nos queda

$$\begin{aligned} I_{1j} = 2G_j + 2\mu - \mu_{2j} - \mu - G_j - E = G_j + \mu - \mu_{2j} - E = \\ = -[\mu_{2j} - \mu - G_j + E] = -I_{2j} \end{aligned} \tag{15.18}$$

por lo que $I_{1j} = -I_{2j} = I_j$. Utilizando los resultados de las ecuaciones (15.16) y (15.18) y sustituyendo en la ecuación (15.15) nos queda una versión más simplificada del modelo

$$\begin{aligned} z_{1jk} &= \mu + G_j + I_j + E + e_{1jk} \\ z_{2jl} &= \mu + G_j - I_j - E + e_{2jl} \end{aligned} \tag{15.19}$$

Las propiedades esenciales de este modelo lineal pueden resumirse de la siguiente manera:

1. La **media total** del modelo, es decir de **todos los genotipos a través de todos los ambientes** es igual a μ.
2. Los **efectos genéticos de los grupos** de individuos (familias) **se asumen aleatorios**, con esperanza de cero y varianza σ^2_G.
3. De la misma manera, **para un ambiente dado los efectos de interacción son variables aleatorias** con esperanza de cero y varianza σ^2_I.
4. **Dentro de los grupos genéticos** el **efecto medio de interacción** está restringido a que debe ser **cero.**
5. Puede existir **correlación entre los valores genotípicos y los efectos de interacción**, en cuyo caso la covarianza es de $\sigma_{G,I}$.
6. **El valor promedio de los dos ambientes es cero.**
7. Las **desviaciones residuales** dentro de los ambientes tienen **esperanza de cero y varianza** σ^2_e para todos los grupos genéticos.

Dentro de cada ambiente, los efectos genotípicos son los directos del grupo genotípico y los de interacción, que como vimos antes estaban restringidos entre sí, por lo que $G_{1j} = G_j + I_j$ y $G_{2j} = G_j - I_j$

Las **varianzas genéticas de los grupos** (familias) en los dos ambientes serán por lo tanto

$$\begin{aligned}\sigma^2_G(1) &= \sigma^2(G_j + I_j) = \sigma^2_G + \sigma^2_I + 2\sigma_{G,I} \\ \sigma^2_G(2) &= \sigma^2(G_j - I_j) = \sigma^2_G + \sigma^2_I - 2\sigma_{G,I}\end{aligned} \tag{15.20}$$

mientras que para la **covarianza genética de los grupos** entre los dos ambientes tenemos

$$\begin{aligned}\sigma_G(1,2) &= \sigma[(G_j + I_j),(G_j - I_j)] \\ &= \sigma(G_j,G_j) + \sigma(G_j,I_j) - \sigma(G_j,I_j) + \sigma(I_j,-I_j) \\ &= \sigma(G_j,G_j) - \sigma(I_j,I_j) \\ &= \sigma^2_G - \sigma^2_I\end{aligned} \tag{15.21}$$

Por otra parte, como vimos en la sección «Semejanza entre parientes» del capítulo «Parentesco y semejanza entre parientes», la varianza entre familias es igual a la covarianza fenotípica entre individuos de la misma familia. Por lo tanto, la contribución **genético-aditiva** a la **varianza genética en los ambientes** 1 y 2, es decir a $\sigma^2_G(1)$ y $\sigma^2_G(2)$ será igual a $2\Theta\sigma^2_A(1) = a_{XY}\sigma^2_A(1)$ y $2\Theta\sigma^2_A(2) = a_{XY}\sigma^2_A(2)$, mientras que el aporte de la covarianza genético-aditiva a la covarianza genética $\sigma_G(1,2)$ será $2\Theta\sigma_A(1,2) = a_{XY}\sigma_A(1,2)$. Θ es el coeficiente de coancestría entre los miembros del grupo familiar, idéntico a lo que en la sección «El concepto de identidad por ascendencia» (IBD) manejamos como f_{XY}, mientras

que a_{XY} es el parentesco aditivo (que en ese contexto notamos con r_{XY}); recordemos que $\Theta = \frac{1}{2}a_{XY}$ o lo que es lo mismo $f_{XY} = \frac{1}{2}r_{XY}$.

A partir de las varianzas y covarianza genético-aditiva, ignorando los efectos genéticos no-aditivos, podemos definir el **coeficiente de correlación genética (aditiva) entre ambientes** como

$$\rho_x = \frac{\sigma_A(1,2)}{\sigma_A(1)\sigma_A(2)} = \frac{2\Theta\ \sigma_A(1,2)}{\sqrt{2\Theta\ \sigma_A^2(1)\ 2\Theta\ \sigma_A^2(2)}} = \frac{\sigma_A(1,2)}{\sigma_A(1)\ \sigma_A(2)} = \frac{\sigma_G(1,2)}{\sigma_G(1)\ \sigma_G(2)} \tag{15.22}$$

La importancia de la ecuación (15.22) radica en que a partir de estimaciones de las varianzas ($\sigma_G^2(1)$ y $\sigma_G^2(2)$) y covarianzas ($\sigma_G(1,2)$) de **familias** podemos estimar la correlación genética de la característica entre ambientes diferentes.

En las ecuaciones anteriores trabajamos con las **varianzas y covarianzas de las medias familiares de grupos genotípicos (familias)**, que tienen componentes de varianza genotípica y de interacción. Ahora, operando con las ecuaciones (15.20) y (15.21) tenemos que podemos obtener la varianza genética como

$$\frac{1}{2}[\frac{\sigma_G^2(1) + \sigma_G^2(2)}{2} + \sigma_G(1,2)] = \tag{15.23}$$

$$\frac{1}{2}[\frac{\sigma_G^2 + \sigma_I^2 + 2\sigma_{G,I} + \sigma_G^2 + \sigma_I^2 - 2\sigma_{G,I}}{2} + \sigma_G^2 - \sigma_I^2] = \tag{15.24}$$

$$\frac{1}{2}[\frac{2\sigma_G^2 + 2\sigma_I^2}{2} + \sigma_G^2 - \sigma_I^2] = \sigma_G^2\ \ \therefore \tag{15.25}$$

$$\sigma_G^2 = \frac{1}{2}[\frac{\sigma_G^2(1) + \sigma_G^2(2)}{2} + \sigma_G(1,2)] \tag{15.26}$$

mientras que la varianza de la interacción la podemos obtener de la siguiente forma

$$\frac{1}{2}[\frac{\sigma_G^2(1) + \sigma_G^2(2)}{2} - \sigma_G(1,2)] = \tag{15.27}$$

$$= \frac{1}{2}[\frac{\sigma_G^2 + \sigma_I^2 + 2\sigma_{G,I} + \sigma_G^2 + \sigma_I^2 - 2\sigma_{G,I}}{2} - \sigma_G^2 + \sigma_I^2] = \tag{15.28}$$

$$= \frac{1}{2}[\frac{2\sigma_G^2 + 2\sigma_I^2}{2} - \sigma_G^2 + \sigma_I^2] = \sigma_I^2\ \ \therefore \tag{15.29}$$

$$\sigma_I^2 = \frac{1}{2}[\frac{\sigma_G^2(1) + \sigma_G^2(2)}{2} - \sigma_G(1,2)] \tag{15.30}$$

De la misma manera, la covarianza entre los efectos genéticos y de interacción se puede obtener a partir de

$$\frac{\sigma_G^2(1) - \sigma_G^2(2)}{4} = \frac{\sigma_G^2 + \sigma_I^2 + 2\sigma_{G,I} - \sigma_G^2 - \sigma_I^2 + 2\sigma_{G,I}}{4} = \frac{4\sigma_{G,I}}{4} = \sigma_{G,I}\ \ \therefore \tag{15.31}$$

$$\sigma_{G,I} = \frac{\sigma_G^2(1) - \sigma_G^2(2)}{4} \tag{15.32}$$

En **términos causales,** como $(\sigma_G^2+\sigma_I^2)^2 = \sigma_G^4+\sigma_I^4+2\sigma_G^2\sigma_I^2$ y $(\sigma_G^2+\sigma_I^2+2\sigma_{G,I})(\sigma_G^2+\sigma_I^2-2\sigma_{G,I}) = \sigma_G^4+\sigma_I^4+2\sigma_G^2\sigma_I^2-4\sigma_{G,I}^2$, el producto de $(\sigma_G^2+\sigma_I^2+2\sigma_{G,I})(\sigma_G^2+\sigma_I^2-2\sigma_{G,I})$ puede escribirse como $(\sigma_G^2+\sigma_I^2)^2-4\sigma_{G,I}^2$ y entonces

$$\rho_x = \frac{\sigma_G^2 - \sigma_I^2}{\sqrt{(\sigma_G^2+\sigma_I^2)^2 - 4\sigma_{G,I}^2}} \tag{15.33}$$

El análisis de la ecuación (15.33) nos permite entender mejor el comportamiento de la característica de interés en dos ambientes, a la luz de la posible interacción genotipo-ambiente. En primer lugar, solo como corroboración, en caso de que la interacción genotipo-ambiente sea nula, entonces $\sigma_I^2 = \sigma_{G,I} = 0$ y por lo tanto

$$\rho_x = \frac{\sigma_G^2 - 0}{\sqrt{(\sigma_G^2+0)^2 - 4\times 0}} = \frac{\sigma_G^2}{\sqrt{\sigma^4}} = \frac{\sigma_G^2}{\sigma_G^2} = 1 \tag{15.34}$$

o lo que es lo mismo, la **correlación genética entre ambientes es igual a uno**, es decir, **se trata de la misma característica** (pese a haberla considerado como dos características diferentes para el análisis).

Por otra parte, si existe una correlación negativa perfecta entre los rankings de los grupos genéticos (familias) a través de los ambientes, entonces la varianza genética de los grupos es cero, es decir $\sigma_G^2 = 0$ y por lo tanto $\sigma_{G,I} = 0$ (ya que es la covarianza de una variable aleatoria con una constante). Colocando esos valores en la ecuación (15.33), tenemos que en ese caso, la **correlación genética entre ambientes** es

$$\rho_x = \frac{0 - \sigma_I^2}{\sqrt{(0+\sigma_I^2)^2 - 4\times 0}} = -\frac{\sigma_I^2}{\sigma_I^2} = -1 \tag{15.35}$$

Finalmente, cuando **las varianzas entre familias en los dos ambientes son iguales**, de acuerdo con la ecuación (15.20), $\sigma_G^2+\sigma_I^2+2\sigma_{G,I} = \sigma_G^2+\sigma_I^2-2\sigma_{G,I} \Leftrightarrow \sigma_{G,I} = 0$. Por lo tanto, de acuerdo con la ecuación (15.33), sustituyendo ese valor tenemos

$$\rho_x = \frac{\sigma_G^2 - \sigma_I^2}{\sqrt{(\sigma_G^2+\sigma_I^2)^2 - 4\times 0}} = \frac{\sigma_G^2 - \sigma_I^2}{\sqrt{(\sigma_G^2+\sigma_I^2)^2}} = \frac{\sigma_G^2 - \sigma_I^2}{\sigma_G^2+\sigma_I^2} \tag{15.36}$$

Claramente, de acuerdo con esta última ecuación, si las varianzas entre familias en los dos ambientes son iguales, entonces (como las varianzas deben ser positivas) la presencia de interacción disminuye el valor de la correlación genética por debajo de 1.

PARA RECORDAR

- Podemos describir los fenotipos de dos individuos k y l, ambos pertenecientes al mismo grupo genotípico j, para la característica de interés en los dos ambientes, 1 y 2 como $z_{1jk} = \mu + G_j + I_j + E + e_{1jk}$
  $z_{2jl} = \mu + G_j - I_j - E + e_{2jl}$
- Las propiedades esenciales de este modelo lineal pueden resumirse en: 1) la **media total** del modelo, es decir de **todos los genotipos a través de todos los**

ambientes es igual a μ, 2) los **efectos genéticos de los grupos** de individuos (familias) **se asumen aleatorios**, con esperanza de cero y varianza σ^2_G, 3) de la misma manera, **para un ambiente dado los efectos de interacción son variables aleatorias** con esperanza de cero y varianza σ^2_I, 4) **dentro de los grupos genéticos** el **efecto medio de interacción** está restringido a que debe ser **cero**, 5) puede existir **correlación entre los valores genotípicos y los efectos de interacción**, en cuyo caso la covarianza es de $\sigma_{G,I}$, 6) **el valor promedio de los dos ambientes es cero**, 7) las **desviaciones residuales** dentro de los ambientes tienen **esperanza de cero y varianza** σ^2_e para todos los grupos genéticos.

- Las **varianzas genéticas de los grupos** (familias) en los dos ambientes serán
  $\sigma^2_G(1) = \sigma^2(G_j + I_j) = \sigma^2_G + \sigma^2_I + 2\sigma_{G,I}$
  $\sigma^2_G(2) = \sigma^2(G_j - I_j) = \sigma^2_G + \sigma^2_I - 2\sigma_{G,I}$
- La **covarianza genética de los grupos** entre los dos ambientes será
  $\sigma_G(1,2) = \sigma^2_G - \sigma^2_I$
- A su vez, el **coeficiente de correlación genética (aditiva) entre ambientes** quedó definido como $\rho_x = \frac{\sigma_G(1,2)}{\sigma_G(1)\ \sigma_G(2)}$
- Entonces, a partir de estimaciones de las varianzas ($\sigma^2_G(1)$ y $\sigma^2_G(2)$) y covarianzas ($\sigma_G(1,2)$) de **familias** podemos **estimar la correlación genética de la característica entre ambientes diferentes.**
- En caso de que la **interacción genotipo-ambiente sea nula**, entonces $\sigma^2_I = \sigma_{G,I} = 0$ y por lo tanto la **correlación genética entre ambientes es igual a uno**, es decir, **se trata de la misma característica** (pese a haberla considerado como dos características diferentes para el análisis). Análogamente, si existe una correlación negativa perfecta entre los rankings de los grupos genéticos (familias) a través de los ambientes, entonces la varianza genética de los grupos es cero, es decir $\sigma^2_G = 0$ y por lo tanto $\sigma_{G,I} = 0$ (ya que es la covarianza de una variable aleatoria con una constante). Por lo que en ese caso, la **correlación genética entre ambientes** es -1.

Procedimientos de estimación

A partir de los resultados obtenidos en la sección anterior resulta claro que la determinación de la existencia de interacción entre genotipo y ambiente a partir del método de la correlación consiste en estimar su valor y determinar si el mismo es **significativamente menor a uno**. Es decir, un valor de $\rho_x = 1$ es indicación de que lo que consideramos como dos características desde el punto de vista genético es en realidad la misma y que por lo tanto no existe interacción G×E. El problema es que aun siendo la correlación igual a uno, el valor estimado es una variable aleatoria, con su propia distribución y que posiblemente casi siempre nos dé un valor estimado menor a uno, aun siendo uno su valor verdadero. Por lo tanto, tenemos dos problemas diferentes, uno de estimación y el otro de significación estadística.

Para obtener un estimado r_x de ρ_x, el procedimiento más sencillo consiste en separar la familia en dos y ensayar cada mitad en uno de los ambientes. Eso nos proporcionará estimados de las varianzas entre familias en los dos ambientes ($\sigma^2_G(1)$ y $\sigma^2_G(2)$), así como un

estimado de la covarianza de familia entre ambientes ($\sigma_G(1,2)$). A partir de esto, alcanza con sustituir los valores observados por los valores esperados en la ecuación (15.22), es decir

$$\rho_x = \frac{\sigma_G(1,2)}{\sigma^2_G(1)\ \sigma^2_G(2)} \tag{15.37}$$

El procedimiento más sencillo para obtener estimados de $\sigma^2_G(1)$ y $\sigma^2_G(2)$ es a través de un análisis de varianza, como el que hemos visto en capítulos precedentes, mientras que para la estimación de la covarianza en procedimiento más sencillo parece ser el cálculo de la covarianza observada entre medias familiares en los dos ambientes. Una advertencia importante aquí es la de evitar el uso de estructuras familiares en las que se compartan otros factores, por ejemplo, los efectos maternales ya que estos contaminarán las estimaciones.

Como la estimación de las varianzas y covarianzas son obtenidas indirectamente, la correlación obtenida por este método no se trata de una correlación producto-momento (la correlación clásica, o de Pearson) y por lo tanto las estimaciones pueden escaparse del intervalo natural (± 1). Como sugieren Lynch y Walsh (1998) sugieren, una forma sencilla de obtener una correlación producto-momento es a través de la regresión de las medias familiares en los dos ambientes, por ejemplo, $\bar{z}_{1j}$ en $\bar{z}_{2j}$.

Otra alternativa usada en la de obtener r_x directamente a partir de la correlación de las medias familiares $r(\bar{z}_{1j},\bar{z}_{2j})$ como un estimador de la correlación genética entre ambientes. Sin embargo, como estos autores remarcan, la correlación estimada de esta forma se encuentra sesgada hacia abajo, ya que en el denominador de la misma aparecen los efectos del error de medición y del ambiente inflando las correspondientes varianzas. Esto lleva a una tendencia a declarar la existencia de interacción genotipo-ambiente ya que puede sugerir valores significativamente menores a uno, aun cuando la correlación genética sea de uno.

Cuando existen datos disponibles para más de dos ambientes una alternativa es el uso de un análisis de varianza de dos vías, siendo los dos factores los genotipos (familias) y los ambientes. Esto nos introduce en un tema de modelos con efectos aleatorios y una mezcla de efectos aleatorios y fijos, cuya interpretación es algo diferente entre ellos. Si consideramos que los genotipos son una muestra aleatoria de una distribución de genotipos posibles (por ejemplo, de «familias»), entonces los efectos macroambientales pueden ser considerados de dos maneras diferentes, como aleatorios o como fijos. En el primer caso se trataría de que los ambientes también son una muestra representativa de los ambientes posibles, como cuando elegimos al azar diferentes tambos, o regiones (macroambientes) para ensayar los genotipos, o cuando elegimos al azar parcelas donde probar variedades de maíz. En este caso, tenemos un modelo aleatorio, ya que tanto los genotipos como los macroambientes son ambos aleatorios. En el segundo caso, si consideramos una serie particular de ambientes de interés, por ejemplo, diferentes países o regiones donde evaluar los genotipos, que son las únicas de interés, entonces vamos a tratar con modelos mixtos, ya que los genotipos serán aleatorios, mientras que los macroambientes serán considerados efectos fijos. Para entender las diferencias entre ambos tipos de modelos se recomienda la lectura del capítulo 22 de Lynch y Walsh (1998), en especial la sección «Two-way analysis of variance».

El uso de modelos mixtos para el estudio de la interacción genotipo ambiente a partir de datos existentes en las evaluaciones genéticas nacionales e internacionales ha sido extenso. Ejemplos locales y regionales incluyen los estudios realizados por Naya, Urioste y Franco (2002) en vacas lecheras a partir de la definición de «clusters ambientales» obtenidos de los propios datos de la evaluación; el trabajo de Espasandin et al. (2013), comparando las

poblaciones Angus de Uruguay y Brasil o el trabajo de Zwald et al. (2003), investigando los factores que afectan el resultado de vacas **holstein** en 17 países que participaban de la evaluación **MACE** realizada por **INTERBULL**.

15.4 Genética cuantitativa de la interacción G×E

En todo el tratamiento que dimos previamente a la genética cuantitativa asumimos que genotipo era sinónimo de la acción de los genes sobre el fenotipo de los individuos y que podíamos aislar ese concepto del concepto de ambiente. Sin embargo, como lo muestra Jong (1990), en el trabajo original de Woltereck (1909) donde aparece por primera vez el concepto de norma de reacción, el concepto del mismo Woltereck era que «Der "Genotypus" (die genotyp. Grundlage) eines Quantitativmerkmals ist die vererbte Reaktionsnorm» («El "genotipo" (la base genotípica) de un rasgo cuantitativo es la norma de reacción heredada»), a tal punto que el encabezado de la sección luce «Genotypus = Reaktionsnorm». Es decir, en la concepción original de Woltereck aparece que no podemos separar genotipo de ambiente ya que, para él, el concepto de genotipo es precisamente la forma específica de reaccionar a los ambientes que tienen los individuos con base genética.

En este capítulo hemos ido avanzando en la comprensión de la importancia que tiene la interacción entre el genotipo y el ambiente, pero apenas nos hemos asomado en el impacto que tiene la misma en los principales resultados que hemos derivado para la genética cuantitativa. De acuerdo con Lynch y Walsh (1998), lamentablemente pocos avances se han hecho en la materia, aunque el trabajo de Jong (1990) sugiere un camino simple e interesante, por lo que vamos a delinear sus principios aquí como forma de conectar el contenido de este capítulo con los capítulos previos.

Supongamos que tenemos una característica fenotípica que se encuentra gobernada por un gen con dos alelos, con frecuencias p y q como usualmente. La única diferencia en nuestro planteo actual respecto a lo que vimos antes en el capítulo «El modelo genético básico» es que ahora el efecto medio de los alelos ya no es independiente del ambiente, sino que es una función lineal del mismo. Por lo tanto, con esta diferencia, los efectos medios de los alelos serán ahora del tipo

$$\begin{aligned} \alpha_1 &= a_1 + c_1 E \\ \alpha_2 &= a_2 + c_2 E \end{aligned} \tag{15.38}$$

con a_1 y a_2 los interceptos y c_1 y c_2 las pendientes de la regresión del valor medio del alelo en el gradiente ambiental.

Como la respuesta genotípica será ahora la norma de reacción, entonces, para una combinación específica de alelos i y j, la misma será

$$R_{ij} = \alpha_i + \alpha_j = (a_i + c_i E) + (a_j + c_j E) \ \therefore \\ R_{ij} = (a_i + a_j) + (c_i + c_j)E \tag{15.39}$$

En características determinadas por varios genes de acción aditiva entre ellos, la norma de reacción será igual a la suma de los R_{ij} de todos los loci involucrados, cada uno con sus frecuencias alélicas y funciones lineales de E.

En el capítulo «El modelo genético básico» vimos que el efecto de sustitución era igual a $\alpha = \alpha_1 - \alpha_2$ y que de acuerdo con la ecuación (8.120) la varianza aditiva era igual a $\sigma^2_A = 2pq\alpha^2$. Por lo tanto, el efecto de sustitución será $\alpha = (a_1 + c_1 E) - (a_2 + c_2 E) = (a_1 - a_2) + (c_1 - c_2)E$ y la varianza aditiva será

$$\sigma^2_A = 2pq[(a_1 - a_2) + (c_1 - c_2)E]^2 \tag{15.40}$$

La expansión de la ecuación anterior nos lleva a

$$\sigma^2_A = 2pq[(a_1 - a_2)^2 + 2E(a_1 - a_2)(c_1 - c_2) + (c_1 - c_2)^2 E^2] = \\ = 2pq(a_1 - a_2)^2 + 2pq[2E(a_1 - a_2)(c_1 - c_2)] + 2pq(c_1 - c_2)^2 E^2 \tag{15.41}$$

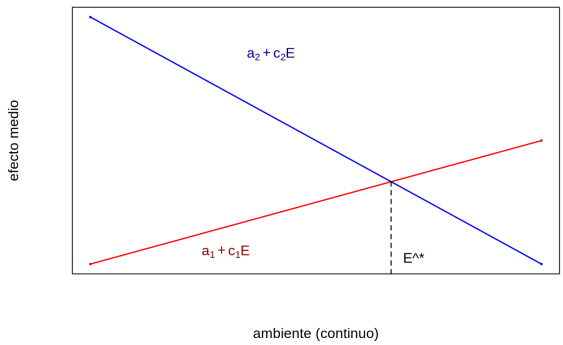


Figura 15.6. Normas de reacción lineales para el efecto medio de un alelo. El punto $E^\star$ representa el valor del macroambiente donde se produce intersección de las dos funciones lineales (es decir, los efectos medios son iguales y el efecto de sustitución igual a cero).

Considerando un punto determinado del macroambiente E, es decir, considerado como una constante y reconociendo cada uno de los términos de la ecuación (15.41) como una varianza o covarianza, podemos expresar la varianza aditiva bajo la norma de reacción como

$$\sigma_A^2 = \sigma_a^2 + 2E\sigma_{a,c} + E^2\sigma_c^2 \tag{15.42}$$

es decir, la suma de las varianzas y covarianza de interceptos y pendientes que forman la norma de reacción.

Por lo tanto, de acuerdo con la ecuación (15.42), la varianza aditiva de la característica no será más un valor solamente dependiente de α y de las frecuencias, sino que ahora será una función del macroambiente. De hecho, mirando la función $\sigma_A^2(E)$ apreciamos que se trata de una función parabólica. Como además las varianzas deben ser positivas, el coeficiente del término cuadrático (σ_c^2) en la función es positivo y por lo tanto se trata de una parábola cóncava (con los brazos extendidos hacia arriba). El mínimo de la función estará dado además por el valor en que la derivada se hace cero, que es

$$\begin{aligned} \frac{d\sigma_A^2(E)}{dE} = \frac{d[\sigma_a^2 + 2E\sigma_{a,c} + E^2\sigma_c^2]}{dE} = 2\sigma_{a,c} + 2\sigma_c^2 E \ \therefore \\ 2\sigma_{a,c} + 2\sigma_c^2 E = 0 \Leftrightarrow E = -\frac{\sigma_{a,c}}{\sigma_c^2} = -\frac{(a_1 - a_2)(c_1 - c_2)}{(c_1 - c_2)^2} \\ E = -\frac{(a_1 - a_2)}{(c_1 - c_2)} \end{aligned} \tag{15.43}$$

En la figura 15.7 se representa la superficie correspondiente a la varianza aditiva, tal cual aparece en la ecuación (15.40). La misma es función de la frecuencia de uno de los dos alelos, p (ya que $q = 1 - p$) y del macroambiente E ya que en este caso las normas de reacción de los genotipos son lineales y no paralelas. Como es de esperar, la varianza tiene forma parabólica convexa con respecto a p y con máximo en $p = q = \frac{1}{2}$, mientras que también tiene forma parabólica, pero cónica con respecto al ambiente y con mínimo en $E^\star$. Una forma complementaria de entender las consecuencias de esta dependencia de la varianza aditiva en el valor del macroambiente es considerar que como α_1 y α_2 son funciones lineales de E, existen dos posibilidades para las rectas que estas ecuaciones describen: a) son parelelas, o b) se cortan. Si son paralelas, entonces las pendientes c_1 y c_2 son iguales, por lo que $c_1 - c_2 = 0$ y la ecuación (15.40) se transforma en

$$\sigma_A^2 = 2pq[(a_1 - a_2) + (c_1 - c_2)E]^2 = 2pq(a_1 - a_2)^2 \tag{15.44}$$

y no hay dependencia de la varianza aditiva en el ambiente, es decir no se observa un efecto de interacción genotipo ambiente. Este resultado también podría obtenerse de la ecuación (15.42), ya que si $c_1 = c_2$ entonces $\sigma_c^2 = \sigma_{a,c} = 0$ (ya que c es constante).

En el caso de que α_1 y α_2 no sean paralelas, entonces las mismas deben cortarse en algún punto. Si igualamos ambas funciones, tenemos que

$$\begin{aligned} \alpha_1 = \alpha_2 \Leftrightarrow a_1 + c_1 E = a_2 + c_2 E \Leftrightarrow E(c_1 - c_2) = (a_2 - a_1) \ \therefore \\ E^\star = \frac{(a_1 - a_2)}{(c_2 - c_1)} \end{aligned} \tag{15.45}$$

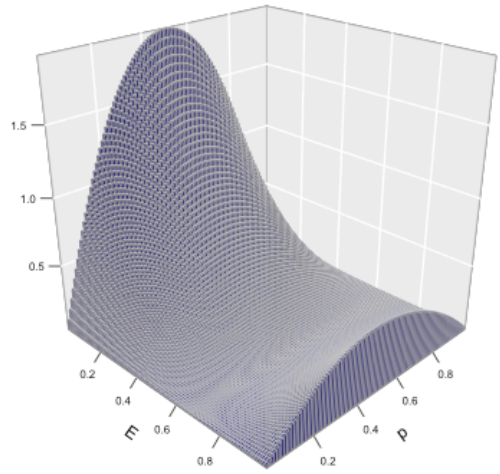


Figura 15.7. Superficie de la varianza aditiva para normas de reacción lineales en función de la frecuencia *p* de uno de los dos alelos y del valor del macroambiente (arbitrario, representado entre cero y uno).

donde $E^\star$ es el ambiente correspondiente al punto de corte de las funciones de efecto medio. Si sustituimos dicho valor en la función de la varianza aditiva dada por la ecuación (15.40) tenemos que

$$\sigma_A^2 = 2pq[(a_1 - a_2) + (c_1 - c_2)E]^2 = 2pq[(a_1 - a_2) + (c_1 - c_2)\frac{(a_1 - a_2)}{(c_2 - c_1)}]^2 = \quad (15.46)$$

$$2pq[(a_1 - a_2) - (a_1 - a_2)]^2 = 2pq[0]^2 \ \therefore \quad (15.47)$$

$$\sigma_A^2 = 0 \quad (15.48)$$

Por lo tanto, si tenemos que las normas de reacción son lineales y existe un efecto de interacción, entonces encontramos un valor ambiental en el que la varianza aditiva de la característica es igual a cero, por lo que no responderá a la selección (la heredabilidad de la característica es igual a cero). Como para cada locus los interceptos y pendientes son específicos, también lo serán los puntos de corte y por lo tanto para características poligénicas resulta difícil que en algún punto la varianza aditiva sea cero. Sin embargo, la posibilidad es firme de que exista uno o más puntos del macroambiente en que la varianza aditiva se vea reducida sustancialmente.

Un punto final al que este análisis nos permite arribar fácilmente es al del estudio de la correlación genética de la característica expresada en dos ambientes, E_j y E_k. En este caso, de forma análoga a lo derivado para la varianza aditiva en la ecuación (15.40), la covarianza genética estará dada por

$$\sigma_{A_j,A_k} = 2pq[(a_1 - a_2) + (c_1 - c_2)E_j][(a_1 - a_2) + (c_1 - c_2)E_k] \quad (15.49)$$

De acuerdo con esta última ecuación, si cualquiera de los dos ambientes es igual al punto de corte de los efectos medios, es decir $E_j = E^\star$ o $E_k = E^\star$, entonces la covarianza genética de la característica será igual a cero y lo mismo para la correlación genética. Por otra parte, si E_j y E_k se encuentran ambos del mismo lado respecto a $E^\star$ (a la izquierda o a la derecha de dicho punto), entonces el producto $[(a_1 - a_2) + (c_1 - c_2)E_j][(a_1 - a_2) + (c_1 - c_2)E_k]$ será positivo y la correlación genética será también positiva. Sin embargo, si los dos puntos se encuentran en lados diferentes respecto a $E^\star$, entonces el producto será negativo y también lo será la correlación genética. Si bien se trata de un modelo teórico elemental, el mismo permite explicar correlaciones genéticas negativas de la misma característica cuando la misma se expresa en dos ambientes diferentes y es un insumo importante para construir modelos más complejos.

Ejemplo 15.2

Una determinada característica fenotípica se encuentra determinada por un locus con dos alelos. Para esta característica existe interacción entre genotipo y ambiente ya que los efectos medios de los dos alelos son $\alpha_1 = 6 - E$ y $\alpha_2 = 4 + 2E$, siendo E la variable (continua) ambiental que determina el valor de los alelos.

Graficar los efectos medios de los genes en función del ambiente (entre $E = 0$ y $E = 1$) y calcular el punto en que son iguales. Determinar la varianza aditiva de la característica

como función del ambiente E y graficarla entre los valores de $E = 0$ y $E = 1$ para una frecuencia del alelo A_1 $p = 0{,}25$ y para $p = 0{,}50$.

La figura 15.8 representa los efectos medios de ambos alelos como función del ambiente E, entre los valores de $E = 0$ y $E = 1$. De acuerdo con la ecuación (15.45), el punto de corte se va a encontrar en

$$E^\star = \frac{(a_1 - a_2)}{(c_2 - c_1)} = \frac{6-4}{2-(-1)} = \frac{2}{3} = 0{,}666 \tag{15.50}$$

lo que aparece marcado en la figura. La varianza aditiva es una función del ambiente. De

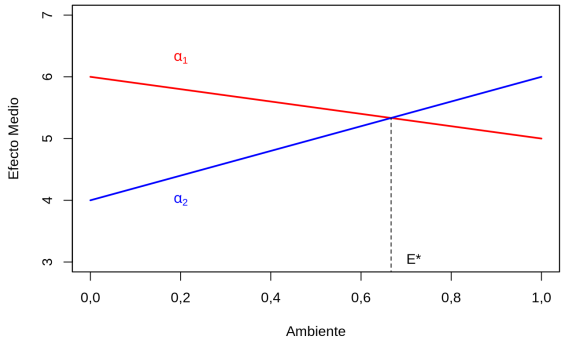


Figura 15.8. Efectos medios de los dos alelos $\alpha_1 = 6 - E$ y $\alpha_2 = 4 + 2E$, como función del ambiente E. El punto de corte aparece marcado por $E^\star$.

acuerdo con la ecuación (15.40) la misma es igual a

$$\sigma^2_A = 2pq[(a_1 - a_2) + (c_1 - c_2)E]^2 = 2pq[(6-4) + (-1-2)E]^2 = 2pq[2 - 3E]^2 \tag{15.51}$$

Sustituyendo ahora el valor de $p = 0{,}25$ nos queda

$$\sigma^2_A = 2pq[2 - 3E]^2 = 2 \times 0{,}25 \times 0{,}75[2 - 3E]^2 = 0{,}375[2 - 3E]^2 \tag{15.52}$$

mientras que para $p = 0{,}5$ nos queda

$$\sigma^2_A = 2pq[2 - 3E]^2 = 2 \times 0{,}5 \times 0{,}5[2 - 3E]^2 = 0{,}5[2 - 3E]^2 \tag{15.53}$$

La representación gráfica de estas funciones de varianza aparece en la figura 15.9, donde se aprecia claramente la forma parabólica de la varianza aditiva, con un mínimo en $E = \frac{2}{3} = E^\star$, ya que en ese valor $\sigma^2_A = 2pq[2 - 3(\frac{2}{3})]^2 = 0$

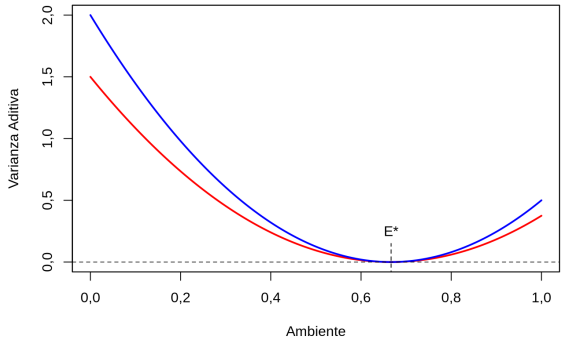


Figura 15.9. Varianza aditiva de la característica en función del ambiente E, para dos frecuencias alélicas del alelo A_1. En rojo $p = 0{,}25$, en azul $p = 0{,}5$. Como la función de varianza aditiva es igual a $\sigma^2_A = 2pq[2 - 3E]^2$, cuando $E = \frac{2}{3} = E^\star$ la varianza se hace cero.

PARA RECORDAR

- Para obtener un **estimado r_x de ρ_x,** el procedimiento más sencillo consiste en **separar la familia en dos y ensayar cada mitad en uno de los ambientes.** Eso nos proporcionará estimados de las varianzas entre familias en los dos ambientes ($\sigma^2_G(1)$ y $\sigma^2_G(2)$), así como un estimado de la covarianza de familia entre ambientes ($\sigma_G(1,2)$).
- Si tenemos una característica fenotípica que se encuentra gobernada por un gen con dos alelos, con frecuencias p y q, el **efecto medio de los alelos ya no es independiente del ambiente** sino que es una función lineal del mismo, por lo que ahora serán $\alpha_1 = a_1 + c_1E$
  $\alpha_2 = a_2 + c_2E$
  con a_1 y a_2 los interceptos y c_1 y c_2 las pendientes de la regresión del valor medio del alelo en el gradiente ambiental.
- Como la **respuesta genotípica** será ahora la norma de reacción, entonces, para una combinación específica de alelos i y j, la misma será $R_{ij} = (a_i + a_j) + (c_i + c_j)E$. En **características determinadas por varios genes de acción aditiva entre ellos**, la norma de reacción será igual a la suma de los R_{ij} de todos los loci involucrados, cada uno con sus frecuencias alélicas y funciones lineales de E.
- La varianza aditiva bajo la norma de reacción será $\sigma^2_A = \sigma^2_a + 2E\sigma_{a,c} + E^2\sigma^2_c$, es decir, la suma de las varianzas y covarianza de interceptos y pendientes que forman la norma de reacción.
- Dada la correlación genética de la característica expresada en dos ambientes, E_j y E_k, la covarianza genética estará dada por $\sigma_{A_j,A_k} = 2pq[(a_1 - a_2) + (c_1 - c_2)E_j][(a_1 - a_2) + (c_1 - c_2)E_k]$. Por lo que, si cualquiera de los dos ambientes es igual al punto de corte de los efectos medios, es decir $E_j = E^\star$ o $E_k = E^\star$, entonces la covarianza genética de la característica será igual a cero y lo mismo para la correlación genética. Por otra parte, si E_j y E_k se encuentran ambos

del mismo lado respecto a $E^\star$ (a la izquierda o a la derecha de dicho punto), entonces el producto $[(a_1 - a_2) + (c_1 - c_2)E_j][(a_1 - a_2) + (c_1 - c_2)E_k]$ será positivo y la correlación genética será también positiva. Sin embargo, si los dos puntos se encuentran en lados diferentes respecto a $E^\star$, entonces el producto será negativo y también lo será la correlación genética.

15.5 Otros ejemplos de interacción genotipo-ambiente

En 1982, Gupta y Lewontin (1982) realizaron un estudio sobre normas de reacción en la mosca *Drosophila pseudoobscura*, variando la densidad de huevos y la temperatura y midiendo el número de setas abdominales. Para ello usaron 32 cepas isocromosómicas para el segundo cromosoma. En la figura 15.10 se ilustra alguno de los resultados obtenidos. Si bien los efectos medios del genotipo, el efecto de la temperatura y de la densidad fueron significativos, también aparecieron importantes efectos de interacción genotipo-ambiente, que en la figura aparecen como la falta de paralelismo en las líneas. Por otra parte, los autores observaron que alrededor del $30 - 45\,\%$ de las comparaciones por pares entre genotipos mostraron un cambio en la posición relativa de los mismos al cambiar la temperatura. Por lo tanto, no es posible caracterizar un genotipo como si tuviera mayor número de setas que otro, ya que esto solo puede ser determinado en forma relativo a un ambiente particular.

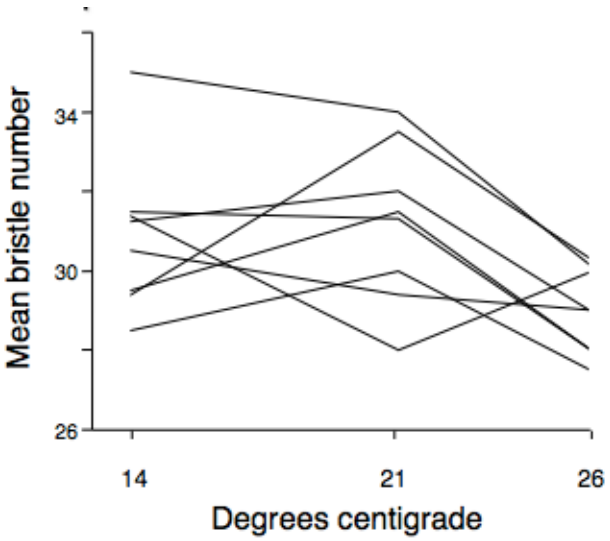


Figura 15.10. Variación en el número de setas abdominales en la mosca *Drosophila pseudobscura* frente al cambio de temperatura durante el desarrollo. La ausencia de paralelismo en las líneas muestra la interacción genotipo-ambiente (Autor: Vbtbio - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36851702).

La interacción entre genotipo y ambiente no solo se manifiesta en características cuantitativas. Por ejemplo, en el caso de patologías, la exposición a determinados elementos del

ambiente puede hacer que genotipos que en otros ambientes no desarrollaban una enfermedad, en uno nuevo ambiente sí lo hagan, mientras que otros genotipos son insensibles a esto. En los seres humanos, los efectos ambientales son complejos y su importancia relativa puede variar considerablemente en el tiempo y depender de la genética de los individuos. Por otro lado, en muchos casos la interacción entre los factores físicos, los sociales y los económicos contribuyen en forma importante a la interacción G×E, muchas veces de formas complejas (como en el caso del tabaquismo, el alcoholismo y la malnutrición). Como resultado, la interacción G×E puede darse en la aparición o no de la enfermedad, en el tiempo para que aparezca, en el ritmo de su progresión y en la respuesta clínica a la intervención farmacológica.

Algunos ejemplos de estas interacciones entre genotipo y ambiente son, donde la combinación aumenta el riesgo de padecer enfermedades son las siguientes:

Pesticidas organofosforados y enfermedad de Parkinson
Si bien la enfermedad de Parkinson (EP) es del tipo multifactorial, con una combinación de influencias que incluyen variantes genéticas, exposición ambiental y estilo de vida, existen casos documentados de aumento sustantivo en el riesgo de desarrollar la enfermedad de Parkinson luego de la exposición organofosforados para genotipos específicos. En particular, el gen **PON1** codifica una enzima paraoxonasa que metaboliza los organofosforados. Algunos individuos son portadores de una variante menos eficiente de PON1 y cuando se exponen a los organofosforados, tienen el doble de riesgo de contraer la enfermedad.

Cáncer de esófago y alcohol
El gen de la aldehído deshidrogenasa 2 (ALDH2) produce una enzima importante para el metabolismo del alcohol, con dos variantes diferentes en el gen, una que produce una enzima que funciona y otra (una mutación estructural generada por el intercambio del aminoácido lisina por glutamato en la posición 487) que produce una enzima que no funciona. La mayoría de los caucásicos tienen dos isozimas principales, mientras que aproximadamente el 50 % de los asiáticos orientales tienen una copia normal del gen ALDH2 y una copia de la forma inactiva. Las personas portadoras de la variante no funcional del gen ALDH2 tienen problemas importantes para metabolizar el alcohol, ya que la enzima defectuosa provoca la acumulación de aldehídos en el organismo (que se conoce como síndrome de enrojecimiento o *flushing*). Los individuos portadores (heterocigotos) tienen una reducción de alrededor de 100 veces en el metabolismo del alcohol. Estos heterocigotos tienen un mayor riesgo de padecer cáncer de esófago cuando se exponen a un consumo moderado o intenso de alcohol etílico en comparación con alguien que tiene ambas variantes funcionales.

Fenilcetonuria
Se trata de una alteración congénita del metabolismo causada reducción o ausencia de la actividad en la enzima fenilalanina hidroxilasa, codificada por el gen PAH, localizado en el cromosoma 12. La misma tiene como cometido metabolizar el aminoácido fenilalanina (Phe).
Se conocen más de 400 mutaciones en ese gen que producen una alteración en la capacidad metabólica del mismo, lo que lleva bajo una dieta normal a un incremento patológico en los niveles de fenilalanina. Sin embargo, en una dieta estrictamente controlada, con bajos niveles de fenilalanina la enfermedad no se manifiesta (en ninguno de los dos genotipos), lo que indica una interacción genotipo-ambiente.

En plantas, los programas de mejora de cultivos disponen cada vez más de datos genómicos, así como de rendimiento en el campo y medioambientales. El análisis de esta información de forma integrada puede facilitar la predicción genómica dentro y entre ambientes, así como proporcionar información sobre la arquitectura genética de rasgos complejos y la naturaleza de las interacciones genotipo-ambiente. Para entender las relaciones entre varianzas aditiva y de dominancia, así como los efectos de la interacción genotipo-ambiente, Rogers et al. (2021) recopilaron y analizaron los datos genotípicos y fenotípicos de 1918 híbridos de maíz (*Zea mays L.*) y los datos ambientales de 65 ambientes de prueba. Para el rendimiento del grano, la varianza de dominancia fue similar en magnitud a la varianza aditiva, y las varianzas G×E fueron más importantes que las varianzas genéticas de efecto principal.

Para otros rasgos, como el tiempo de floración, la varianza G×E solo representó una pequeña parte de la varianza fenotípica, lo que sugiere que puede ignorarse con seguridad al realizar selección para estas características. Los resultados de este estudio sugieren que, para la predicción genómica de rendimiento de los híbridos en función del ambiente, la dominancia y las interacciones G×E deberían incorporarse a los modelos de predicción. Por otra parte, los modelos que incluían tanto los efectos aditivos como los de dominancia fueron los que mejor se ajustaron a los datos, y la modelización de las covarianzas genéticas únicas entre todos los ambientes proporcionó la mejor caracterización de los patrones de interacción genotipo-ambiente. A su vez, la similitud del rendimiento relativo de los híbridos entre los ambientes se modificó en función de las variables climáticas subyacentes, lo que permitió identificar las covariables climáticas que impulsan las correlaciones de los efectos genéticos entre los entornos.

PARA RECORDAR

- Debemos tener en cuenta que la interacción entre genotipo y ambiente **no solo se manifiesta en características cuantitativas**. Por ejemplo, en el caso de patologías, la exposición a determina- dos elementos del ambiente puede hacer que genotipos que en otros ambientes no desarrollaban una enfermedad, en uno nuevo ambiente sí lo hagan, mientras que otros genotipos son insensibles a esto.

15.6 Conclusión

Para finalizar esta sección (y el libro) nos adentramos en un tema de suma complejidad: la interacción genotipo-ambiente. Hasta este momento hemos considerado los componentes de la varianza como genéticos (aditivos) o ambientales, incluyendo en este último componente errores de medición y todo aquello que escapa a la herencia como la conocemos clásicamente. No obstante, como desarrollamos en este capítulo, la biología es compleja y un mismo genotipo puede mostrar diferentes fenotipos en ambientes disímiles. Esta interacción es difícil de modelar, y a su vez posee una importancia conceptual muy grande en campos distintos. A nivel práctico, resulta claro que es atractivo a nivel industrial buscar genotipos que sean lo más estables posibles en diferentes ambientes. Pero incluso para la ciencia básica tener en cuenta estos factores es sumamente importante: ¿podemos trasladar nociones obtenidas en el laboratorio a diferentes condiciones? ¿Cómo afecta la noción de plasticidad fenotípica a la evolución de las especies, siendo que este concepto escapa al componente hereditario clásico sobre el cual se suele parar la teoría evolutiva más sencilla? Adentrarse en estos problemas no es sencillo, ni a nivel matemático ni conceptual. Por lo tanto, este último capítulo puede ser entendido como una primera aproximación a un problema complejo. Lo que es más, podemos tomarlo de ejemplo para entender que, si nos adentramos lo suficiente en la complejidad de los temas planteados en los capítulos anteriores, la misma idea puede surgir. Con los conocimientos adquiridos en este libro, la persona interesada puede referirse a la bibliografía citada para comenzar a profundizar en la riqueza de dichos temas.

15.7 Actividades

15.7.1 Control de lectura

1. ¿Qué es la plasticidad fenotípica? ¿Y la norma de reacción?
2. ¿A qué refiere el concepto de canalización? ¿Cómo se vincula con el proceso evolutivo?
3. Desarrolla alguna aproximación para el estudio de la interacción genotipo-ambiente.
4. ¿Qué estimador conoces del coeficiente de correlación genética (aditiva) entre ambientes (ρ_x)?
5. Plantea ejemplos donde la interacción genotipo-ambiente posee un rol preponderante en el fenómeno biológico.

15.7.2 ¿Verdadero o falso?

1. La presencia de interacción G×E necesariamente implica una reclasificación de los genotipos.
2. La interacción entre genotipo y ambiente se manifiesta tanto en características cuantitativas como cualitativas.
3. En caso de que la interacción genotipo-ambiente sea nula, la correlación genética entre ambientes es igual a cero.
4. Cuando el coeficiente de correlación genético (aditiva) estimado para dos características es $\rho_x = 1$ no existe interacción genotipo-ambiente.
5. Si bien en cada ambiente un genotipo puede expresar un determinado fenotipo de manera óptima, existen rangos ambientales donde las respuestas poblacionales son aproximadamente equivalentes en diferentes ambientes.

Soluciones

1. **Falso.** Puede existir un cambio de escala sin reclasificación.
2. **Verdadero.** Podemos pensar en las enfermedades como un caso claro donde la interacción genotipo-ambiente puede llevar a efectos cualitativos: la presencia de ciertos metabolitos puede inducir cambios en estados celulares que lleven a una patología, y la respuesta dada por un genotipo puede variar para un rango dado del factor ambiental, manifestándose la patología solo luego de cierto umbral. El mismo razonamiento lleva a pensar también en casos como el desarrollo de determinadas morfologías, donde el pasaje de un determinado punto de corte puede inducir una respuesta celular que a través de una cascada lleve al desarrollo de un carácter.
3. **Falso.** Si el coeficiente es igual a uno se trata de la misma característica (pese a haberla considerado como dos características diferentes para el análisis).
4. **Verdadero.** En efecto, si $\rho_x = 1$ lo que estamos considerando como dos características son en efecto una. Por lo tanto, no puede existir interacción genotipo-ambiente en este caso.

5. **Verdadero.** Tomemos como ejemplo la respuesta a la disponibilidad hídrica en especies de plantas para ilustrar este concepto. Es posible notar que más allá que cada genotipo tenga un óptimo de producción, generalmente los extremos (anegamientos o sequías prolongadas) generarán que ambos genotipos tengan el mismo comportamiento. Si bien desde el punto de vista cuantitativo los óptimos pueden variar levemente, desde una visión más biológica es razonable pensar que las respuestas serán razonablemente parecidas en los rangos ambientales medios entre los extremos planteados.

15.7.3 Ejercicios

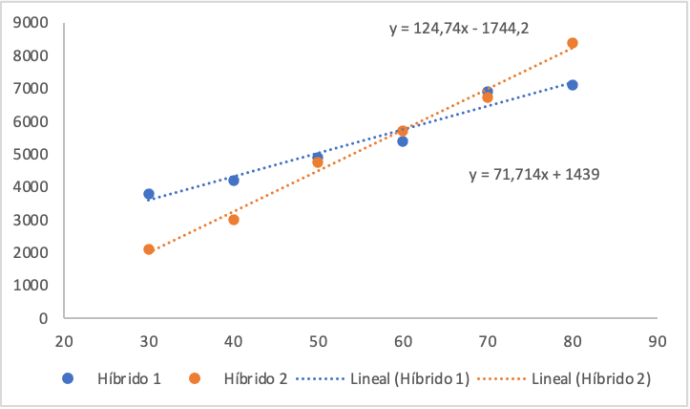
Ejercicio 15.1

A continuación se presenta el comportamiento en rendimiento (kg/ha) de dos híbridos de maíz bajo diferente disponibilidad hídrica en el rango de 30 a 80 % de agua disponible (AD).

a. Desarrolla un modelo lineal que describa el comportamiento de cada híbrido frente a la variación de disponibilidad hídrica. **b.** ¿Cuál de los híbridos tiene mayor respuesta al aumento en la disponibilidad hídrica?

Solución

a. Valiéndonos del uso de una herramienta de cálculo (planilla Excel, etcétera) podremos graficar y posteriormente obtener el modelo lineal asociado a cada híbrido como se observa a continuación.



Como se observa, el modelo lineal que mejor predice el comportamiento del híbrido 1 es Rend(kg/ha)$= 1439 + 71{,}714$ %AD, mientras que para el híbrido 2 es
Rend(kg/ha)$= -1744{,}2 + 124{,}74$ %AD

b. El híbrido con mayor respuesta al aumento en la disponibilidad de agua es el 2 debido a que su rendimiento aumenta casi 125 kg/ha por cada unidad de aumento en el %AD.

Ejercicio 15.2

Dos genotipos de avena negra (*Avena strigosa*) presentan diferente respuesta a la temperatura en el rango entre 7 y 30 ºC, lo que afecta su capacidad de generar biomasa foliar. La variedad **A** posee un rendimiento en kg/ha que puede ser descrito por la función $A = 3200(t22)^2$ kg/ha mientras que para la variedad **B** sería $B = 2900 + 15t$

¿A qué valor o valores de temperatura ambas variedades producen la misma biomasa foliar?

Solución

Para evaluar a qué temperaturas ambas variedades producen lo mismo deberemos igualar las funciones que describen el comportamiento de cada una

$$\begin{aligned} & [3200 - (t-22)^2] = [2900 + 15t] \\ \Leftrightarrow \quad & 3200 - 2900 = (t-22)^2 + 15t \\ \Leftrightarrow \quad & 300 = t^2 - 44t + 22^2 + 15t \\ \Leftrightarrow \quad & t^2 - 29t + 184 = 0 \end{aligned} \tag{15.54}$$

Obtuvimos como resultado una ecuación cuadrática, y revolviéndola obtendremos las dos raíces que representan las intersecciones entre ambas funciones

$$\begin{aligned} t^2 - 29t + 184 = 0 \Leftrightarrow t = \frac{29 \pm \sqrt{29^2 - 4 \times 184}}{2} \Leftrightarrow \\ t = \frac{29 \pm \sqrt{105}}{2} \Leftrightarrow t = \frac{29 \pm 10{,}2}{2} \Leftrightarrow \\ t = 9{,}4^{\circ}\mathrm{C};\ t = 19{,}6\,^{\circ}\mathrm{C} \end{aligned} \tag{15.55}$$

Por lo tanto, ambas variedades producirán la misma cantidad de biomasa a los 9,4 y a los 19,6 °C.

Ejercicio 15.3

Un grupo de investigación se dispone a estudiar la interacción genotipo-ambiente de una especie que crece en un rango de temperaturas amplio. En las poblaciones estudiadas encuentran tres genotipos para un locus dado, el cual se ve correlacionado con la capacidad metabólica a diferentes temperaturas. El alelo A es dominante, por lo que el grupo decide contrastar el comportamiento de organismos de genotipo $A./A.$ respecto al de organismos homocigotas recesivos aa/aa.

Luego de un arduo trabajo, determinan que la capacidad metabólica (medida por satélite como producción de CO_2 por área del organismo, $g \times cm^{-2} CO_2/m^3$) sigue las siguientes ecuaciones de dependencia con la temperatura (T, en grados Celsius)

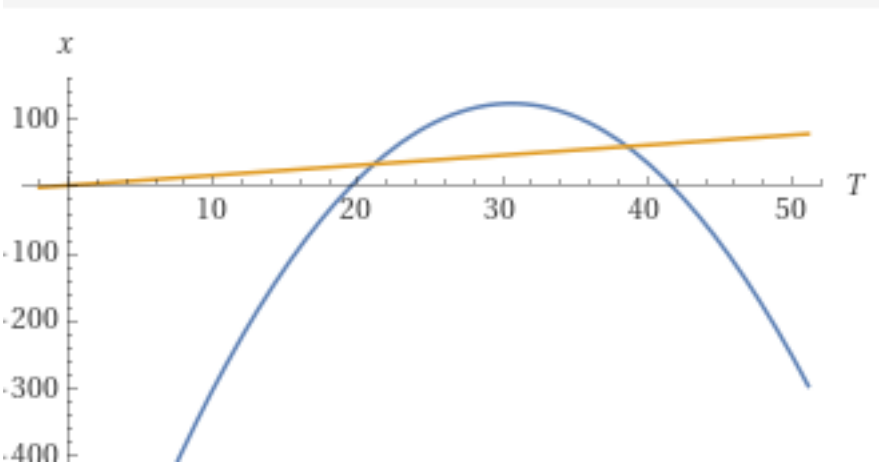
$$\begin{cases} \text{Capacidad metabólica (A./A.) (T)} = [T^2 + 61{,}2T812{,}55]g \times cm^{-2} CO_2/m^3 \\ \text{Capacidad metabólica (aa/aa)} = [1.5T + 2]g \times cm^{-2} CO_2/m^3 \end{cases} \tag{15.56}$$

¿Reaccionan igual ambos genotipos frente al ambiente?

Solución

Claramente ambos genotipos no reaccionan de forma similar al ambiente: una de las normas de reacción sigue una ecuación de segundo grado (parábola), mientras que la otra una ecuación lineal. Esto constituiría un caso bastante peculiar de interacción genotipo-ambiente.

Si graficamos ambas normas de reacción obtenemos el siguiente gráfico



Observamos que existe un rango inicial en el cual el Genotipo 2 posee mayor capacidad metabólica que el Genotipo 1. La situación se revierte, pero a medida que aumenta la temperatura la capacidad metabólica de este último genotipo alcanza un pico, volviendo hacia el final a ser superada por la capacidad del Genotipo 1.

Ejercicio 15.4

En un laboratorio se estudia la norma de reacción frente a la temperatura de dos fenotipos (largo de ala y largo de antenas) en individuos de *D. melanogaster*. El grupo de investigación sabe de la literatura que dos genotipos son la base de cada genotipo observado en cada caso, por lo que se disponen a analizar dos grupos de individuos de líneas puras para rangos de temperaturas controladas en cada caso.

Un análisis estadístico arroja los siguientes datos:

Largo de ala

$$\begin{cases} \sigma_T^2 = 0{,}39 \\ \sigma_{largo}^2 = 0{,}67 \\ \sigma_{T,largo} = 0{,}09 \end{cases} \tag{15.57}$$

Largo de antenas

$$\begin{cases} \sigma_T^2 = 0{,}34 \\ \sigma_{largo}^2 = 0{,}45 \\ \sigma_{T,largo} = 0{,}149 \end{cases} \tag{15.58}$$

¿Existe interacción genotipo-ambiente en alguno de estos casos?

Solución

Recordemos que podemos estimar la correlación de las medidas de temperatura y cada fenotipo para los grupos genotípicos según

$$\rho_x = \frac{\sigma_G(1,2)}{\sigma_G^2(1)\,\sigma_G^2(2)} \tag{15.59}$$

Tenemos así que para el caso del largo de alas

$$\rho_x = \frac{0{,}09}{0{,}39 \times 0{,}67} \approx 0{,}34 \tag{15.60}$$

esta correlación se encuentra claramente alejada de un valor absoluto de 1, lo cual sugiere la existencia de interacción genotipo-ambiente en este caso.

De forma análoga, tenemos para el caso del largo de antena

$$\rho_x = \frac{0{,}149}{0{,}34 \times 0{,}45} \approx 0{,}97 \tag{15.61}$$

En el caso de largo de antena observamos que la estimación de correlación obtenida es altísima, a efectos prácticos considerable en 1. Esto no permite sugerir la existencia de una interacción genotipo-ambiente en el caso de este carácter.

Ejercicio 15.5

(Inspirado en Feldman y Lewontin 1975)

Supongamos que varios loci influencian a un carácter fenotípico que controla la producción de leche de un mamífero en poblaciones naturales, teniendo una altísima heredabilidad. La norma de reacción de los dos genotipos existentes frente a la temperatura (en grados Celsius) es la siguiente

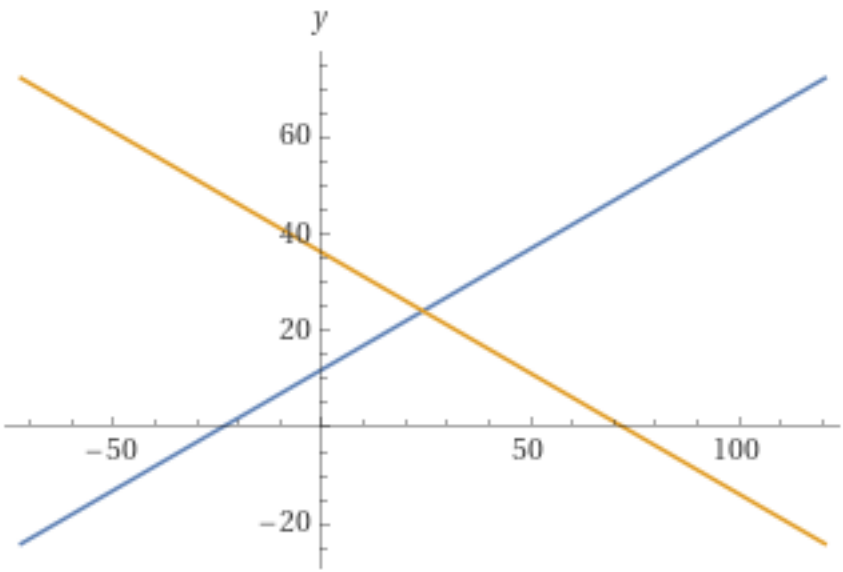
$$\begin{cases} \text{Genotipo 1} : y = [0{,}5T + 12]L/da \\ \text{Genotipo 2} : y = [-0{,}5T + 36{,}2]L/da \end{cases} \tag{15.62}$$

Para sorpresa de un grupo de investigadores, un estudio del carácter fenotípico a lo largo de un área estudiada muestra un promedio estable, a pesar de que se observa variación en el carácter fenotípico. ¿Cómo es esto posible, y cuál es el valor medido por los investigadores?

Pista: Es conveniente graficar ambas normas de reacción para observar su comportamiento.

Solución

A continuación se observa el resultado de graficar ambas normas de reacción



Como se puede observar, ambas normas de reacción parecen simétricas respecto a un punto de intersección. ¿Qué implica esto? A medida que cambia un rango de temperaturas, uno de los genotipos aumentará el valor del carácter fenotípico exactamente lo mismo que lo disminuye el otro. Si en una población se observa el reemplazo de un genotipo por el otro, pero esto se ve acompañado por una disminución del carácter fenotípico (producción de leche), entonces el promedio en la población se mantendrá estable aun cuando varíe el genotipo.

El valor en el que se debería encontrar este promedio que se mantiene estable se obtiene de igualar ambas ecuaciones. Tenemos

$$0{,}5T + 12 = -0{,}5T + 36{,}2 \tag{15.63}$$

$$1T = 36{,}2 - 12 \tag{15.64}$$

$$T = 24{,}2 \tag{15.65}$$

por lo que estimamos que los investigadores miden aproximadamente 24,2 litros de leche por día en cada animal lactante.

Apéndice A: Conceptos matemáticos básicos

Fracciones

Una fracción, representada de la forma $\frac{n}{m}$, es un cociente entre dos números enteros (n y m) que representa un valor numérico. Al número m se lo llama **denominador** e indica en cuántas partes iguales se divide una unidad. Por su parte n es el **numerador** e indica cuántas partes se deben tomar.

Fracciones equivalentes y fracciones irreducibles

Se dice que dos fracciones son equivalentes si representan el mismo número, por más que tengan distinto numerador y denominador. Para comprobar si dos fracciones son equivalentes se utiliza el método de productos cruzados. Se multiplica el numerador de una con el denominador de la otra y viceversa, y si ambos productos dan igual, las fracciones son equivalentes. Cada fracción posee infinitas fracciones equivalentes a ella y podemos obtenerlas mediante dos métodos: *amplificación* y *simplificación*.

La amplificación consiste en la multiplicación de numerador y denominador por un mismo número mientras que la simplificación consiste en la división de los factores por un número que se podrá lograr únicamente si ambos son divisibles por el mismo. Si el número entre el que se divide es el máximo común denominador, se obtiene una *fracción irreducible*. Una fracción irreducible es aquella que no se puede simplificar y esto sucede cuando numerador y denominador son primos entre sí.

Operaciones con fracciones

Sean $\frac{n}{m}$ y $\frac{p}{q}$ dos fracciones. Se definen las siguientes operaciones entre ellas:

Suma:

$\frac{n}{m} + \frac{p}{q} = \frac{nq+pm}{mq}$ donde m y q son dos números naturales distintos.

En caso de ser iguales, basta con sumar los numeradores y el denominador del resultado será el denominador de los sumandos. Se escribe de la forma $\frac{n}{m} + \frac{p}{m} = \frac{n+p}{m}$

Ejemplo: Un campo tiene un total de 3500 ha y se encuentra dividido en diversas fracciones. Como superficie cultivable se cuenta con dos fracciones que poseen $\frac{1}{7}$ y $\frac{3}{17}$ del total del campo. ¿Cuál es la superficie cultivable?

La fracción de la superficie total que es cultivable es la suma de las dos fracciones, es decir $\frac{1}{7}+\frac{3}{17} = \frac{1\times17+3\times7}{7\times17} = \frac{38}{119}$. Como la superficie total es de 3500 ha, entonces dicha superficie será $3500 \times \frac{38}{119} = \frac{3500\times38}{119} \approx 1117{,}6$ ha

Resta:

Al igual que la suma, se definen dos casos diferenciando si las fracciones tienen igual denominador o distinto. Si m y q son distintos, se define la resta como $\frac{n}{m} - \frac{p}{q} = \frac{nq-pm}{mq}$ mientras que, si son iguales, el resultado de la resta es $\frac{n-p}{m}$

Ejemplo: Las reservas forrajeras del establecimiento son de 1200 fardos, que asumimos vamos a utilizar durante tres meses en igual proporción. Sin embargo, por razones ajenas a nuestra voluntad, antes de entrar en el último de los meses debimos vender la cuarta parte del total de nuestras reservas. ¿Con cuánto forraje me quedé para el último mes?

Al último mes le correspondería un tercio de las reservas originales, pero debimos vender un cuarto antes, por lo que me queda para este último mes:

$\frac{1}{3} - \frac{1}{4} = \frac{1\times4-1\times3}{3\times4} = \frac{4-3}{12} = \frac{1}{12}$

Como el total eran 1200 fardos, tenemos que $1200 \times \frac{1}{12} = \frac{1200}{12} = 100$ fardos.

Multiplicación:

$\frac{n}{m} \times \frac{p}{q} = \frac{np}{mq}$

Ejemplo: Las reservas forrajeras del establecimiento son de 1200 fardos, que asumimos vamos a utilizar durante tres meses en igual proporción, pero que debemos distribuirlas en forma equitativa entre dos rodeos de cría separados que posee el establecimiento. ¿Cuánto forraje recibe cada rodeo mensualmente?

La fracción mensual que recibe cada rodeo sería la mitad de un tercio del total, es decir $\frac{1}{3}\frac{1}{2} = \frac{1\times1}{3\times2} = \frac{1}{6}$. Como el total de las reservas era de 1200 fardos, entonces cada rodeo recibirá mensualmente $1200 \times \frac{1}{6} = \frac{1200}{6} = 200$ fardos.

División:

La división de fracciones consiste en la multiplicación de la fracción que corresponde al dividendo por el inverso a la que corresponde al divisor; o lo que es lo mismo, es la multiplicación cruzada de numerador y denominador. Se expresa de la siguiente manera: $\frac{\frac{n}{m}}{\frac{p}{q}} = \frac{nq}{mp}$

Común denominador

Transformar varias fracciones a común denominador consiste en convertir cada una de ellas en otra fracción equivalente, pero buscando que todas tengan el mismo denominador. A continuación, se presentan dos métodos para hallar el común denominador:

a) Multiplicar los denominadores entre sí:

Para aplicar este método lo primero que se hará es multiplicar entre sí los denominadores de las fracciones que queremos convertir. A continuación, será necesario modificar los numeradores para no modificar el valor de la fracción; para ello, se divide el denominador común encontrado entre cada uno de los denominadores de cada fracción inicial. Por último, el

resultado de cada una de esas divisiones se multiplica por el numerador de cada fracción inicial.

Este método a pesar de ser práctico puede resultar en trabajar con números bastante grandes lo que puede complicar futuros cálculos.

Ejemplo: Sumar las fracciones $\frac{3}{12}$ y $\frac{4}{18}$. El producto de los denominadores es $12 \times 18 = 216$. Para mantener las fracciones originales expresadas en este denominador, hacemos para la primera $216/12 = 18$ que sería el número por el que también hay que multiplicar el numerador si el nuevo denominador va a ser 216, es decir $3 \times 18 = 54$ y por lo tanto la primera fracción quedará expresada como $\frac{54}{216}$

Para la segunda tenemos que $216/18 = 12$ y por lo tanto el numerador será $4 \times 12 = 48$ y la fracción quedará expresada como $\frac{48}{216}$. Ahora podemos sumar directamente las fracciones porque tienen el mismo denominador y por lo tanto $\frac{3}{12} + \frac{4}{18} = \frac{54}{216} + \frac{48}{216} = \frac{102}{216}$

b) Mínimo común múltiplo (mcm)

Para aplicar este método lo primero que se debe hacer es descomponer cada denominador en factores primos.

A continuación, será necesario multiplicar todas las potencias producto de la descomposición, pero en caso de existir bases repetidas únicamente se utilizará la de mayor exponente.

Para modificar los numeradores se aplica el mismo proceso que en el método anterior.

El mcm tiende a ser un método más laborioso que el primero, sin embargo, nos permite operar con valores más pequeños.

Ejemplo: Como en el ejemplo anterior, sumar las fracciones $\frac{3}{12}$ y $\frac{4}{18}$.

$12 = 2 \times 2 \times 3 = 2^2 \times 3$, mientras que $18 = 2 \times 3 \times 3 = 2 \times 3^2$. O sea, tenemos los números primos 2 y 3 constituyendo ambos denominadores. La mayor potencia de ambos es 2, por lo que el **mcm** será $2^2 \times 3^2 = 36$. Claramente $36 = 12 \times 3$, mientras que $36 = 18 \times 2$. Ahora, tenemos que $\frac{3}{12} \equiv \frac{3}{3}\frac{3}{12} = \frac{9}{36}$, mientras que $\frac{4}{18} \equiv \frac{2}{2}\frac{4}{18} = \frac{8}{36}$. Sumando directamente (porque los denominadores son iguales), tenemos $\frac{9}{36} + \frac{8}{36} = \frac{17}{36}$

Como verificación de que por ambos métodos llegamos a fracciones equivalentes, podemos notar que en $\frac{102}{216}$ tanto numerador como denominador son divisibles entre 6 (porque ambos son pares y además en ambos la suma de dígitos es divisible entre 3), por lo que $102/6 = 17$ y $216/6 = 36$ y por lo tanto $\frac{102}{216} \equiv \frac{17}{36}$, que es lo mismo que obtuvimos por el método del **mcm**.

Saber aplicar todas las transformaciones detalladas en esta sección será fundamental al momento de operar con fracciones.

Operaciones combinadas con fracciones

Al momento de realizar operaciones combinadas, las prioridades a seguir son las mismas que cuando se calculan utilizando números enteros, así pues:

1) Calcular potencias y raíces.
2) Realizar las operaciones entre paréntesis o llaves.
3) Calcular los productos y cocientes.
4) Calcular sumas y restas.

Proporcionalidad

Dos magnitudes mantienen una proporcionalidad directa si cuando una aumenta la otra también lo hace a igual relación ocurriendo lo mismo si se reduce una de ellas. A modo de ejemplo, si una se divide a la mitad, la otra también. Se dice que dos magnitudes x e y mantienen una proporcionalidad directa si hay una constante a, llamada **constante de proporcionalidad** tal que $y = ax$

Regla de tres

La regla de tres es un método utilizado para resolver problemas de proporcionalidad. Se utiliza para hallar el cuarto término de una proporción. Se basará la explicación del método con un ejemplo.

Ejemplo:
Sabiendo que una guitarra tiene 6 cuerdas, determinar cuántas cuerdas necesito si quiero cambiar completamente el encordado de 17 guitarras.

Solución
Para resolver este problema, se debe plantear la regla de tres simple. Para eso, se disponen los datos en forma de tabla, teniendo en cuenta que en una misma columna se deben disponer datos de un mismo tipo (en el ejemplo, una columna lleva los datos de las guitarras y la otra las cuerdas). El valor a hallar se representará con una x que simboliza la incógnita.

Guitarras	Cuerdas
1	6
17	x

Para hallar el valor de x, se realiza la siguiente operación: $x = \frac{17 \times 6}{1} = 102$

Generalizando entonces, la regla de tres se plantea de la siguiente forma:

a_1	b_1
a_2	x

$\Rightarrow x = \frac{a_2 b_1}{a_1}$ donde a_1 y a_2 son del mismo tipo y las magnitudes a y b son directamente proporcionales.

Cifras significativas y redondeo

Las cifras significativas son los números que se obtienen al realizar una medición. Se encuentran limitados por el instrumento de medición. Las cifras significativas ciertas son las que se pueden observar directamente del instrumento de medición, mientras que las dudosas son las que se estiman. Una medida no puede tener más de una cifra significativa dudosa.

Por ejemplo, si se tiene una balanza con aguja cuyas divisiones son de un kilogramo y la aguja marca entre los 75 y 76 kg, la medida deberá ser expresada con un solo decimal (75,5 kg a modo de ejemplo).

Los ceros a la izquierda del valor no serán considerados como cifras significativas.

Para reducir la cantidad de cifras utilizadas para expresar una medida y que quede con el número correcto de cifras significativas se debe redondear, siguiendo las siguientes reglas: 1) Si el número a eliminar es menor a 5, la cifra a su izquierda no se modifica. 2) Si el número a eliminar es mayor o igual a 5, la cifra a su izquierda incrementa en uno.

Ejemplo: Se tiene el número 3,4581 para redondearlo a distintas cantidad de cifras significativas, como se muestra en la siguiente tabla:

Núm. Cifras significativas	Resultado	Regla utilizada
1	3	Regla 1
2	3,5	Regla 2
3	3,46	Regla 2
4	3,458	Regla 1

Operaciones con cifras significativas

Al operar con resultados de mediciones, es importante tener en cuenta las cifras significativas dado que los resultados no pueden tener mayor precisión que las medidas originales. Dependiendo de la operación a realizar, se siguen distintas reglas.

Suma y resta

Para el caso de la suma y la resta, lo primero que se debe hacer es redondear todas las medidas involucradas en la operación hasta que tengan igual cantidad de decimales que aquel número con menos decimales. Luego se realiza la operación buscando mantener la misma cantidad de decimales que los números a sumar (o restar).

Ejemplo: Se quiere sumar las siguientes masas: 21,34 kg, 5,3 kg y 9,37 kg. Primero se redondeará todas las medidas para que queden expresadas con un decimal.

Medida	Redondeo	Regla utilizada
21,34 kg	21,3 kg	Regla 1
5,3 kg	5,3 kg	No se redondea
9,37 kg	9,4 kg	Regla 2

Hechos todos los redondeos pertinentes, se procede a sumar

$$21{,}3\ \text{kg} + 5{,}3\ \text{kg} + 9{,}4\ \text{kg} = 36{,}0\ \text{kg} \tag{15.66}$$

Producto y división

Para el producto y la división, se realiza primero la operación para luego redondear el resultado hasta que tenga la misma cantidad de cifras significativas que las medidas utilizadas.

Ejemplo: Se quiere calcular la fuerza sobre un elemento de masa 9,1 kg que tiene una aceleración de 3,12 m/s^2 Primero se realiza la operación:

$$3{,}12\,\text{m/s}^2 \times 9{,}1\,\text{kg} = 28{,}392\,\text{N}$$

Luego se debe redondear el resultado hasta obtener la misma cantidad de cifras significativas que el operando con menos cifras significativas. En este caso, es 9,1 que tiene dos, por lo que el resultado se expresa: 28 N (utilizando la regla 1 de redondeo).

Notación científica

La notación científica se utiliza para expresar medidas en potencias de base 10, con especial utilidad para expresar medidas muy grandes o muy chicas. El formato para expresar un número en notación científica es el siguiente:

$$A \times 10^n$$

donde A es un número entre 1 y n indica la cantidad de veces que se debe correr la coma decimal de un número para obtener el valor de A. Si n es positivo, el resultado es un número mayor a 1 (por ejemplo, con $n = 3$ se deben agregar 3 ceros a A). En cambio, si n es negativo, el resultado es un número menor a 1, corriendo la coma n lugares a la izquierda.

Ejemplos:

Notación decimal	Notación científica
0,03	3×10^{-2}
6000	6×10^3

Ecuaciones

Las ecuaciones son expresiones que se utilizan para modelar diversas situaciones y problemas. Se definen como una igualdad entre dos expresiones conteniendo uno (ecuación de una variable) o más (ecuación en varias variables) valores desconocidos. A los valores desconocidos se los denomina **incógnitas** y suele representarse con la letra x.

Se le llama **solución** de la ecuación a todo número que haga que se cumpla la igualdad. Puede existir una única solución, múltiples soluciones o ninguna. Cuando una ecuación es utilizada para modelar un escenario concreto y tiene más de una solución, deberán elegirse las soluciones que tengan sentido en el contexto del problema (descargando el resto).

Se dice que dos ecuaciones son equivalentes si tienen el mismo conjunto solución. La clave para resolver una ecuación es reducirla hasta la ecuación equivalente más simple utilizando la **propiedad uniforme**. Esta establece que «si se realiza la misma operación con el mismo número a los dos lados de la igualdad, se mantiene la igualdad». Coloquialmente, esto implica que se pueden «pasar» números de un lado a otro de la ecuación realizando la operación opuesta (si está sumando de un lado, al otro pasa restando). Se utiliza la propiedad hasta tener «despejada» la incógnita, que implica que las incógnitas quedan de un lado de la ecuación.

Ecuaciones polinómicas

Una ecuación es de grado *n* si es de la forma $ax^n + bx^{n-1} + ... + cx + d = 0$ con $a \neq 0$. Este polinomio, de tener solución, tendrá como máximo *n* soluciones. En particular, a continuación se explica cómo hallar las soluciones de una *ecuación cuadrática* (ecuación de segundo grado, $n = 2$) que es de la siguiente forma

$$ax^2 + bx + c = 0$$

Existen tres escenarios posibles: 1) b es igual a cero. La ecuación es de la forma $ax^2 + c = 0$ y se resuelve como si fuera una ecuación con una sola incógnita recordando que la operación opuesta a la potencia es la raíz. En ese caso, $x = \pm\sqrt{\frac{-c}{a}}$. Es importante notar que si $\frac{-c}{a}$ es un número menor a cero, la ecuación no tiene solución. 2) c es igual a cero. La ecuación es de la forma $ax^2 + bx = 0$. Se saca x de factor común, por lo que $x = 0$ es una de las soluciones del problema. la otra se halla con el método explicado anteriormente, llegando a qué $x = \frac{-b}{a}$. 3) Todos los coeficientes son no nulos. La ecuación es de la forma $ax^2 + bx + c = 0$. Las dos raíces, si existen, se hallan aplicando la fórmula de Bhaskara. Esta fórmula dice lo siguiente:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Si el término $b^2 - 4ac$ es mayor a cero se puede afirmar que existen soluciones a la ecuación de segundo grado. Si es igual a cero, se tiene una única solución (a la que se le llama **solución doble**). Si es menor a cero, no existen soluciones.

Ejemplos:

Se tiene un campo de 35 has distribuidas en 700 m × 500 m. Se quiere comprar x m de largo y x m de ancho suficientes para extender el terreno a aproximadamente 50 has. ¿Cuánto debe valer x?

Las nuevas medidas del campo serán

$$(700 + x)(500 + x) = 50\,\text{ha} = 500.000\,\text{m}^2$$

Primero se calcula el producto de los binomios, donde se obtiene $350e^3 + 1200x + x^2 = 500e^3$. Operando, se llega a la siguiente ecuación $-150e^3 + 1200x + x^2 = 0$, a la que, al ser una ecuación de segundo grado con todos sus coeficientes no nulos, se le aplicará la fórmula de Bhaskara para hallar la solución.

De ahí salen dos soluciones,
$x_1 = \frac{-1200+\sqrt{20.400.000}}{2} = 114\,\text{m}^2$ y $x_2 = \frac{-1200-\sqrt{20.400.000}}{2} = -1314$

La solución x_2 es un número negativo que será descartado debido a que las medidas siempre son números mayores a cero. Por lo tanto, serán necesarios agregar 114 m de largo y 114 m de ancho para obtener un campo de aproximadamente 50 has. Para saber la superficie final obtenida, basta con sustituir x_1 en la ecuación inicial:

$$\text{Superficie} = (700 + 114)(500 + 114) = 49{,}98\,\text{ha}$$

Funciones

Una función es una regla que asigna a cada elemento de un conjunto A un elemento de un conjunto B. Se escribe

$$\begin{aligned} f : A &\to B \\ x &\to f(x) \end{aligned}$$

Lo anterior se lee «f es una función de A a B», donde a cada valor de $x \in A$ se le asigna un valor de B representado por $f(x)$ (que se lee «f de x»). Al conjunto A se lo denomina **dominio** o **conjunto de partida** mientras que a B se le llama **codominio** o **conjunto de llegada**. Se dice que una cantidad y es una función de otra cantidad x si el valor de la primera depende del valor que tome la segunda y se escribe $y = f(x)$

Para definir una función, hay dos condiciones que deben cumplirse: **existencia** y **unicidad**. Para cada valor x en el dominio debe existir un único valor $f(x)$ en el codominio que sea imagen de x.

Existen tres formas de representar las funciones: con una ecuación, una tabla o una gráfica. La más utilizada y con la que se trabajará en este libro principalmente es con la ecuación, ya que a partir de ella se puede llegar a las otras dos.

Ceros de una función

Hallar los ceros (también llamados raíces) de una función significa tomar la ecuación que la define e igualarla a cero para hallar aquellas x para las cuales la función vale cero. Matemáticamente, se definen como aquellos x tal que $f(x) = 0$. Es importante tener en cuenta que no siempre es sencillo (o posible) hallar las raíces de manera analítica, sino que puede ser necesario utilizar métodos numéricos para los cálculos.

El caso más sencillo es el de funciones polinómicas de primer y segundo orden. En esos casos, hallar los ceros implica hallar directamente las raíces del polinomio.

Tipos de funciones

Función polinómica

Una función polinómica de grado n (donde $n \in N \cup \{0\}$) es una función definida por una ecuación polinómica, tal que

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$$

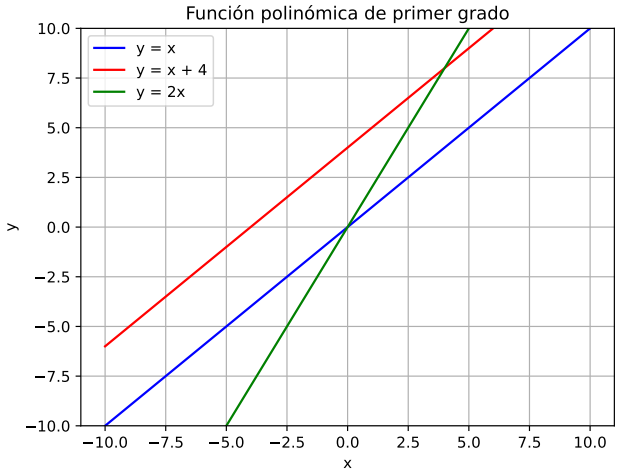
donde a_n, a_{n-1},..., a_1, $a_0 \in R$ y se denominan coeficientes de la función polinómica. Una función polinómica tendrá como dominio a todo R.

a) Polinomios de primer grado

Un polinomio de primer grado es una función de forma $f(x) = ax+b$ (también denominado *función lineal*) donde su representación gráfica será siempre una recta. Al parámetro a se lo denomina *pendiente de la recta* y marca su verticalidad. Si $a = 0$ la representación de la función es una recta horizontal que pasa por el punto b, mientras que, si a es un número muy grande, la recta es vertical (establece un ángulo de 90° respecto al eje horizontal). Al parámetro b se lo denomina intercepto y establece a qué altura del eje vertical cortará la recta. Si $b = 0$ la recta pasará por el origen. Si es mayor, la recta pasará por encima del origen mientras que, si es menor, pasará por debajo.

Las funciones lineales tienen una única raíz, ya que el polinomio tiene tantas raíces como su orden.

En la figura se muestran tres ejemplos de funciones lineales, donde se puede observar las diferencias que resultan de establecer distintos valores para a y b. Para la función $y = x$ e $y = 2x$ el cero de se da en $x = 0$. Para $y = 2x + 4$ el cero se encuentra en $x = -4$.



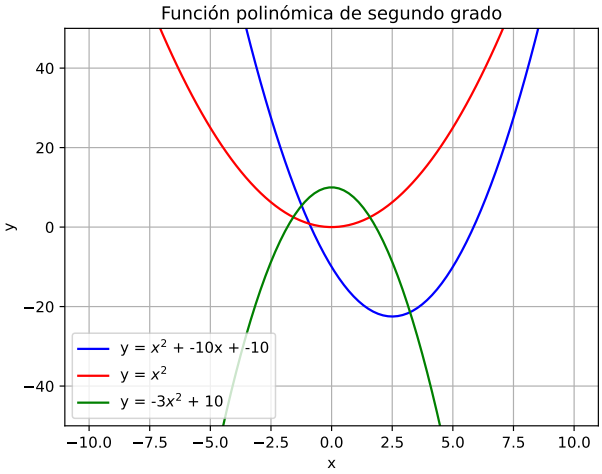
b) Polinomios de segundo grado

Un polinomio de segundo grado es una función de forma $f(x) = ax^2 + bx + c$ donde $a \neq 0$

Su representación gráfica será siempre una parábola cuyo vértice estará por debajo de la parábola si $a > 0$ mientras que para $a < 0$ se encontrará por encima de la misma. El vértice se sitúa en el punto $x = \frac{b}{2a}$. En la figura se muestran tres ejemplos de parábolas, donde se puede ver qué ocurre con el vértice cuando a toma valores positivos o negativos. También se muestra el efecto que tiene tener distintos valores de b y c. En este caso, se puede ver que cada función tiene dos raíces.

Las funciones polinómicas de segundo orden tienen dos posibles x para cada $f(x)$. Es importante destacar que esto no se contrapone con la *unicidad* descrita previamente, que dice que para un valor x solo puede existir un solo $f(x)$. Para resolución de problemas de la vida real, es importante tener en cuenta que existen dos puntos x preimagen de $f(x)$ ya que dependiendo del campo de aplicación puede ser necesario elegir uno de ellos.

En el ejemplo utilizado para ecuaciones polinómicas, el valor x negativo fue descartado ya que una medida de tierra siempre será positiva. Este tipo de análisis suelen ser muy frecuentes al momento de trabajar con funciones.



Función racional

Una función racional es una función del tipo $h(x) = \frac{f(x)}{g(x)}$

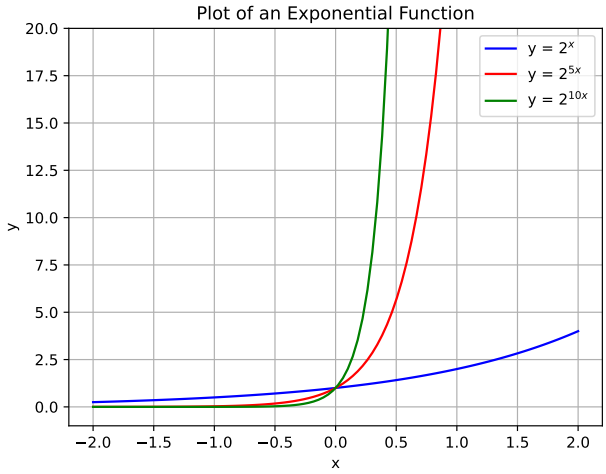
en donde f y g son funciones polinómicas. Podemos definir entonces la función racional como el cociente de dos funciones polinómicas. El dominio de este tipo de funciones será R exceptuando los valores en los que $g(x)$ es nula.

Función exponencial

Una función exponencial es de la forma $f(x) = a^{bx}$, donde $a \neq 1$ (dado que $1^x = 1$) y $a > 0$. La variable de la función aparece en el exponente. Tiene gran importancia para el modelado de distintos fenómenos, desde económicos hasta fenómenos de la naturaleza (como el crecimiento de una colonia de bacterias).

Estas funciones presentan un crecimiento más rápido en comparación con el de las funciones lineales a medida que x crece, denominándose **crecimiento exponencial**. Además, toman siempre valores positivos para cualquier x.

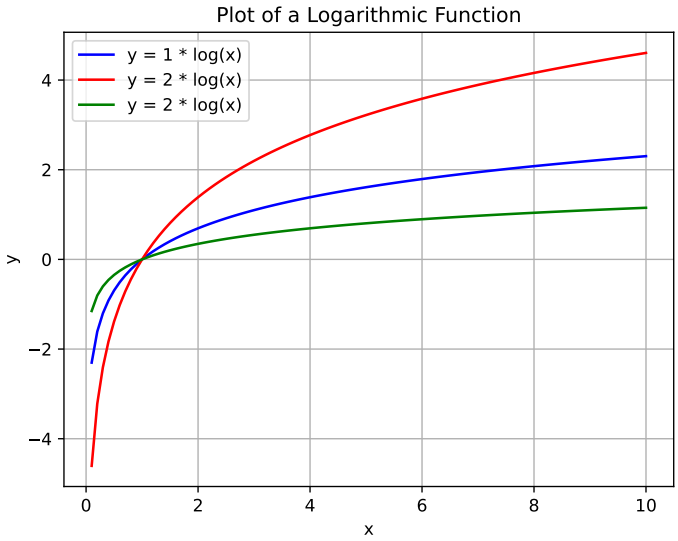
En la figura se observan tres curvas exponenciales. Hacia la izquierda (cuánto más negativo el eje x) la curva se acerca más a cero. Toma valores más grandes cuánto más grande es x. Además, más grandes son los valores adoptados por la función cuánto mayor es b. En todos los casos, para $x = 0$ el valor de la función es 1. Estas funciones no tienen ceros, se acercan cuando mayor es el valor absoluto de los $x < 0$ pero nunca llegan a igualarlo.



Función logarítmica

La función logarítmica es aquella en la que la variable aparece en el argumento del logaritmo, por lo que son de la forma $f(x) = log_a x$ con la base $a \neq 1$ y $a > 0$. Teniendo en cuenta que el logaritmo solo existe para números positivos, el dominio de la función deberá pertenecer al conjunto de R^+.

En la figura se muestran tres funciones logarítmicas. Se observa que están definidas únicamente para $x > 0$



Sumatoria y productoria

Se define la **sumatoria** como un operador matemático utilizado para expresar la suma de una gran cantidad de números que sigue una ley general de formación. Se representa con la letra griega $\sum$. Formalmente, se define como

$$\sum_{i=m}^{n} f(i) = f(m) + f(m+1) + ... + f(n)$$

donde f es una función, i es el índice de la sumatoria, m su límite inferior y n el límite superior. Esto significa que el índice tomará valores entre desde m hasta n y se sumarán todos los valores que toma la función al evaluar el índice.

Ejemplos:

1)
$$\sum_{i=2}^{5} i = 2 + 3 + 4 + 5 = 14$$

2)
$$\sum_{i=2}^{5} i^2 = 4 + 9 + 16 + 25 = 54$$

3)
$$\sum_{i=1}^{4} 2(i+8) = 18 + 20 + 22 + 24 = 84$$

4)
$$\sum_{i=0}^{7} 2^i = 1 + 2 + 4 + 8 + 16 + 32 + 64 + 128 = 255$$

La **productoria** es el análogo a la sumatoria, pero para representar el producto de una gran cantidad de números que siguen una ley general de formación. En este caso, se utiliza la letra griega $\prod$ para su representación.

Ejemplos:

1)
$$\prod_{i=2}^{5} i = 2 \times 3 \times 4 \times 5 = 120$$

2)
$$\prod_{i=0}^{3} 2^i = 1 \times 2 \times 4 \times 8 = 64$$

Apéndice B: Cálculo diferencial e integral

Este apéndice se inspiró en material generado por el IMERL (Facultad de Ingeniería), el cual se encuentra disponible para su libre consulta [ver Castigeras (n.d.) e IMERL (n.d.c)].

Límites

Límite finito

Se considera $f : X \subset \mathbb{R} \to \mathbb{R}$ con X una unión de intervalos. El límite de la función f en un punto a refiere a qué pasa con los valores de $f(x)$ para valores de x muy cercanos a a. Se escribe de la siguiente forma:

$$\lim_{x \to a} f(x) = L$$

Cuando se estudia el límite en a, se busca saber cómo se comporta la función en puntos cercanos a él, pero no importa específicamente la imagen de la función en a. Esto implica que a no necesariamente tiene que ser parte del dominio de la función. El límite de una función en un punto a nos indica el comportamiento de f cerca de a sin importar el punto a.

Sea L el límite de f en a, pueden darse tres situaciones respecto a la imagen $f(a)$:
1. Si f continua en a, entonces $L = f(a)$
2. Si f no es continua en a, entonces L es el límite pero $L \neq f(a)$
3. Si f no está definida en a, existe el límite L pero no está definido $f(a)$

Proposiciones importantes:

Proposición 1 Si el límite $\lim_{x \to a} f(x)$ existe, es único.

Proposición 2 Sean $\lim_{x \to a} f(x) = b$ y $\lim_{x \to a} g(x) = c$
Entonces
1. $\lim_{x \to a} (f + g)(x) = b + c$
2. $\lim_{x \to a} (f - g)(x) = b - c$
3. $\lim_{x \to a} (f \times g)(x) = b \times c$

Límite infinito

Intuitivamente, se habla de un límite infinito cuando a medida que la función se acerca al punto a, la imagen $f(a)$ es un valor muy grande. Cuanto más se acerca x a a, $f(x)$ es un número cada vez mayor, pero si $x = a$ la operación es indeterminada (por ejemplo, queda una división entre cero).

También puede ser importante estudiar el límite de la función cuando $x \to \pm\infty$ (valores de x con módulo muy grande, tanto positivos como negativos).

Para ver los distintos casos, se trabajará con un ejemplo.

Ejemplo:

Se define

$$f(x)/f(x) = \frac{1}{1-x} \forall x \in \{\mathbb{R}\} - \nVdash.$$

¿Qué pasa con el límite en 1? ¿Y en $\pm\infty$?

Solución:

A) Para saber el límite de $x \to 1$ se divide en dos casos: 1. Por un lado se estudia el límite cuando nos acercamos a 1 por la derecha (con valores mayores a uno, pero cada vez más cercanos).

x	$f(x)$
1,01	-100
1,001	-1000
1,0001	-10000
1,00001	-100000

Se puede ver que cuánto más cerca está x de 1, mayor es el módulo de $f(x)$. Se puede ver también que $f(x)$ es negativa, por lo que se dice que el límite de $f(x)$ cuando x tiende a 1 por derecha es $-\infty$ y se escribe:

$$\lim_{x \to 1^+} f(x) = -\infty$$

2. Al estudiar el límite por izquierda, acercándonos a valores menores a él, se puede ver que

x	$f(x)$
0,9	10
0,99	100
0,999	1000
0,9999	10000

Cuánto más cerca de uno, mayor es el valor de $f(x)$ pero con valores positivos por lo que, en este caso:

$$\lim_{x \to 1^-} f(x) = +\infty$$

B) Se estudiarán ahora los límites para $\pm\infty$ separando nuevamente en los dos casos.

1. Para el caso $+\infty$, tomaremos valores de x muy grandes:

x	$f(x)$
1×10^5	-1×10^{-5}
1×10^{10}	-1×10^{-10}
1×10^{15}	-1×10^{-15}
1×10^{20}	-1×10^{-20}

En este caso, se puede ver que cuánto más grande el valor de x, más pequeño es el valor de $f(x)$ pero manteniéndose positivo. Decimos entonces que el límite es 0^+.

$$\lim_{x \to +\infty} f(x) = 0^+$$

2. Realizando la misma tabla que en el caso anterior pero para valores negativos, se obtiene que $f(x)$ se acerca más a cero cuanto mayor el valor de x, pero en este caso los valores obtenidos son negativos. Se dice que el límite es 0^-.

$$\lim_{x \to -\infty} f(x) = 0^-$$

Se muestra a continuación la gráfica de la función estudiada. En azul, se muestran las asíntotas $x = 1$ y $y = 0$ que son aquellas rectas que la función nunca corta.

Gráfica de límites$_{f_X}$

Se puede ver que cuanto más grandes en módulo son los valores de x más se acerca la función a cero. Si bien el eje x en esta imagen no llega a valores tan altos, invitamos al lector a verificar esto con su calculadora. Se puede observar lo mismo para $x = 1$, donde se puede observar que la función no llega a cortar la asíntota y toma valores cada vez más grandes.

Proposiciones importantes:

Proposición 3a: Sean $\lim_{x \to a} f(x) = +\infty$ y $\lim_{x \to a} g(x) = c$, entonces

Suma:

$$\lim_{x \to a} (f \times g)(x) = +\infty$$

Producto:
- Si $c > 0$, $\lim_{x \to a} (f \times g)(x) = +\infty$
- Si $c < 0$, $\lim_{x \to a} (f \times g)(x) = -\infty$

Proposición 3b: Sean $\lim_{x\to a} f(x) = \infty$ y $\lim_{x\to a} g(x) = \infty$, entonces

Suma:

$$\lim_{x\to a}(f+g)(x) = \infty$$

Producto:

$$\lim_{x\to a}(f\times g)(x) = \infty$$

Proposición 4a: Si $\lim_{x\to a} f(x) = \infty \Rightarrow \lim_{x\to a} \frac{1}{f(x)} = 0$

Proposición 4b: Si $\lim_{x\to a} f(x) = 0 \Rightarrow \lim_{x\to a} \frac{1}{f(x)} = \infty$

Derivadas

Derivada puntual

La derivada puntual describe con qué rapidez se produce la variación de una función para un determinado punto a. La definición formal de la derivada dice que f es derivable en a si existe el límite $\lim_{x\to a} \frac{f(x)-f(a)}{x-a}$. Si existe el límite, $f'(a)$ es el resultado y se dice que $f'(a)$ es la derivada de f en el punto a.

Gráficamente, se puede ver como la pendiente de la gráfica en el punto.

Derivada de una función

A partir de la definición anterior, se puede definir la derivada de una función. Tomando un punto x en el intervalo (a,b) y f una función definida, por lo menos, en ese intervalo abierto (a,b), se define la derivada f' como el límite del cociente de diferencias:

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$$

Este cociente representa la variación media de f en el intervalo que une a x y $x+h$. Este límite define una nueva función llamada *primera derivada de f* y, como fue mencionado previamente, se simboliza como f'.
Si se deriva f', obtenemos la segunda derivada de f, f''.

Ejemplo:

Se tiene una explotación ganadera que cría vacas para la producción de carne. Para obtener un modelo simple, se considera que la tasa de crecimiento de las vacas depende únicamente de la cantidad de alimento suministrado por día (que será una función de x).

Se tiene entonces que el crecimiento C depende de x:

$$C(x) = e^{3x} + x^2$$

Esta ecuación nos dice que el crecimiento queda ligado a la cantidad de alimento mediante una función exponencial más una función polinómica. La tasa de crecimiento se calculará a partir de las derivadas total de la función, obteniendo que $C'(x) = 3e^{3x} + 2x$

Máximos y mínimos locales y absolutos

Tomando la noción de la derivada como la pendiente de la curva en un punto, se puede ver que en aquellos puntos donde la derivada es cero la tangente es horizontal. Esto implica tres posibilidades para ese punto:
1. La función f tiene un *máximo local* en ese punto.
2. La función f tiene un *mínimo local* en ese punto.
3. La función f tiene un punto silla, que no es ni máximo ni mínimo de la función.

Un punto a es un máximo local si para un entorno alrededor, la función toma valores menores a $f(a)$. Si además $f(a)$ es el máximo de los máximos, es un *máximo absoluto*.

Lo inverso ocurre si en a hay un mínimo. Existe un entorno alrededor de él para el cual la función toma valores mayores a $f(a)$. Si además es el mínimo de los mínimos, la función presenta un *mínimo absoluto*.

En la siguiente imagen se muestra un ejemplo de una función que tiene un máximo relativo, un mínimo relativo y un punto silla. El mínimo relativo se encuentra en $(1,-2)$, el máximo relativo en $(-1{,}2)$ y el punto silla en $(0{,}0)$. Se puede observar como en los tres casos la tangente al punto es horizontal (derivada nula). Sin embargo, la función toma valores mayores y menores a los del entorno, por lo que no son extremos absolutos.

Gráfica de puntos críticos

Para hallar si un punto es mínimo o máximo se puede trabajar con dos métodos:

Método de la derivada primera

- Para que un punto sea máximo, los puntos adyacentes a él deben ser menores, a la vez que la pendiente a la izquierda de este será positiva y a la derecha negativa.
- Para que un punto sea mínimo, los puntos adyacentes a él deben ser mayores, a la vez que la pendiente a la izquierda de este será negativa y a la derecha positiva.

Establecido esto, el criterio de la primera derivada dice que una vez que se encuentra un punto donde $f'(a)=0$ se miran las derivadas a la derecha y a su izquierda, concluyendo:
- Si a la izquierda la derivada es positiva y a la derecha negativa, el punto es un máximo.
- Si por su parte, a la izquierda del punto la derivada es negativa y a la derecha positiva, el punto es un mínimo.

En el caso que a ambos lados del punto la derivada sea positiva o negativa, no hay ni máximo ni mínimo.

Ejemplo: Sea la función $f(x)=4x^2-8x$, su derivada es $f'(x)=8x-8$. Para hallar los candidatos a máximos y mínimos, igualamos a cero y obtenemos que $f'(x)=0 \Leftrightarrow x=1$. Sustituyendo en $f(x)$ obtenemos la otra coordenada del punto, por lo que pasaremos a estudiar si $(1,-4)$ es un máximo o un mínimo.

Para ver la derivada a los costados de $x=1$ tomamos un punto a la izquierda (por ejemplo, cero) y otro a la derecha (por ejemplo, dos) y evaluamos el signo de la derivada.

$$f'(0)=-8$$

$$f'(2)=8$$

Se observa que a la izquierda la derivada es negativa y a la derecha positiva, por lo que el punto $(1, -4)$ es un mínimo.

Método de la derivada segunda Para utilizar el método de la derivada segunda, se calcula la derivada de $f'(x)$ y se evalúa en x. Si $f''(x)$ es negativa, en x hay un máximo. Por el contrario, si $f''(x)$ es positiva, en x hay un mínimo.

Ejemplo:

Tomando nuevamente la función $f(x) = 4x^2 - 8x$, calculamos su derivada primera y derivada segunda:

$$f'(x) = 8x - 8 \Rightarrow f''(x) = 8$$

La derivada segunda es la función constante positiva, por lo que en $(1, -4)$ la función $f(x)$ presenta un mínimo.

Proposiciones importantes

Proposición 1: Álgebra de las derivadas
Sean $f, g : \mathbb{X} \to \mathbb{R}$ derivables en a.
Las funciones $f + g,\ f - g,\ fg,\ \frac{f}{g}$ con $g(a) \neq 0$ son derivables en a y:

1. $(f + g)'(a) = f'(a) + g'(a)$
2. $(f - g)'(a) = f'(a) - g'(a)$
3. $(fg)'(a) = f'(a)g(a) + g'(a)f(a)$
4. $(\frac{f}{g})'(a) = \frac{f'(a)g(a)+g'(a)f(a)}{g^2(a)}$

Ejemplos:

1. Se tiene $f(x) = x$ y $g(x) = x^2$, por lo que $(f + g)(x) = x + x^2$.
Entonces, la derivada $(f + g)'(x) = 1 + 2x$, donde $f'(x) = 1$ y $g'(x) = 2x$

2. Se tiene $f(x) = e^x$ y $g(x) = 4x$, por lo que $(f - g)(x) = e^x - 4x$.
Entonces, la derivada $(f - g)'(x) = e^x - 4$, donde $f'(x) = e^x$ y $g'(x) = 4$

3. Se tiene $f(x) = e^x$ y $g(x) = x^2$, por lo que $(fg)(x) = x^2e^x$.
Entonces, como $f'(x) = e^x$ y $g'(x) = 2x$, la derivada $(fg)'(x) = x^2e^x + 2xe^x$

4. Se tiene $f(x) = log(x)$ y $g(x) = x^2$, por lo que $(\frac{f}{g})(x) = \frac{e^x}{x^2}$.
Entonces, como $f'(x) = \frac{1}{x}$ y $g'(x) = 2x$, la derivada $(\frac{f}{g})'(x) = \frac{\frac{x^2}{x} - 2xlog(x)}{x^4}$,
que simplificando queda: $(\frac{f}{g})'(x) = \frac{1-2log(x)}{x^3}$

En todos los casos, para obtener la derivada de cada una de las funciones en el punto a se evalúa en $x = a$

Proposición 2: Regla de la cadena

Sean $f : \mathbb{X} \to \mathbb{R}$ y $g : \mathbb{Y} \to \mathbb{R}$.
Además, $a \in X$ y $b\ inY$, $f(a) = b$. Si f es derivable en a y g derivable en b,
entonces $(g \circ f)'(a) = g'(f(a)) \times f'(a)$

Ejemplo:

Sean $f(x) = 4x^2 - 6x - 5$ y $g(x) = x^8$
La función compuesta $h(x) = (g \circ f)(x) = (4x^2 - 6x - 5)^8$
Aplicando la regla de la cadena, $h'(x)$ será $g'(x)$ evaluada en la función f y luego multiplicada por $f'(x)$
Se obtiene entonces - $g'(x) = 8x^7 \Rightarrow g'(f(x)) = 8(4x^2 - 6x - 5)^7$ - $f'(x) = 8x - 6$

$$\Rightarrow h'(x) = g'(f(x))f'(x) = 8(4x^2 - 6x - 5)^7(8x - 6)$$

Para obtener la derivada en el punto a se evalúa en $x = a$

Tabla de derivadas

A continuación, se presenta una tabla con las derivadas típicas, para facilitar los cálculos. Es importante destacar que tanto u como v son funciones de x, pero se denotan como u y v por simplicidad en la notación.

Simples		Compuestas	
Función	*Derivada*	*Función*	*Derivada*
k	0		
x^n	nx^{n-1}	u^n	nuu'
$\frac{1}{x}$	$\frac{-1}{x^2}$	$\frac{1}{u}$	$\frac{-u'}{u^2}$
$\sqrt[n]{x}$	$\frac{1}{n-1\sqrt[n]{x^{n-1}}}$	$\sqrt[n]{u}$	$\frac{u'}{n-1\sqrt[n]{u^{n-1}}}$
e^x	e^x	e^u	$u'e^u$
a^x	$a^x lna$	a^u	$u'a^u lna$
lnx	$\frac{1}{x}$	lnu	$\frac{u'}{u}$
$log_a x$	$\frac{1}{xlna}$	$log_a u$	$\frac{u'}{ulna}$
$sen(x)$	$cos(x)$	$sen(u)$	$u'cos(u)$
$cos(x)$	$-sen(x)$	$cos(u)$	$-u'sen(u)$
$tan(x)$	$\frac{1}{cos^2(x)}$	$tan(u)$	$\frac{u'}{cos^2(u)}$
		$u + v$	$u' + v'$
		$u - v$	$u' - v'$
		$u \times v$	$u'v + uv'$
		$\frac{u}{v}$	$\frac{u'v - uv'}{v^2}$

Derivadas parciales

Para trabajar con derivadas de funciones definidas en intervalos de $\mathbb{R}^n$ con $n > 1$ es necesario definir la *derivada parcial*. La definición formal para funciones de $\mathbb{R}^2$ dice que Sea $f : \mathbb{R}^2 \to \mathbb{R}$ y (x_0, y_0) un punto de $\mathbb{R}^2$. Se define la derivada parcial de f respecto a la variable x como el siguiente límite (si existe):

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

La definición es análoga para la derivada parcial de f respecto a la variable y. Se puede definir el límite para todas las variables que tenga la función en caso de ser una función en $\mathbb{R}^n$.

A nivel práctico, para calcular la derivada parcial respecto a una variable, se consideran las otras como constantes, que se sabe de secciones previas que su derivada es cero. Se muestran a continuación un ejemplo en $\mathbb{R}^3$.

Ejemplos: Sea la función $f(x,y,z) = x + xy + z$. Se calculan las derivadas parciales respecto a las tres variables:

$$\frac{\partial f}{\partial x}(x,y,z) = 1 + y$$

donde la derivada de z es cero porque la consideramos constante y la derivada de xy es y por el mismo motivo.

$$\frac{\partial f}{\partial y}(x,y,z) = x$$

$$\frac{\partial f}{\partial z}(x,y,z) = 1$$

Ejemplo de uso: Se decide ahora complejizar el modelo del que depende el crecimiento de las vacas para acercarse más a la realidad. Se considerará, además la cantidad de alimento suministrado por día (que será una función de x), las condiciones ambientales, como temperatura y humedad (que serán una función de y).

Se tiene entonces que el crecimiento C depende de x e y de la siguiente forma:

$$C(x,y) = e^{3x} + y^3$$

Esta ecuación nos dice que el crecimiento queda ligado a la cantidad de alimento mediante una función exponencial y a las condiciones ambientales por una función polinómica.

La tasa de crecimiento dependiendo de cada una de estas variables se calculará a partir de las derivadas parciales. Se podrá obtener la tasa de crecimiento asociado a la variación de la cantidad de alimento calculando $\frac{\partial C}{\partial x} = 3x^2 e^{3x}$ y la tasa asociada a las condiciones climáticas calculando $\frac{\partial C}{\partial y} = 3y^2$. Para calcular cada una de las derivadas, se considera constante la variable que no es de interés.

Integrales

El cálculo integral se utiliza fundamentalmente con dos fines: el primero consiste en la obtención de la función original a partir de su derivada (integral indefinida o primitiva) o bien obtener el área comprendida bajo una curva (integral definida).

Integral indefinida o primitiva

Sea $f : [a,b] \to \mathbb{R}$, se dice que F es una primitiva de f si $F'(x) = f(x),\ a \leq x \leq b$. Si F_1 y F_2 son dos primitivas de f, $F_1 = F_2 + \mathbb{C}$. Esto implica que una función tiene infinitas primitivas que difieren únicamente por una constante.

Se define entonces la integral indefinida cómo:

$$\int f(x)\,dx = F(x) + k$$

De las definiciones anteriores se puede considerar a la integral como la operación inversa de la derivada o *antiderivada*. A continuación se presenta una tabla con las primitivas más comunes, donde se puede hacer la comparación con la tabla de derivadas previa para confirmar la idea de «antiderivada».

$f(x)$	$F(x)$
k	$kx + \mathbb{C}$
kx^n	$\frac{kx^{n+1}}{n+1} + \mathbb{C}$
kx^{-1}	$kln\|x\| + \mathbb{C}$
e^x	$e^x + \mathbb{C}$
a^x	$\frac{a^x}{ln(a)} + \mathbb{C}$
$sen(x)$	$-cos(x) + \mathbb{C}$
$cos(x)$	$sen(x) + \mathbb{C}$
$tan(x)$	$-ln\|cos(x)\| + \mathbb{C}$

Integral definida

Cuando se quiere hallar el área algebraica que hay debajo de una curva se trabaja con la integral definida, que se expresa de la siguiente forma: $\int_a^b f(x)\,dx$, donde $[a,b]$ es el intervalo de integración. Se define como área algebraica el número real positivo o negativo obtenido contando con un signo positivo las áreas por arriba del eje x, mientras que son negativas las áreas por debajo de este.

Para hallar el área debajo de la curva f se usa el método de Riemann que consiste en aproximar la región de la curva por regiones rectangulares más pequeñas. El área total se calcula sumando el área de cada uno de estos rectángulos. Se calcula el área por defecto (la altura del rectángulo está determinada por el mínimo entre $f(a)$ y $f(b)$) y por exceso (la altura del rectángulo está determinada por el máximo entre $f(a)$ y $f(b)$). Tomando particiones cada vez más angostas, las áreas por exceso y defecto se asemejan cada vez más

al área real de la curva y se dice que una función es integrable si estas áreas coinciden. Finalmente, se puede definir el área de la función como el límite de la suma de las áreas de las aproximaciones con el ancho de los intervalos teniendo a cero.

Relación entre primitiva e integral definida

El segundo teorema fundamental del cálculo dice:

Sea f continua en un intervalo abierto I, F una primitiva cualquiera y $a,\ b$ dos puntos cualesquiera dentro de I. Para todo a y b, $F(b) = F(a) + \int_a^b f(x)\,dx$

Despejando la integral de la igualdad anterior, se puede calcular su valor como una simple substracción entre la primitiva evaluada en dos puntos de interés, por lo que el problema de integración pasa a ser cómo hallar la primitiva de la función objetivo.

Ejemplo:

La tasa de consumo de alimento de una vaca en un día está dada por una función $f(t)$, donde t representa el tiempo en días. Queremos determinar la cantidad total de alimento consumido por todas las vacas en el rebaño durante un mes, que es equivalente a calcular el área bajo la curva de la función $f(t)$ durante ese período de tiempo y luego multiplicarlo por la cantidad de vacas.

Para eso primero se calcula la primitiva de $f(t)$, obteniendo la función $F(t)$ para cualquier periodo de tiempo. Si consideramos un mes de treinta días, calculamos luego la integral definida en el intervalo [0,30] para así obtener la función deseada. Para eso, utilizamos la expresión que vincula la integral definida con su primitiva, evaluando $F(t)$ en $t = 0$ y $t = 30$.

Métodos de integración

A continuación se realiza una breve descripción de distintos métodos para calcular integrales.

Integración por sustitución

Sea g una función derivable y F una primitiva de f
Si $u = g(x) \Rightarrow \int f(g(x))g'(x)\,dx = \int f(u)\,du = F(u) + C = F(g(x)) + C$

Esto implica que, si la función a integrar es una función compuesta y está multiplicada por su derivada, se puede hacer un cambio de variable sencillo para simplificar el cálculo de la primitiva. Si se tiene una integral definida en el intervalo (a,b) los extremos de integración pasan a ser $(g(a), g(b))$.

Ejemplo:

Se quiere calcular la integral:

$$\int_0^{\sqrt{\pi}} 2xcos(x^2)\,dx$$

Se observa que tomando $g(x) = x^2$, que tiene como derivada $3x^2$ podemos expresar la integral para aplicar el método de sustitución. Realizando el cambio de variable $u = g(x)$, $g'(x)dx = du$ y cambiando los límites de integración a $(0, \pi)$ (que resultan de calcular $g(2)$ y $g(4)$), se obtiene:

$$\int_0^{\pi} cos(u)\, du = sen(\pi) - sen(0) = 0$$

Integración por partes

De la expresión de la derivada del producto, se puede deducir otra forma de calcular una integral que se denomina *Integración por partes*. Si se quiere calcular la integral de una función $k(x)$, se buscarán dos funciones f y g tales que k se pueda expresar como el producto de f y g'. La integral entonces de $f \times g'$ se puede calcular a partir de la siguiente expresión:

$$\int f(x) \times g'(x)\, dx = f(x)g(x) - \int f'(x)g(x)\, dx + \mathbb{C}$$

En el caso de una integral definida, la expresión a calcular es

$$\int_a^b f(x) \times g'(x)\, dx = f(x)g(x)|_a^b - \int_a^b f'(x)g(x)\, dx + \mathbb{C}$$

Ejemplo:

Calcular la primitiva de $\int xsen(x)\, dx$.

En primera instancia debe elegirse cuál será la función f y cual será g'. Es importante aclarar que la elección de estas funciones es a conveniencia del lector, considerando cuál de las opciones tiene primitiva más simple de hallar. En este caso, es simple calcular la primitiva de cualquiera de las dos funciones, por lo que la elección es indistinta.
Se elige $f(x) = x$ y $g'(x) = sen(x)$

Se procede a calcular $f'(x) = 1$ y $g(x) = -cos(x)$ (se deduce de la tabla), para luego aplicar la función de integración por partes:

$$\int xsen(x)\, dx = -xcos(x) - \int -cos(x) + \mathbb{C} = -xcos(x) + sen(x) + \mathbb{C}$$

donde en el último paso se calculó la integral del $cos(x)$ a partir de la tabla de primitivas típicas.

Ecuaciones diferenciales

Una *ecuación diferencial* es una igualdad en la que la incógnita es una función desconocida y (derivable hasta orden k) y en la que aparece alguna de las derivadas de la función desconocida. El orden de la ecuación queda determinado por el orden de la máxima derivada que aparece. Por ejemplo, es de orden uno si la única derivada que aparece es la derivada primera, de orden dos si aparece la derivada segunda (independientemente de si está la derivada primera o no). Son ecuaciones utilizadas para modelar muchos problemas físicos del mundo real, por ejemplo, las oscilaciones de un péndulo.

Una *ecuación diferencial ordinaria* tiene como función incógnita una función que depende únicamente de una variable, por lo que $y = f(x)$. Por otra parte, las *ecuaciones diferenciales en derivadas parciales* tienen como función incógnita una función que depende de más de una variable, por ejemplo, $y = f(x,y)$.

Ejemplos: A continuación, se muestran algunos ejemplos de ecuaciones diferenciales que modelan problemas físicos. En cada caso se detalla si la ecuación es ordinaria o parcial y el orden.

Problema	Ecuación	Tipo	Orden
Movimiento circular uniforme	$\frac{d\theta(t)}{dt} = \omega$	Ordinaria	1
Movimiento armónico simple	$y''+k^2y= 0$	Ordinaria	2
Ecuación del calor en una barra infinita	$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}$	Parcial	2
Ecuación de la cuerda	$\frac{\partial^2 u}{\partial^2 t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}$	Parcial	2

Ejemplo de ecuación diferencial de primer orden con solución

Se tiene una población de vacas y se quiere modelar en crecimiento de la misma. Para eso, se utiliza el modelo Malthus desarrollado en 1798 que relaciona el crecimiento de la población con su mortalidad y natalidad, de acuerdo con la siguiente ecuación: $\frac{dP(t)}{dt} = mP(t) - nP(t)$. $P(t)$ es el número de individuos mientras que m y n la natalidad y mortalidad respectivamente. Reescribiendo y denominando $m - n = r$, se obtiene $\frac{dP(t)}{dt} = rP(t)$

Para hallar la solución a la ecuación, se llevan a un lado todos los términos dependientes del tiempo, obteniendo así que $\frac{dP(t)}{dt}\frac{1}{P(t)} = r$. Integrando a ambos lados de la ecuación, se llega a que $ln(P(t)) + c = rt$ y aplicando la ecuación inversa al logaritmo a ambos lados de la ecuación se obtiene:

$$P(t) = ce^{rt}$$

.

Ejemplo de ecuación diferencial de segundo orden con solución

Se tiene una ecuación diferencial de segundo orden que modela el crecimiento de una población ganadera, sujeta a un factor de amortiguación (como, por ejemplo, la competencia por recursos naturales). Sea P la población y α y β la tasa de amortiguación y crecimiento de la población, respectivamente. Se tiene entonces la siguiente ecuación:

$$\frac{dP^2}{dt^2} + 2\alpha\frac{dP}{dt} + \beta P = 0$$

Para una ecuación diferencial de segundo orden, basta con hallar dos soluciones y_1 y y_2, cuya combinación lineal va a ser la solución inicial. Conviene siempre comenzar analizando si es posible hallar dos soluciones exponenciales de la forma $y_1 = \lambda_1 e^{\lambda_1 t}$ y $y_2 = \lambda_2 e^{\lambda_2 t}$. Realizando el cambio $y'' = \lambda^2$ y $y' = \lambda$ en la ecuación inicial, bastará con hallar el λ que satisface $\lambda^2 + 2\alpha\lambda + \beta = 0$

Por lo tanto, $\lambda = \frac{-2\alpha \pm \sqrt{(2\alpha)^2 - 4\beta}}{2}$ por lo que la solución de la ecuación diferencial será

$$P(t) = C_1 \frac{-2\alpha + \sqrt{(2\alpha)^2 - 4\beta}}{2} e^{\frac{-2\alpha + \sqrt{(2\alpha)^2 - 4\beta}}{2}} + C_2 \frac{-2\alpha - \sqrt{(2\alpha)^2 - 4\beta}}{2} e^{\frac{-2\alpha - \sqrt{(2\alpha)^2 - 4\beta}}{2}}$$

Apéndice C:
Álgebra lineal y geometría analítica

Este apéndice se inspiró en material generado por el IMERL (Facultad de Ingeniería), el cual se encuentra disponible para su libre consulta [ver IMERL (n.d.a) e IMERL (n.d.b)].

Matrices

Se define una matriz A de m filas y n columnas ($m \times n$) como una función A : 1,...,$m \times$ 1,...,$n \to \mathbb{R}$ con entradas a_{ij} donde la i indica a qué fila pertenece y la j la columna. Se representa de la siguiente forma:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Ejemplo:

$$A = \begin{pmatrix} 1 & 1 & 2 & 1 \\ 4 & 3 & -1 & 1 \\ -1 & 1 & 3 & 3 \end{pmatrix}$$

Matriz de 3×4, tres filas y cuatro columnas.

Relación con los sistemas de ecuaciones

Sea el sistema de ecuaciones

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m \end{cases}$$

se le llama *matriz del sistema de ecuaciones* a la matriz A de dimensión (m,n) que contiene los coeficientes del mismo:

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix}$$

y *matriz ampliada del sistema* a la matriz de dimensión $(m,n+1)$ que contiene los coeficientes y el término independiente:

$$\left(\begin{array}{cccc} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & ... & a_{mn} \end{array}\right| \left.\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}\right.$$

Ejemplo: El sistema de ecuaciones

$$\begin{cases} x + y + 2z = 9 \\ 3x + 6y - 5z = 0 \\ 2x + 4y - 3z = 1 \end{cases}$$

La matriz del sistema de ecuaciones en este caso es

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 3 & 6 & -5 \\ 2 & 4 & -3 \end{pmatrix}$$

y la matriz ampliada es

$$A = \left(\begin{array}{ccc} 1 & 1 & 2 \\ 3 & 6 & -5 \\ 2 & 4 & -3 \end{array}\right| \left.\begin{pmatrix} 9 \\ 0 \\ 1 \end{pmatrix}\right.$$

A partir de la matriz ampliada se puede aplicar un método sencillo para resolver el sistema de ecuaciones. Aplicando transformaciones elementales (intercambio, suma y resta entre filas de la matriz), se puede llevar la matriz a su formato *escalerizado* y así despejar las incógnitas del sistema.

Una matriz se encuentra escalerizada si cumple con las siguientes condiciones:
1. Todas las filas, salvo (quizás) la primera, comienzan con una sucesión de ceros.
2. Cada fila tiene al principio por lo menos un cero más que la fila inmediata superior.

Se sugiere que el lector revise a fondo la bibliografía sobre el método de escalerización, cómo realizarlo y cómo usarlo para resolver el sistema de ecuaciones.

Ejemplo: Se muestra a continuación un ejemplo de matriz escalerizada:

$$E = \left(\begin{array}{cccccc} 3 & 6 & -5 & 1 & 1 & 2 \\ 0 & -2 & -1 & -3 & 3 & 2 \\ 0 & 0 & 1 & 2 & 4 & -3 \\ 0 & 0 & 0 & -2 & -4 & 3 \\ 0 & 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & -2 \end{array}\right) \left| \begin{pmatrix} -1 \\ 1 \\ 3 \\ 1 \\ -2 \\ 1 \end{pmatrix}\right.$$

Álgebra de matrices

A continuación se definirán algunas operaciones con matrices.

Suma de matrices:

Sean A y B dos matrices de $m \times n$ con coeficientes (a_{ij}) y (b_{ij}), $i = 0,...,m$, $j = 0,...,n$. Se define la suma de matrices $(+)$ de manera que $A + B = C$ con C de coeficientes (c_{ij}) y $c_{ij} = a_{ij} + b_{ij}$. Es importante notar que para sumar dos matrices, deben ser de la misma dimensión.

Ejemplo: La suma de las matrices $A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$ y $B = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 5 & 5 \\ 0 & 4 & \frac{1}{2} \end{pmatrix}$ da como resultado la matriz

$$C = \begin{pmatrix} 1 & 4 & 4 \\ 1 & 5 & 5 \\ 0 & 4 & 1 \end{pmatrix}$$

Cada coeficiente de la matriz resultante es la suma de los coeficientes en la misma posición de las matrices sumando.

Resta de matrices:

Sean A y B dos matrices de $m \times n$ con coeficientes (a_{ij}) y (b_{ij}), $i = 0,...,m$, $j = 0,...,n$. Se define la resta de matrices $(-)$ de manera que $A - B = C$ con C de coeficientes (c_{ij}) y $c_{ij} = a_{ij} - b_{ij}$ Es importante notar que para restar dos matrices deben ser de la misma dimensión.

Ejemplo: La resta de $A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$ y $B = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 5 & 5 \\ 0 & 4 & \frac{1}{2} \end{pmatrix}$ da como resultado la matriz

$$C = \begin{pmatrix} 1 & 0 & 2 \\ 1 & -5 & -5 \\ 0 & -4 & 0 \end{pmatrix}$$

Cada coeficiente de la matriz resultante es la resta de los coeficientes en la misma posición de las matrices sumando.

Producto de matriz y escalar:

Sea $\lambda \in \mathbb{R}$ y A una matriz de dimensión $m \times n$ y coeficientes (a_{ij}). Se define el producto de λ y A tal que $\lambda A = B$, donde cada coeficiente $b_{ij} = \lambda a_{ij}$

Ejemplo: Sean $A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$ y $\lambda = 3$, la matriz B que resulta del producto λA es

$$B = \begin{pmatrix} 3 & 6 & 9 \\ 3 & 0 & 0 \\ 0 & 0 & \frac{3}{2} \end{pmatrix}$$

Producto de matrices:

Sea A una matriz de $m \times p$ y B de dimensión $p \times n$, se define el producto entre ellas como $A \times B = C$ de dimensión $m \times n$, con los coeficientes $c_{ij} = \sum_{h=1}^{p} a_{ih} b_{hj}$. A diferencia de la suma y la resta, las matrices a multiplicar no necesariamente deben tener iguales dimensiones. Lo que deben cumplir es que sean *matrices conformables*, la cantidad de columnas de la primera debe ser igual a la cantidad de filas de la segunda.

El producto entre matrices no es conmutativo. Para matrices conformables de distintas dimensiones, si se da vuelta el orden de las matrices a multiplicar, la operación queda mal definida. En caso de poder realizarse el producto con factores conmutados, el resultado es distinto.

Ejemplo: Se muestra el producto entre matrices con un ejemplo, que ayudará a entender mejor las operaciones a realizar.

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 3 & -1 & 2 \\ 1 & -1 & 1 \\ -2 & -3 & -1 \end{pmatrix} \text{ y } B = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ -1 & 1 \end{pmatrix}$$

Para hallar el coeficiente de la fila y columna uno, se toma la primera fila de A y la primera columna de B. Se multiplican posición a posición y se suman los resultados. La primera fila de A es $\begin{pmatrix} 1 & 2 & 1 \end{pmatrix}$ mientras que la columna de B es $\begin{pmatrix} 1 & 2 & -1 \end{pmatrix}$. El producto entre ellos posición a posición es $\begin{pmatrix} 1 & 4 & -1 \end{pmatrix}$. La suma de estos tres valores es 4, por lo que $c_{11} = 4$. Para c_{21} se toma la segunda fila de A y la primera columna de B y se procede de la misma forma.

Repitiendo el procedimiento para todos los c_{ij}, se obtiene una matriz de dimensión 4×2,

$$C = \begin{pmatrix} 4 & 6 \\ -1 & 4 \\ -2 & 1 \\ -7 & -6 \end{pmatrix}$$

Trasposición:

Sea la matriz A de dimensiones $m \times n$ con coeficientes (a_{ij}), se define la matriz traspuesta A^t como $A^t = ((a^t_{ij})) = ((a_{ji}))$

Ejemplo: Sea

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 3 & -1 & 2 \\ 1 & -1 & 1 \\ -2 & -3 & -1 \end{pmatrix}$$

su matriz traspuesta queda:

$$A^t = \begin{pmatrix} 1 & 3 & 1 & -2 \\ 2 & -1 & -1 & -3 \\ 1 & 2 & 1 & -1 \end{pmatrix}$$

Propiedades de la matriz traspuesta: 1. $(A+B)^t = A^t + B^t$ 2. $(\lambda^t A)^t = \lambda A^t$ 3. $(A \times B)^t = B^t \times A^t$ 4. $(A^t)^t = A$

Inversión:

Para definir la inversión de matrices es importante previamente definir la *matriz identidad*. Se le llama *identidad* de tamaño n a la matriz de dimensiones $n \times n$ tal que

$$\mathbb{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Tiene unos únicamente en la diagonal, mientras que el resto de los coeficientes son cero. Esta matriz cumple la propiedad de neutro del producto matricial, por lo que $A \times \mathbb{I}_n = \mathbb{I}_n \times A = A$.

Se define, ahora sí, una *matriz invertible*. Se dice que una matriz A de $n \times n$ es invertible si y solo si existe una matriz B de $n \times n$ tal que $A \times B = \mathbb{I}_n = B \times A$

Las matrices invertibles son sí o sí matrices cuadradas. La matriz B que multiplicada por A devuelve la identidad es una matriz única y se le denomina *matriz inversa de A*, cumpliendo que $B = A^{-1}$

Propiedad de la matriz inversa: Sean A y B dos matrices cuadradas de dimensión n:

$$(A \times B)^{-1} = B^{-1} \times A^{-1}$$

Determinante de una matriz

El *determinante* de una matriz cuadrada es un número, real o complejo, que se calcula a partir de cada matriz. Se trabajará con la definición inductiva del determinante, donde para una matriz de tamaño $n \times n$ se define a partir del determinante de una matriz de $(n-1) \times (n-1)$.

El determinante de A, $|A|$, de una matriz de 1×1 es el propio número. Para una matriz de 2×2, $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ el determinante es el número $|A| = ad - bc$. Para una matriz de $n \times n$, el determinante se define como

$$\Rightarrow |A| = (-1)^{1+1} a_{11} |A_{11}| + \ldots + (-1)^{i+1} a_{i1} |A_{i1}| + \ldots + (-1)^{n+1} a_{n1} |A_{n1}|$$

Ejemplos:

1. Calcular el determinante de la siguiente matriz: $A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ -1 & 4 & -1 \end{pmatrix}$

$$\Rightarrow |A| = (-1)^2(1) \left| \begin{pmatrix} 1 & 1 \\ 4 & -1 \end{pmatrix} \right| + (-1)^3(0) \left| \begin{pmatrix} 1 & 2 \\ 4 & -1 \end{pmatrix} \right| + (-1)^4(-1) \left| \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \right|$$
$$= -5 + 0 + (-1)(-1) = -4$$

2. Calcular el determinante de $B = \begin{pmatrix} 1 & 1 & 2 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 0 & 4 & 3 \\ 3 & 3 & 3 & 0 \end{pmatrix}$

$$\Rightarrow |B| = (-1)^2(1) \left| \begin{pmatrix} 0 & 2 & 1 \\ 0 & 4 & 3 \\ 3 & 3 & 0 \end{pmatrix} \right| + (-1)^3(1) \left| \begin{pmatrix} 1 & 2 & 1 \\ 0 & 4 & 3 \\ 3 & 3 & 0 \end{pmatrix} \right|$$
$$+ (-1)^4(0) \left| \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \\ 3 & 3 & 0 \end{pmatrix} \right| + (-1)^5(3) \left| \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & 4 & 3 \end{pmatrix} \right|$$
$$= 6 + 3 - 6 = 3$$

Los determinantes de las matrices 3×3 se calcularon como en el ejemplo (1).

Si bien se definió el determinante utilizando los elementos de la primera columna y sus matrices adjuntas, por propiedades del determinante puede utilizarse cualquier fila o columna que el lector crea conveniente.

El determinante tiene gran importancia en la inversión de matrices. En primera instancia, una matriz cuadrada que tenga determinante cero no es invertible. Además, si se tiene una matriz de 2×2 $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, la matriz inversa se puede calcular como

$$A^{-1} = \frac{1}{|A|} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Norma, producto escalar y producto vectorial

Vectores

Un vector $\vec{x}$ es un elemento geométrico que consiste de una flecha que parte desde el origen y va hasta un punto de dimensión n determinado por las coordenadas x_i. Los vectores son elementos con módulo (determinado por el largo), dirección y sentido (hacia donde va la flecha), siendo estos dos representables por el ángulo del vector.

El ángulo también se puede utilizar para medir cuánto difieren dos vectores entre sí. Si el ángulo entre ellos es cero, son colineales y apuntan en el mismo sentido. Si es de 180ř, son colineales pero tienen sentidos opuestos. Si es de 90ř o 270ř se dice que son ortogonales (figura 15.11).

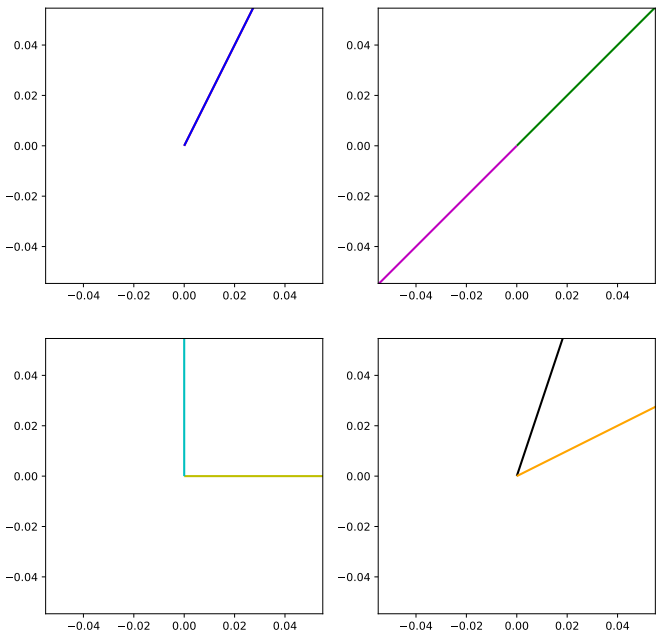


Figura 15.11. Gráfica de diferentes configuraciones de vectores en un plano cartesiano. Se muestran ejemplos de vectores colineales con igual (figura A) y opuesto (figura B) sentido, vectores ortogonales (figura C) y vectores con un ángulo arbitrario (α) entre ellos (figura D).

Distancia y norma

Se define la *distancia entre dos puntos* como la longitud del segmento entre ellos, una función $d : E \times E \to \mathbb{R}^+$, que satisface:

1. $d(A,B) = d(B,A)\ \forall A \in E$

2. $d(A,B) = 0 \Leftrightarrow A = B$

3. $d(A,C) \leq d(A,B) + d(B,C)\ \forall A,B,C \in E$

A partir de la distancia, se define la *norma de un vector* $v = \vec{AB}$ como la distancia $d(A,B)$ y se denota como $||v||$. La norma cumple que

1. $||v|| \geq 0; ||v|| = 0 \Leftrightarrow v = \vec{0}$

2. $||av|| = |a|||v||\ \forall a \in \mathbb{R},\ \forall v \in V$

3. $||u + v|| \leq ||u|| + ||v||\ \forall u,v \in V$

En particular, siendo $A = (x_a, y_a)$ y $B = (x_b, y_b)$, se define la norma dos como

$||v||_2 = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$

Producto escalar

Dados los vectores u y v de V, se llama *producto escalar* o *producto interno* de u por v al número escalar $||u||\,||v||cos\theta$, donde θ es el ángulo BAC de tres puntos tales que $\vec{AB} = u$ y $\vec{AC} = v$

Propiedades:

1. $u \times v = v \times u$

2. $(a \times u) \times v = a \times (u \times v) = u \times (a \times v)$

3. $(u_1 + u_2) \times v = u_1 \times v + u_2 \times v$

4. $v\dot{v} \geq 0 y v \times v = 0 \Leftrightarrow v = 0$

Ortogonalidad y producto escalar:

Dos vectores u, v son ortogonales si $u \times v = 0$

Producto vectorial

Se define el *producto vectorial* como una función $V \times V \to V$ que a cada par de vectores v, u le asocia un vector, que se denota $v \wedge u$, que cumple:

1. $||v \wedge u|| = ||v|| \times ||u||sen\theta$, donde θ es el ángulo entre v y u

2. $(v \wedge u) \times v = 0$ y $(v \wedge u) \times u = 0$

Además, $v \wedge u = 0$ si y solo si v y u son colineales (en particular, si alguno de ellos es el vector $\vec{0}$).

Transformaciones lineales

Sean V y W dos espacios vectoriales sobre el mismo cuerpo K. Se define como una *transformación lineal* a una función $T : V \to W$ que cumple:

1. $T(u + v) = T(u) + T(v)\ \forall u, v \in V$

2. $T(a \times v) = aT(v)\ \forall a \in K, \forall v \in V$

Ejemplo: Sea $C^1 = \{f : \mathbb{R} \to \mathbb{R} / f \text{ es derivable}\}$ y $F = \{f : \mathbb{R} \to \mathbb{R}\}$

La transformación $T : C^1 \to F$ tal que $T(f) = f'$ es una transformación lineal ya que cumple:

1. $T(f_1 + f_2) = (f_1 + f_2)' = f_1' + f_2' = T(f_1) + T(f_1)$
2. $T(\alpha f) = (\alpha f)' = \alpha f' = \alpha T(f)$

Núcleo e imagen de una transformación lineal

Sean V y W espacios vectoriales sobre un mismo cuerpo K y $T : V \to W$ una transformación lineal. Se define el *núcleo de la transformación lineal* T al conjunto de V cuya imagen por T es el vector nulo $\Rightarrow N(T) = \{v \in V :\ T(v) = \vec{0}\}$ La imagen de T es el conjunto $Im(T) = \{w \in W :\ \exists v \in V \text{ con } w = T(v)\}$

Geometría de matrices

Se definen un vector $\vec{x}$ y otro vector $\vec{y}$ tal que $y = Ax$. El vector $\vec{y}$ tendrá coordenadas distintas a las del vector $\vec{x}$ a menos que A sea una matriz cuadrada. Por lo tanto, se dice que A describe una transformación (en particular rotación y escalado) del sistema de coordenadas de x a un nuevo sistema de coordenadas y.

Para definir las componentes de rotación y escalado de una matriz, conviene comenzar definiendo una matriz ortonormal. Una matriz U es ortonormal si:

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1\ if\ i = j \\ 0\ if\ i \neq j \end{cases}$$

También se denominan *unitarias* y cumplen que $\mathbf{U}^T\mathbf{U} = \mathbf{I}$

Si se multiplica un vector por estas matrices, no hay escalado si no únicamente rotación dado que las matrices ortonormales preservan los productos internos. Las matrices de rotación son de la forma

$$\begin{pmatrix} cos\phi & -sen\phi \\ sen\phi & cos\phi \end{pmatrix}$$

donde ϕ determina el ángulo de rotación.

La transformación de escalado quedará descrita por los valores y vectores propios a tratar en la siguiente sección.

Valores y vectores propios

De una transformación lineal:

Sea V un espacio vectorial sobre el conjunto de escalares $\mathbb{K}$ y $T : V \to V$ una transformación lineal. Se llama *vector propio* de T, asociado al *valor propio* $\lambda \in \mathbb{K}$ a todo vector $\vec{v} \neq 0$ tal que $T(v) = \lambda v$.

Los valores propios describen como será el escalado de los vectores en su nuevo espacio vectorial que queda determinado por los vectores propios.

De una matriz:

Sea A una matriz cuadrada con entradas en $\mathbb{K}$. Se llama *vector propio* de A asociado al *valor propio* $\lambda \in \mathbb{K}$ a todo vector no nulo de $\mathbb{K}^n$ tal que

$$A \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Proposición: Sean V un espacio vectorial sobre el cuerpo $\mathbb{K}$, $T : V \to V$ una transformación lineal y A una matriz asociada a la transformación lineal T. Entonces λ es valor propio de T si y solo si $\lambda \in \mathbb{K}$ y $\det(A - \lambda I) = 0$

Se le llama *polinomio característico* de una matriz cuadrada al polinomio representado por $x_A\lambda$ que resulta de calcular $\det(A - \lambda I)$. La *ecuación característica* es $x_A\lambda = 0$ y a todas las soluciones de ella se les llama *raíces características*. Estas soluciones, como se vio previamente, son los valores propios.

Ejemplo: Hallar los valores propios de la matriz A asociada a la transformación lineal T:

$$A = \begin{pmatrix} 2 & -1 & 0 \\ 3 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Se calcula el determinante de $A - \lambda I$:

$$\det(A - \lambda I) = \left| \begin{pmatrix} 2-\lambda & -1 & 0 \\ 3 & 2-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix} \right| = (1-\lambda)(\lambda^2 - 4\lambda + 7) = 0$$

Hallando las raíces de la ecuación previa se obtiene:

$$\Rightarrow \begin{pmatrix} \lambda_1 = 1 \\ \lambda_2 = 2 + i\sqrt{3} \\ \lambda_3 = 2 - i\sqrt{3} \end{pmatrix}$$

Para hallar los vectores propios, se sustituye cada uno de los valores propios en la matriz $A - \lambda I$ para resolver el sistema

$$(A - \lambda I) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Será vector propio el vector $v = (x_1, x_2,...,x_n)$ con $x_1, x_2,...,x_n$ soluciones del sistema.

Diagonalización de una matriz

Se le denomina *matriz diagonal* a aquellas matrices que todos sus términos por fuera de la diagonal principal son nulos. Se busca llevar una matriz a su forma diagonal, para lo que se calcularán sus valores propios. Si A tiene n valores propios reales, se puede definir una matriz D diagonal, tal que $A = P^{-1}DP$ donde P es una matriz invertible compuesta por los vectores propios en columnas. Además, en ese caso, $D = \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \lambda_n \end{pmatrix}$

La matriz D es única a menos del orden de $\lambda_1, \lambda_2, ..., \lambda_n$. A la matriz P la llamamos *base* de la transformación a diferencia de D no es única.

Derivada de la forma cuadrática

Queremos estudiar la derivada de la forma cuadrática: $f(x) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ donde $\mathbf{x}$ es un vector en $\mathbf{R^m}$ y A una matriz de dimensión $m \times m$. Reescribiendo de forma tal que se nombra a la transformada $\mathbf{A}\mathbf{x}$ como una función $y(x)$ se llega a que $f(x,y(x)) = \mathbf{x}^T y(x)$

Se puede derivar $f(x)$ usando la fórmula de derivada total, obteniendo que $f'(x,y(x)) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} y'(x)$.

Para llegar a la expresión final, se realizará el análisis de cada sumando de manera independiente.

En primera instancia, se calcula la derivada $\frac{\partial f}{\partial x} = \frac{\partial x^T y(x)}{\partial x}$. Tomando $y(x)$ fijo, la derivada de la expresión es la derivada de un producto interno, por lo que resulta en $\frac{\partial x^T y(x)}{\partial x} = y(x)^T = \mathbf{x}^T\mathbf{A}^T$

El segundo sumando tiene por un lado la derivada de la función f según y, que siguiendo el procedimiento del previamente explicado es $\mathbf{x}^T$. Por otro lado, $y'(x)$ es la derivada del producto de una matriz y un vector, que resulta en la matriz. Por lo tanto, el segundo sumando nos queda $\frac{\partial f}{\partial y}.y'(x) = \mathbf{x}^T\mathbf{A}$

Sustituyendo y operando, se obtiene $f'(x,y(x)) = 2\mathbf{x}^T\mathbf{A}^T$

Apéndice D: Probabilidad y estadística

Este apéndice se inspiró en material generado por el IMERL (Facultad de Ingeniería), el cual se encuentra disponible para su libre consulta (ver IMERL (n.d.d)).

Probabilidad y conteo

Intuitivamente, se define la probabilidad como la chance de que un evento ocurra frente a todo el universo de posibilidades. Si se tiene un espacio muestral, definido como el conjunto de todos los resultados posibles que pueden ocurrir con igual probabilidad, y un evento A, se define la probabilidad del evento A como $P(A) = \frac{casos\ favorables}{casos\ posibles}$

Las probabilidades cumplen tres reglas: 1. Son siempre positivas ($P(A) \geq 0$) para cualquier evento. 2. $P(\Omega) = 1$ donde Ω es el espacio muestral. 3. Dados k eventos incompatibles (que no pueden ocurrir en simultáneo), $P(A \cup B \cup ... \cup K) = P(A) + P(B) + ... + P(K)$

Sin embargo, contar los casos favorables no siempre es una tarea sencilla, por lo que se utilizan elementos de la teoría combinatoria para facilitar el conteo.

Número factorial: El número n factorial, denotado por $n!$, se define como el producto de todos los números entre 1 y n: $\Rightarrow n! = n \times (n-1) \times (n-2) \times ... \times (n-n+1)$

Arreglos: Los arreglos definen el número total de formas en que se pueden elegir k elementos ordenados de un total de n elementos. Para elegir el primer elemento se tienen n opciones. Para el segundo elemento ya se descartó una opción, por lo que se pasa a tener $n-1$ posibilidades. Así sucesivamente hasta llegar al elemento k, para el cual se tienen $n-k+1$ opciones. El número total de posibilidades se define entonces como

$$(n)_k = (A)_k^n = n \times (n-1) \times (n-2) \times ... \times (n-k+1) = \frac{n!}{(n-k)!}.$$

Combinaciones: Las combinaciones definen el número total de formas en que se pueden elegir k elementos no ordenados de un total de n elementos. Si se tiene una lista no ordenada de k elementos, se pueden formar $k!$ listas ordenadas. Esto quiere decir que por cada lista no ordenada hay $k!$ listas ordenadas, lo que implica que las combinaciones no son más que arreglos divididos por $k!$. Por tanto:

$$\binom{n}{k} = C_k^n = \frac{(A)_k^n}{k!} = \frac{n!}{k!(n-k)!}.$$

Probabilidad condicional

La *probabilidad condicional* define la probabilidad de que ocurra un evento cuando se sabe qué ocurrió previo a él. Formalmente, se tienen dos eventos A y B del espacio muestral Ω y se conoce que $P(B) > 0$. Se define la probabilidad de A dado B como $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Ejemplo:
Se tiran dos datos y se quiere saber la probabilidad de que la suma sea siete (a este evento lo llamaremos A) dado que uno de ellos muestra un uno (evento B). Por la definición de probabilidad condicional, queremos calcular $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Comenzamos calculando la probabilidad de B. De los 36 resultados posibles (6^2), 11 de ellos tienen un uno en al menos uno de los dados $\Rightarrow P(B) = \frac{11}{36}$

Para la probabilidad de la intersección, hay que tener en cuenta que hay 6 combinaciones posibles con las cuales la suma de los dos dados da 7. Sin embargo, solo dos de ellas contienen un uno en alguno de los dos dados, por lo tanto: $\Rightarrow P(A \cap B) = \frac{2}{36}$

Juntando los dos resultados, obtenemos que $P(A|B) = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$

Fórmula de probabilidad total

Sean $C_1, C_2, ..., C_n$ una partición numerable de Ω cuyos eventos tienen probabilidades positivas. Sea A un evento cualquiera. Entonces $P(A) = \sum_{i=1}^{n} P(A|C_i)P(C_i)$

Teorema de Bayes

Sean $C_1, C_2, ..., C_n$ una partición numerable de Ω cuyos eventos tienen probabilidades positivas. Sea A un evento con probabilidad positiva, entonces $P(C_k|A) = \frac{P(C_k)P(A|C_k)}{\sum_{i=1}^{n} P(A|C_i)P(C_i)}$

Variables aleatorias

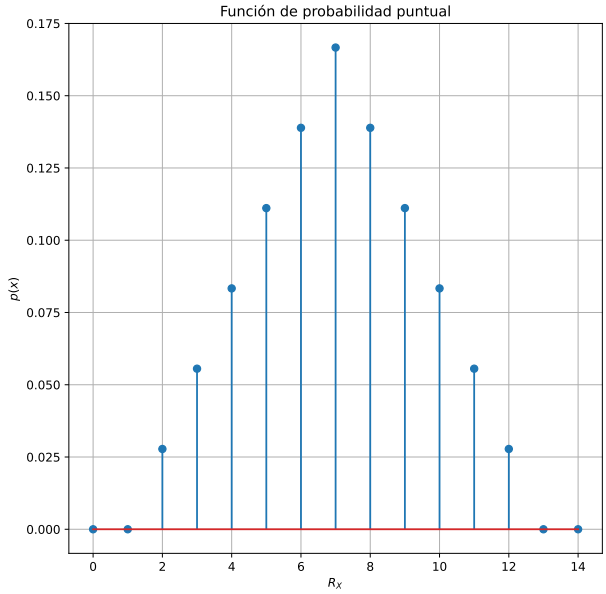
Una variable aleatoria es una función $X : \Omega \to \mathbb{R}$ que a cada elemento ω del espacio muestral le asigna un número real $X(\omega)$. Se definen variables aleatorias discretas o continuas en función de cómo está conformado el espacio muestral. Una variable aleatoria es discreta si el recorrido, los valores que la variable puede tomar, es numerable. Es decir, se pueden ordenar en una sucesión. En cambio, es continua si sus valores consisten en uno o más intervalos de la recta de los reales.

Variables aleatorias discretas

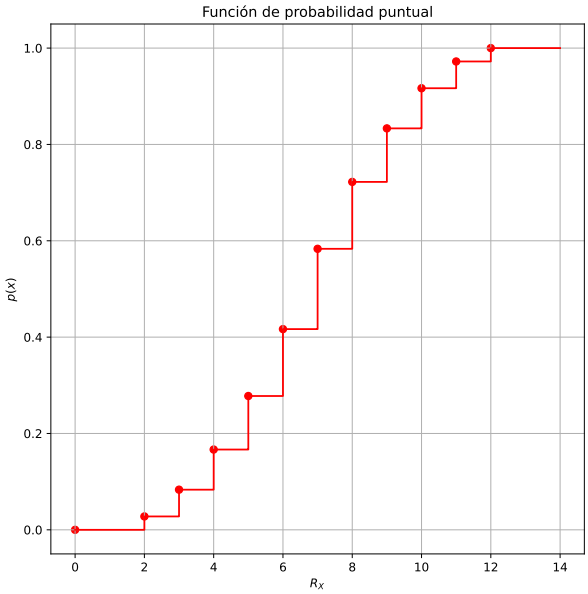
Para las variables aleatorias discretas podemos definir dos formas de representación:

Función de probabilidad puntual: La función de probabilidad puntual de una variable aleatoria discreta $X : \Omega \to X$ es una función $p : \mathbb{R} \to [0,1]$, definida por $p(x) = P(X = x)$, que indica para cada punto de x de $\mathbb{R}$ la probabilidad de que X tome ese valor. Si x no es un posible valor de X, $p(x) = 1$. Además, $\sum_{x \in \mathbb{R}_X} p(x) = 1$ debido a que en R_X están todos los posibles valores de X.

Ejemplo: Función de probabilidad puntual de la suma del resultado de lanzar dos dados.



Función de distribución acumulada: La función de distribución acumulada de una variable aleatoria X es una función $F : \mathbb{R} \to [0,1]$ definida por $F(x) = P(X \leq x)$. F da la probabilidad acumulada al sumar las probabilidades $p(y)$ con $y \leq x$.



Distribución de probabilidad conjunta

La distribución de probabilidad conjunta muestra cómo se comportan las variables X e Y y las probabilidades de los distintos resultados posibles. Formalmente, es la función $p : \mathbb{R}^2 \to [0,1]$ definida por $p(x,y) = P(X = x, Y = y)$

Las distribuciones de X e Y obtenidas a partir de la distribución conjunta se llaman *distribuciones marginales*.

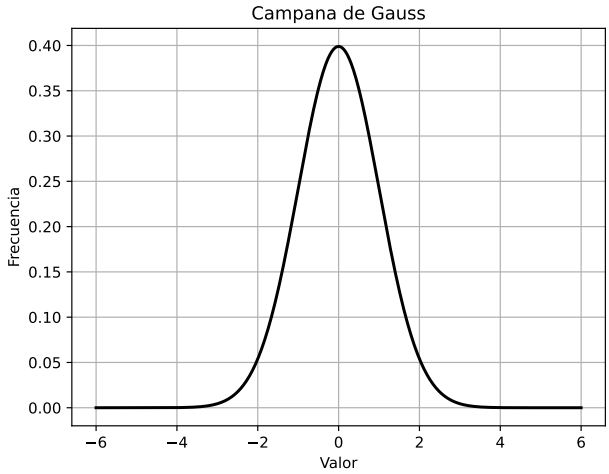
Variables aleatorias continuas

Se trabajará específicamente con las variables aleatorias absolutamente continuas, pero serán nombradas *variables aleatorias continuas*. Para definirlas, es necesario previamente definir qué es una *densidad de probabilidad*.

Densidad de probabilidad: Una densidad de probabilidad es una función $p : \mathbb{R} \to \mathbb{R}$ integrable que cumple las siguientes dos condiciones: 1. Es positiva $p(x) \geq 0$ para todo $x \in \mathbb{R}$. 2. $\int_{-\infty}^{+\infty} p(x)dx = 1$

Una variable aleatoria es absolutamente continua si existe una densidad de probabilidad p tal que $P(X \in I) = \int_I p(x)dx$ para todo intervalo I de $\mathbb{R}$.

Un ejemplo conocido de variable aleatoria continua es la *campana de Gauss*, donde la probabilidad de que una variable aleatoria se encuentre entre dos puntos $[a,b]$ está dada por la gráfica bajo la curva y se puede calcular usando la tabla de la distribución normal (que desarrollaremos más adelante).



Densidad de probabilidad conjunta

Al igual que se definió a la distribución de probabilidad conjunta para las variables aleatorias discretas, se puede definir la densidad de probabilidad conjunta para variables continuas. La densidad de probabilidad en dos variables cumple que $p(x,y) \geq 0$ y $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x,y)dxdy = 1$

Variables independientes

Para el caso de variables aleatorias discretas, se dice que son independientes si la distribución conjunta es igual al producto de las marginales: $p(x,y) = p(x)p(y)$

La definición de variable independiente en el caso continuo es análoga, pero para la densidad conjunta y densidades marginales.

Valor esperado

Si repetimos un experimento la suficiente cantidad de veces, el valor que tomará la variable aleatoria es el *valor esperado*. A continuación se define formalmente para variables aleatorias discretas y continuas.

El valor esperado de una variable aleatoria discreta X con recorrido R_X se define como $E(X) = \sum_{x \in R_X} x \times P(X = x)$

Para variables aleatorias continuas con densidad $p(x)$ se define como $E(X) = \int_{-\infty}^{+\infty} xp(x)dx$

Propiedades importantes:

Se definen a continuación las principales propiedades del valor esperado, que aplican tanto para variables aleatorias discretas como continuas.

Propiedad 1: Sea X una variable aleatoria y a una constante cualquiera. Entonces $E(aX) = aE(x)$

Propiedad 2: Sean X e Y dos variables aleatorias. Entonces $E(X+Y) = E(X) + E(Y)$

Propiedad 3: Sean X e Y dos variables aleatorias independientes, estas cumplen que $E(XY) = E(X)E(Y)$

Varianza y desviación estándar

La *varianza* es una medida de qué tan dispersos están los valores de X respecto a su valor esperado. Para una variable aleatoria discreta X con función de probabilidad puntual $p(x)$, la varianza se define como $Var(X) = \sum_{x \in R_X} (x - E(X))^2 p(x)$

En el caso de una variable aleatoria continua X con función de densidad $p(x)$, la varianza se define como $Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 p(x) dx$

Otra notación común de la varianza es σ^2. La varianza tiene unidades de X^2 por lo que se defina otra medida de desviación con las mismas unidades de X que resulta en un valor con mejor interpretabilidad. Se le llama *desviación estándar* y se define como $\sigma = \sqrt{Var(X)}$

Propiedades importantes:

Se desarrollan propiedades de la varianza, que aplican tanto para variables aleatorias continuas como discretas.

Propiedad 1: Sea X una variable aleatoria y c una constante cualquiera. Entonces $Var(X + c) = Var(X)$

Propiedad 2: Sea X una variable aleatoria y c una constante cualquiera. Entonces $Var(cX) = c^2 Var(X)$

Propiedad 3: La varianza de una variable aleatoria X se puede calcular como $Var(X) = E(X^2) - E^2(X)$

Propiedad 4: Sean X e Y dos variables aleatorias independientes, la varianza de $X + Y$ queda definida por $Var(X + Y) = Var(X) + Var(Y)$

Covarianza y coeficiente de correlación

La *covarianza* mide la variabilidad conjunta de X e Y. Mide qué tan asociadas están estas dos variables aleatorias y sus respectivas dispersiones. Está definida como

$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

que desarrollando queda

$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

Si a la covarianza se la divide por el producto de las desviaciones estándar de X e Y se obtiene el *coeficiente de correlación*: $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Propiedades importantes:

Propiedad 1: Desigualdad de Cauchy - Schwarz Sean X e Y dos variables aleatorias cualesquiera $\Rightarrow |Cov(X,Y)| \leq \sqrt{Var(X)}\sqrt{Var(Y)} \leq 1$

Propiedad 2: Sean X e Y dos variables aleatorias cualesquiera $\Rightarrow Cov(X,Y) = Cov(Y,X)$

Propiedad 3: Sea X una variable aleatoria constante $\Leftrightarrow Cov(X,X) = 0$

Propiedad 4: Sean X e Y dos variables aleatorias independientes $\Rightarrow Cov(X,Y) = 0$ dado que $E(XY) = E(X)E(Y)$

Propiedad 5: Sean X, Y y Z tres variables aleatorias discretas y a una constante $\Rightarrow Cov(aX + Y, Z) = aCov(X,Z) + Cov(Y,Z)$

Ley de los grandes números

La ley de los grandes números establece que el promedio de muchas realizaciones independientes de un mismo experimento está, con alta probabilidad, cerca de la esperanza de la distribución subyacente. El enunciado dice:

Sean $X_1, X_2,...,X_n$ variables aleatorias independientes e idénticamente distribuidas (i.i.d), con esperanza $\mu = E(X_i)$ y varianza $\sigma^2 = Var(X_i)$. Entonces, para todo $\epsilon > 0 \Rightarrow P(|\overline{X}_n - \mu| < \epsilon) \to 1$ cuando $n \to \infty$

$\overline{X}_n$ es el promedio de las variables aleatorias i.i.d que tiene como esperanza $E(\overline{X}_n) = \mu$ y varianza $Var(\overline{X}_n) = \frac{\sigma^2}{n}$

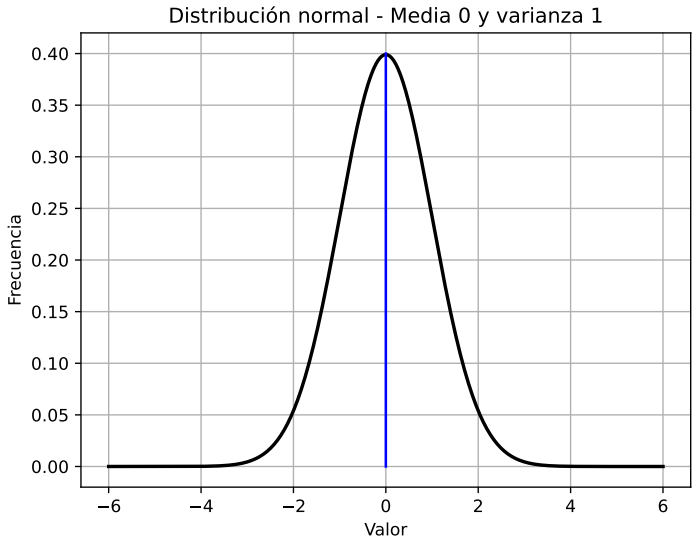
Distribuciones conocidas

A continuación se muestra una tabla donde se presentan algunas variables aleatorias típicamente utilizadas, qué experimentos modelan, su distribución, esperanza y varianza.

Variable	Caso Prototipo	$p_X(x)$	$F_X(x)$	Esperanza	Varianza
Uniforme: $X \sim U(n)$	Todos los resultados equiprobables	$\frac{1}{n}$	1	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli: $X \sim Ber(p)$	Éxito o fracaso	$(1-p)^{1-k}p^k$	$\begin{cases} 1-p & \text{si } k=0 \\ 1 & \text{si } k=1 \end{cases}$	p	$p(1-p)$
Geométrica: $X \sim Geo(p)$	Repetir ensayos de Bernoulli hasta uno exitoso	$(1-p)^{k-1}p$	$\begin{cases} 1-(1-p)^{k+1} & \text{si } k=0 \\ 1 & \text{si } k=1 \end{cases}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial: $X \sim Bin(n,p)$	k éxitos en n ensayos	$\binom{n}{k}(1-p)^{n-k}p^k$	–	np	$np(1-p)$
Poisson: $X \sim Pois(\mu)$	k ocurrencias con media μ en tiempo t	$\frac{e^{-\mu}\mu^k}{k!}$	–	μ	μ
Hipergeométrica: $X \sim H(N,q,n)$	Urna con N bolillas, q rojas. Probabilidad de k rojas sin reposición	$\frac{\binom{q}{k}\binom{N-q}{n-k}}{\binom{N}{n}}$	–	$\frac{nq}{N}$	$(\frac{nq(N-q)}{N^2}) \times \left(1-\frac{n-1}{N-1}\right)$

Distribución normal

La distribución normal es de las distribuciones más conocidas en probabilidad y estadística. Su función de densidad es la conocida como **Campana de Gauss** y está dada por la función $\varphi = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ definida para todo $x \in \mathbb{R}$. Esta distribución normal tiene esperanza 0 y varianza 1 por lo que la función de densidad es una campana de Gauss centrada en 0. Se representa por $X \sim N(0,1)$



De la gráfica de la distribución normal se observa que es simétrica respecto a la media. Para hallar la función de distribución acumulada debe integrarse, pero φ no tiene una primitiva elemental (no es expresable mediante funciones conocidas). Para calcular probabilidades usando esta distribución se utiliza la **Tabla de la distribución normal** que, para cada valor, indica el área bajo la curva que hay desde $-\infty$ hasta ese punto.

Se observa que no indica el valor de probabilidad para valores menores a 0. Sin embargo, dada la simetría mencionada previamente y que la probabilidad del complemento de un conjunto que tiene probabilidad p es $1-p$, se puede deducir que $\phi(-x) = 1 - \phi(x)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Ejemplo:

Se quiere saber la probabilidad de que $X < 2{,}38$. Para eso, en la tabla buscamos la fila correspondiente a 2,3, ya que las filas indican unidades y primer decimal. Posicionados en esa fila, buscamos la columna que corresponda a 0,08, dado que las columnas indican el segundo decimal. Se llega entonces a que $P(X < 2{,}38) = \phi(2{,}38) = 0{,}9913$. Si se quisiera calcular la probabilidad de $X > 2{,}38$ basta con calcular la probabilidad del complemento, por lo que $P(X > 2{,}38) = 1 - \phi(2{,}38) = 0{,}0087$

En el caso de querer calcular la probabilidad de que $X < -2{,}38$ se utiliza la propiedad de simetría. Se observa que la cola gaussiana a la izquierda de $-2{,}38$ es igual a la cola a la derecha de 2,38 por lo que $P(X < -2{,}38) = P(X > 2{,}38) = 0{,}0087$.

Normal de media μ y desviación estándar σ

La distribución normal puede tener media $\mu \neq 0$ y desviación estándar $\sigma \neq 1$. Que la media sea distinta de cero implica que la distribución tendrá su máximo en ese valor μ, lo que genera que esté centrada sobre otro valor del eje x. En el caso de la desviación estándar, cuanto mayor sea, más ancha queda la campana, ocurriendo lo contrario para valores más pequeños. Sin embargo, al ser una densidad de probabilidad debe siempre integrar uno, por lo que a mayor desviación estándar menor es el máximo de la campana. En este caso, la función de densidad está dada por $\varphi = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

La desviación estándar indica que en el intervalo $[\mu - \sigma, \mu + \sigma]$ se encuentra aproximadamente el **68 %** de los datos. A 2σ de distancia de μ se encuentra **94 %** de los datos, aproximadamente.

Estandarización: Si bien existen distribuciones normales $N(\mu, \sigma^2)$, la tabla de distribución normal conocida con la que se trabaja es la de la normal $N(0,1)$. Se buscará modelar la distribución que se tiene como una normal $N(0,1)$ mediante el proceso de estandarización.

Si se tiene una distribución normal de parámetros μ y σ y además $a < b$:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

Se puede ver que $\frac{X-\mu}{\sigma}$ tiene distribución $N(0,1)$ por lo que

$$P(a < X < b) = \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$$

.

Ejemplo: Sea $X \sim N(2,2)$, se quiere calcular la probabilidad de $X \in [1.5, 3]$. Estandarizando, se obtiene que los valores a buscar en la tabla son

$$\begin{cases} \frac{a-\mu}{\sigma} = \frac{1{,}5-2}{2} = -0.25 \\ \frac{b-\mu}{\sigma} = \frac{3-2}{2} = 0.5 \end{cases}$$

Recordar que al tener un valor negativo, $\phi(-x) = 1 - \phi(x)$ por lo que $\phi(-0{,}25) = 1 - \phi(0{,}25)$. Sustituyendo en $\phi(0{,}5) - \phi(-0{,}25)$, se obtiene que $P(1{,}5 < X < 3) = \phi(0{,}5) - 1 + \phi(0{,}25) = 0{,}6915 - 1 + 0{,}5987 \Rightarrow P(1{,}5 < X < 3) = 0{,}2902$.

Combinación lineal de normales independientes

Combinaciones lineales de normales independientes generan distribuciones normales. En particular, es de interés mencionar qué pasa en el caso de la suma de normales.

Si X_1,...,X_n son independientes con $X_i \sim N(\mu_i, \sigma_i^2)$ entonces

$$X_1 + ... + X_n \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

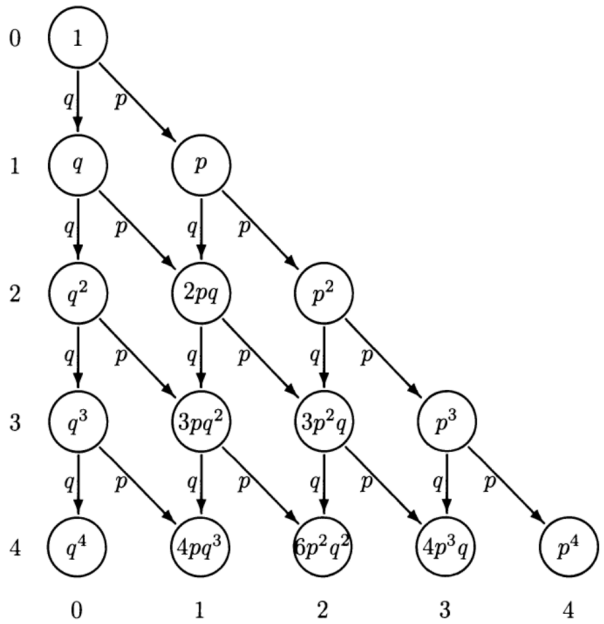
Si además las normales son i.i.d., se puede saber la distribución del promedio: Si X_1,...,X_n son i.i.id con $X_i \sim N(\mu, \sigma^2)$ entonces

$$\frac{X_1 + ... + X_n}{n} \sim N\left(\mu, \frac{\sigma_i^2}{n}\right)$$

Distribución binomial

Para definir la distribución Binomial previamente necesitamos definir un ensayo de Bernoulli. Los ensayos de Bernoulli miden la probabilidad de éxito (y por ende la de fracaso) de que ocurra un evento. Por ejemplo, tiramos una moneda y queremos medir la probabilidad de que salga *cara*. Llamaremos p a la probabilidad de éxito y q a la de no éxito, que por propiedades de probabilidad será $q = 1 - p$

La distribución binomial mide la probabilidad de obtener k éxitos en n ensayos Bernoulli independientes entre sí. A continuación se muestra un árbol de probabilidades para los primeros cuatro niveles, donde en cada nivel se muestra la probabilidad de que ocurra ese evento y se llega con probabilidad p o q si el evento es éxito o fracaso.



Si queremos la probabilidad de tener dos éxitos en cuatro eventos, se buscan en el árbol aquellos caminos en que en todo el recorrido —desde el primer evento hasta el último— haya dos éxitos.

Extendiendo el árbol a n pasos, si quiero k éxitos debo recorrer el camino donde tenga k veces el evento exitoso y $n-k$ veces el evento fracaso. Cada uno de estos caminos tendrá probabilidad p^k (correspondiente a los éxitos) multiplicado por q^{n-k} (correspondiente a los fracasos), o lo que es lo mismo $(1-p)^{n-k}$. Extendiéndolo a la cantidad de caminos posibles, la probabilidad de tener k éxitos en n ensayos está dada por

$$P(S_n = k) = (\begin{cases} n \\ k \end{cases})p^k(1-p)^{n-k}$$

Recordando que la esperanza de una Bernoulli de parámetro p es p y la varianza es $p(1-p)$, se llega a que la esperanza y varianza de una Binomial (S_n) son

$$\begin{cases} E(S_n) = np \\ Var(S_n) = np(1-p) \end{cases}$$

Distribución de Poisson

La distribución de Poisson mide la cantidad de éxitos en un intervalo de tiempo t. Se puede considerar como una extensión del proceso de Bernoulli pero que considera los éxitos en un intervalo continuo en lugar de realizaciones discretas del experimento.

Para que se ajuste a un proceso de Possion de parámetro $\mu > 0$ se deben cumplir las siguientes tres condiciones:
– La cantidad de éxitos en subintervalos disjuntos deben ser independientes.
– La probabilidad de éxito de un intervalo de lo suficientemente chico de longitud h es μh.
– La probabilidad de dos o más éxitos en un intervalo lo suficientemente chico es esencialmente cero.

Se considera un experimento que cumple las tres condiciones previas, para buscar la probabilidad P que modele el experimento se considera un intervalo de longitud 1 que se divide en n intervalos de igual longitud. Como se considera n lo suficientemente grande, se sabe que en cada intervalo va a haber un éxito con probabilidad $\frac{\mu}{n}$. El éxito o no en cada intervalo se modela como una Bernoulli, por lo que conseguir k éxitos en los n intervalos se puede modelar como una binomial. Tomando la probabilidad de la binomial y calculando su límite para $n \to \infty$ se obtiene que

Una variable discreta S tiene distribución de Poisson de parámetro μ si toma valores enteros mayores o iguales a cero con probabilidad

$$P(S = k) = \frac{\mu^k}{k!}e^{-\mu}$$

con $k = 0,1,2....$ Se escribe como $S \sim Poiss(\mu)$.

Ejemplo:
Se tiene un rebaño de ovejas del que se sabe que en promedio $\mu = 2$ de ellas dan a luz por día. Se quiere calcular la probabilidad de que cinco vacas den a luz, por lo que sustituyendo $k = 5$ y $\mu = 2$ en la función de la probabilidad de Poisson se obtiene la probabilidad deseada.

Estimadores

Estimadores frecuentistas

Un modelo paramétrico consiste en suponer que se conoce la densidad de probabilidad salvo algunos parámetros. Por ejemplo, se supone que las muestras que se tienen de un experimento se ajustan a una distribución normal, pero no se conoce su media ni su varianza. Un parámetro se denota con la letra θ y la densidad correspondiente por $p(x;\theta)$.

Un *estimador* es un estadístico que se cree tiene información suficiente sobre un parámetro θ de la distribución X. ¿Cómo evaluar si un estimador es bueno? Se definirán dos propiedades con las que se pueden establecer dos criterios.

Sesgo:

Se define el sesgo de un estimador T como la diferencia entre el parámetro θ y la esperanza del estimador: $Sesgo(\theta) = E(T) - \theta$. Se dice que un estimador es insesgado si $E(T) = \theta$.

Se dice que un estimador es *asintóticamente insesgado* si $\lim_{n\to+\infty} E(T) = \theta$, siendo n el número de muestras de la distribución.

Ejemplo: Se quiere estimar el parámetro μ de una distribución. Se elige como estimador el promedio $\overline{X}_n$. ¿Es este estimador insesgado?

Calculamos su esperanza:

$$E(\overline{X}_n) = E(\frac{1}{n}\sum_{n=1}^{n} X_i) = \frac{1}{n}\sum_{n=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu$$

Por lo tanto, es un estimador insesgado.

Error cuadrático medio:

El error cuadrático medio de un estimador T de un parámetro θ está definido como $MSE(T) = E\left((T-\theta)^2\right)$. Se puede reescribir como $MSE(T) = Var(T) + Sesgo(T)^2$

En general, los estimadores que minimizan el MSE dependen del parámetro desconocido, dado que dependen del sesgo. Por lo tanto, suelen ser no realizables. Un estimador insesgado suele tener mayor varianza, resultando en lo que se conoce como *compromiso sesgo-varianza*. Se debe decidir si buscar minimizar el sesgo (que resulta en un estimador más exacto) o la varianza (que resulta en un estimador más preciso).

Otras propiedades:

Consistencia: Se dice que un estimador es consistente si converge en probabilidad al valor del parámetro que está intentando estimar, conforme crece la muestra.

Invarianza: Un estimador se dice invariante si una función del estimador $\hat{\theta}$ es igual al estimador de la función del parámetro θ. O sea que $f(\hat{\theta}) = f\hat{(\theta)}$

Suficiencia: Un estimador de θ se considera suficiente si utiliza toda la información contenida en la muestra aleatoria con respecto a θ

Robustez: Un estimador robusto es aquel que sigue siendo efectivo incluso si las suposiciones iniciales del modelo estadístico no se cumplen completamente.

Asintóticamente normal: Un estimador se dice que es asintóticamente normal si, a medida que el tamaño de la muestra tiende a infinito, la distribución del estimador se aproxima a una distribución normal.

Métodos de estimación

A continuación, se presentan distintas técnicas de estimación. Puede convenir la utilización de unas u otras en función de qué se busca (estimadores insesgados o estimadores de varianza pequeña, por ejemplo).

Estimadores insesgados de varianza mínima (MVU)

Dado que minimizar el MSE puede llevar a estimadores no realizables, una opción es elegir el estimador de menor varianza del conjunto de estimadores insesgados. No necesariamente va a ser el estimador que minimice el MSE, pero sí es el que tiene menor MSE dado que el sesgo es cero. También se le llama *estimador eficiente*.

Mejor estimador lineal insesgado (BLUE)

Otra estrategia para elegir el mejor estimador dado el problema es restringir el estimador a ser lineal con los datos y buscar el estimador lineal insesgado con menor varianza. Un estimador es lineal con los datos si: $\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n]$. Se deberá hallar los coeficientes a_n para que el estimador sea insesgado y de mínima varianza.

Estimadores de máxima verosimilitud (MLE)

El estimador de máxima verosimilitud es una herramienta muy popular para obtener estimadores prácticos ya que tiene características asintóticas deseables: es asintóticamente eficiente y es consistente (con probabilidad alta se acerca al parámetro buscado dada una cantidad grande de muestras). Suelen utilizarse métodos numéricos para hallarlo.

Sea una muestra $x_1,...,x_n$ obtenida por un muestreo aleatorio $X_1,...,X_n$ de una variable X con distribución $p(x;\theta)$ que depende de un parámetro θ a estimar. La probabilidad de observar la muestra que observamos es $p(x_1;\theta)p(x_2;\theta)...p(x_n;\theta)$. El principio de *máxima verosimilitud* se basa en hallar un $\hat{\theta}$ estimador de θ que maximice este producto.

Definimos la *función de verosimilitud* como $L_n(\theta) = \prod_{i=1}^{n} p(X_i;\theta)$ donde $X_1,...,X_n$ son muestras i.i.d con distrbución $p(x;\theta)$. Es usual trabajar con el logaritmo ya que facilita las operaciones: $l_n(\theta) = \sum_{i=1}^{n} ln(p(X_i;\theta))$

Sea $X_1,...,X_n$ i.i.d $\sim p(x;\theta)$. El estimador T_n de θ es el valor que maximiza la función de verosimilitud (o de forma equivalente, su logaritmo): $T_n = argmax_\theta L_n(\theta) = argmax_\theta l_n(\theta)$

Estimadores bayesianos

El enfoque bayesiano toma al parámetro θ no como un valor desconocido (como es el caso del enfoque trabajado hasta el momento) sino que lo considera una variable aleatoria. Esta visión permite agregar información sobre θ como una distribución a priori $p(\theta)$. Si la información brindada por la distribución a priori es razonable, puede llevar a estimaciones más precisas. Sin embargo, la elección de esta distribución deberá ser muy cautelosa ya que información errónea puede tener el efecto opuesto. Por lo tanto, si no se tiene ningún tipo de noción previa, conviene utilizar el enfoque clásico.

Un ejemplo de información a priori es que el parámetro θ no pueda ser negativo o que sea cercano a un cierto valor a conocido.

Estimador del error cuadrático medio mínimo (MMSE)

El *error cuadrático medio bayesiano* se define como
$bmse(\hat{\theta}) = \mathbb{E}_{x,\theta}\left[(\hat{\theta}-\theta)^2\right] = \int_\theta \int_x (\hat{\theta}(x)-\theta)^2 p(x,\theta)dxd\theta$

El estimador MMSE es el que minimiza el BMSE. Realizando las operaciones pertinentes, se llega a que $\hat{\theta}_{MMSE} = \mathbb{E}[\theta|x]$ que es la media de la distribución *a posteriori* $p(\theta|x)$

La distribución *a posteriori* es la distribución de θ una vez observados los datos. Este estimador depende de la *prior* de θ y de los datos. Si el conocimiento previo que se asume sobre θ es débil respecto a los datos, el estimador ignora el conocimiento *a priori*. Lo opuesto ocurre si el conocimiento previo es fuerte respecto a los datos.

Estimador Máximo a Posteriori (MAP)

Es el análogo bayesiano al estimador del MLE. En este caso, en lugar de maximizar la verosimilitud se maximiza la distribución *a posteriori*. Entonces, el estimador queda definido por
$\hat{\theta}_{MAP} := argmax_\theta p(\theta|x) = argmax_\theta\{log(p(x|\theta)) + log(p\theta)\}$
donde el primer término es la función de verosimilitud y el segundo, la *prior*.

Ejemplo conceptual de estimadores:

Supongamos que un ganadero está interesado en estimar la cantidad promedio de leche producida por sus vacas en un determinado período de tiempo. Se presenta a continuación los dos posibles enfoques, el frecuentista y el bayesiano:

Estimador frecuentista:

El ganadero podría utilizar un enfoque frecuentista y estimar la cantidad promedio de leche producida por las vacas basándose únicamente en los datos observados de su muestra. Por ejemplo, podría calcular la media muestral de la cantidad de leche producida por cada vaca en un período de tiempo específico y utilizar esta media como estimación del verdadero valor promedio de leche producida por todas las vacas en la población. *Estimador bayesiano:*

Por otro lado, el ganadero también podría adoptar un enfoque bayesiano. En este caso, el ganadero podría tener información previa sobre la producción de leche de las vacas en temporadas anteriores. Esta información previa se utilizaría para establecer una distribución de probabilidad inicial (prior) sobre la cantidad promedio de leche producida por las vacas.

Bibliografía

Abzhanov, A., W. P. Kuo, C. Hartmann, B. R. Grant, P. R. Grant, and C. J. Tabin. 2006. «The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches». *Nature* 442 (7102): 563–67.

Abzhanov, A., M. Protas, B. R. Grant, P. R. Grant, and C. J. Tabin. 2004. «Bmp4 and morphological variation of beaks in Darwin's finches». *Science* 305 (5689): 1462–65.

Achtman, M. 1975. «Mating aggregates in Escherichia coli conjugation». *J. Bacteriol* 123 (2): 505–15.

Aïssani, B., and G. Bernardi. 1991. «CpG Islands, Genes and Isochores in the Genomes of Vertebrates». *Gene* 106 (2): 185–95. http://www.ncbi.nlm.nih.gov/pubmed/1937049

Allais-Bonnet, A., C. Grohs, I. Medugorac, S. Krebs, A. Djari, A. Graf, S. Fritz, et al. 2013. «Novel insights into the bovine polled phenotype and horn ontogenesis in Bovidae». *PLoS ONE* 8 (5): e63512.

Aminov, R. I. 2011. «Horizontal gene exchange in environmental microbiota». *Front Microbiol* 2: 158.

Arias, J. A., M. Keehan, P. Fisher, W. Coppieters, and R. Spelman. 2009. «A high density linkage map of the bovine genome». *BMC Genet* 10 (April): 18.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy. 2010. «PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice». *Science* 327 (5967): 836–40.

Bell, W., J. I. Urioste, N. Barlocco, A. Vadell, and R. Pérez Clariget. 2015. «Genetic and Environmental Factors Affecting Reproductive Traits in Sows in an Outdoor Production System». *Livestock Science* 182: 101–7. https://doi.org/10.1016/j.livsci.2015.10.025

Belle, E. M. S., N. Smith, and A. Eyre-Walker. 2002. «Analysis of the Phylogenetic Distribution of Isochores in Vertebrates and a Test of the Thermal Stability Hypothesis». *Journal of Molecular Evolution* 55 (3): 356–63. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12187388

Bennett, G. L., R. G. Tait, S. D. Shackelford, T. L. Wheeler, D. A. King, E. Casas, and T. P. L. Smith. 2019. «Enhanced estimates of carcass and meat quality effects for polymorphisms in myostatin and μ-calpain genes». *J Anim Sci* 97 (2): 569–77.

Benzécri, J. P. 1980. «Introduction à l'analyse des Correspondances d'apres un Exemple de Données Médicales». *Les Cahiers d'Analyse Des Données* V (3): 283–310.

Bernardi, G, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. 1985. «The Mosaic Genome of Warm-Blooded Vertebrates». *Science* 228 (4702): 953–58. http://www.ncbi.nlm.nih.gov/pubmed/4001930

Bezdíček, J., J. Šubrt, and R. Filipčík. 2008. «The Effect of Inbreeding on Milk Traits in Holstein Cattle in the Czech Republic». *Archives Animal Breeding* 51 (5): 415–25. https://doi.org/10.5194/aab-51-415-2008

Bigelow, H., and B. Rost. 2009. «Online Tools for Predicting Integral Membrane Proteins». Edited by Matthew J Peirce and Robin Editors Wait. *Methods In Molecular Biology Clifton NJ* 528: 3–23. https://link.springer.com/protocol/10.1007/978-1-60327-310-7_1

Blasco, A. 2025. *Mejora genética animal*. Editorial Síntesis.

Blows, M., and B. Walsh. 2009. «Spherical Cows Grazing in Flatland: Constraints to Selection and Adaptation». In *Adaptation and Fitness in Animal Populations: Evolutionary and Breeding Perspectives on Genetic Resource Management*, edited by Julius van der Werf, Hans-Ulrich Graser, Richard Frankham, and Cedric Gondro, 83–101. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-9005-9_6

Boggs, C. L. 2001. «Species and Speciation,» 14855–61. https://doi.org/10.1016/b0-08-043076-7/03073-4

Bourdieu, P. 2002. *Le Bal des célibataires. Crise de la Société Paysanne En Béarn*. Éditions du Seuil.

Bulmer, M. G. 1971. *The Effect of Selection on Genetic Variability*. *The American Naturalist* 105 (943): 201–11.

Bulmer, M. G. 1980. *The Mathematical Theory of Quantitative Genetics*. Clarendon Press.

Byars, S. G., D. Ewbank, D. R. Govindaraju, and S. C. Stearns. 2010. «Colloquium papers: Natural selection in a contemporary human population». *Proc Natl Acad Sci U S A* 107 Suppl 1 (January): 1787–92.

Caballero, A. 2020. *Quantitative Genetics*. Cambridge University Press.

Cacciò, S, K. Jabbari, G. Matassi, F. Guermonprez, J. Desgrès, and G. Bernardi. 1997. «Methylation Patterns in the Isochores of Vertebrate Genomes». *Gene* 205 (1-2): 25–34. https://doi.org/10.1016/S0378-1119(97)00560-X

Campbell, R. F., P. T. McGrath, and A. B. Paaby. 2018. «Analysis of Epistasis in Natural Traits Using Model Organisms». *Trends in Genetics* 34 (11): 883–98. https://doi.org/10.1016/j.tig.2018.08.002

Cannings, C., and A. W. Edwards. 1968. «Natural selection and the de Finetti diagram». *Ann. Hum. Genet.* 31 (4): 421–28.

Cardellino, R., and J. Rovira. 1987. *Mejoramiento genético animal*. 1.ª ed. Editorial Hemisferio Sur.

Castells Montes, D. 2008. «Evaluación de resistencia genética de ovinos corriedale a los nemátodos gastrointestinales en Uruguay: Heredabilidad y correlaciones genéticas entre el recuento de huevos de nematodos y características productivas». Tesis de Maestría, Facultad de Agronomía - Universidad de la República.

Castigeras, E. n.d. «Ecuaciones diferenciales. Una introducción para los cursos de Cálculo I y II». https://eva.fing.edu.uy/pluginfile.php/497334/mod_resource/content/1/Calcu2EcuacionesDiferenciales.pdf

Castle, W. E. 1903. «The laws of Galton and Mendel and some laws governing race improvement». *Proceedings of the American Academy of Arts and Sciences* 35 (8): 233–42.

Cellon, C., R. R. Amadeu, J. W. Olmstead, M. R. Mattia, L. F. V. Ferrao, and P. R. Munoz. 2018. «Estimation of Genetic Parameters and Prediction of Breeding Values in an Autotetraploid Blueberry Breeding Population with Extensive Pedigree Data». *Euphytica* 214 (87).

Chan, K. M., Y. T. Liu, C. H. Ma, M. Jayaram, and S. Sau. 2013. «The 2-micron plasmid of Saccharomyces cerevisiae: a miniaturized selfish genome with optimized functional competence». *Plasmid* 70 (1): 2–17.

Charlesworth, B., R. Lande, and M. Slatkin. 1982. «A Neo-Darwinian Commentary on Macroevolution». *Evolution* 36 (3): 474–98.

Charlesworth, D., and J. H. Willis. 2009. «The genetics of inbreeding depression». *Nat Rev Genet* 10 (11): 783–96.

Chimpanzee Sequencing and Analysis Consortium. 2005. «Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome». *Nature* 437 (7055): 69–87. http://www.ncbi.nlm.nih.gov/pubmed/16136131

Ciappesoni, G., I. De Barbieri, D. Gimeno, and F. Montossi. 2011. «Parámetros genéticos para índices de selección de la raza merino australiano». *Revista Argentina de Producción Animal*.
Cipolla, C. 1988. *Allegro ma non troppo con le leggi fondamentali della stupidità umana*. Il Mulino, Milano.
Cortazzo, P., C. Cerveñansky, M. Marín, C. Reiss, R. Ehrlich, and A. Deana. 2002. «Silent Mutations Affect in Vivo Protein Folding in *Escherichia coli*». *Biochemical and Biophysical Research Communications* 293 (1): 537–41. http://www.ncbi.nlm.nih.gov/pubmed/12054634
Cortes-Hernández, J., A. García-Ruiz, C. G. Vásquez-Peláez, and F. J. Ruiz-López. 2021. «Correlation of Genomic and Pedigree Inbreeding Coefficients in Small Cattle Populations». *Animals (Basel)* 11 (11).
Costantini, M., R. Cammarano, and G. Bernardi. 2009. «The Evolution of Isochore Patterns in Vertebrate Genomes». *BMC Genomics* 10 (i): 146. https://doi.org/10.1186/1471-2164-10-146
Council on Dairy Cattle Breeding (CDCB). n.d. https://www.uscdcb.com/
Crow, J. F., and M. Kimura. 1970. *An introduction to population genetics theory*. Harper & Row, New York.
Cunningham, E. P. 1987. «Crossbreeding the Greek Temple Model». *Journal of Animal Breeding and Genetics* 104 (1-2): 2–11.
Darwin, C. 1876. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. John Murray, Albemarle Street.
Denny, M., and S. Gaines. 2000. *Chance in Biology: Using Probability to Explore Nature*. Princeton University Press.
Dickerson, G. E. 1973. «Inbreeding and Heterosis in Animals». *Journal of Animal Science* Issue Symposium: 54–77.
Dobzhansky, T. 1964. «Biology, Molecular and Organismic». *American Zoologist* 4 (4): 443–52.http://icb.oxfordjournals.org/cgi/reprint/4/4/443.pdf
Dohm, M. R. 2002. «Repeatability Estimates Do Not Always Set an Upper Limit to Heritability». *Functional Ecology* 16: 273–80.
Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. «Rates of spontaneous mutation». *Genetics* 148 (4): 1667–86.
Drögemüller, C., T. Leeb, B. Harlizius, I. Tammen, O. Distl, M. Höltershinken, A. Gentile, A. Duchesne, and A. Eggen. 2007. «Congenital syndactyly in cattle: four novel mutations in the low density lipoprotein receptor-related protein 4 gene (LRP4)». *BMC Genet* 8 (February): 5.
Duret, L., D. Mouchiroud, and C. Gautier. 1995. «Statistical Analysis of Vertebrate Sequences Reveals That Long Genes Are Scarce in GC-Rich Isochores». *Journal of Molecular Evolution* 40 (3): 308–17. https://doi.org/10.1007/bf00163235
Duval, M., D. Dar, F. Carvalho, E. P. C. Rocha, R. Sorek, and P. Cossart. 2018. «HflXr, a homolog of a ribosome-splitting factor, mediates antibiotic resistance». *Proc Natl Acad Sci U S A* 115 (52): 13359–64.
Echeverría, P., A. Espasandin, P. Gaiero, A. L. Sánchez, and J. I. Urioste. 2012. *Manual de prácticas del curso de Zootecnia*. Facultad de Agronomía.
Edwards, A. W. 1994. «The fundamental theorem of natural selection». *Biol Rev Camb Philos Soc* 69 (4): 443–74.
Ermolaeva, M. D. 2001. «Synonymous Codon Usage in Bacteria». *Current Issues in Molecular Biology* 3 (4): 91–97. http://www.ncbi.nlm.nih.gov/pubmed/11719972

Espasandin, A. C., J. I. Urioste, H. Naya, and M. M. Alencar. 2013. «Genotype × production Environment Interaction for Weaning Weight in Angus Populations of Brazil and Uruguay». *Livestock Science* 151 (2): 264–70. https://doi.org/10.1016/j.livsci.2012.11.015

Eyre-Walker, A. 1992. «Evidence That Both G + C Rich and G + C Poor Isochores Are Replicated Early and Late in the Cell Cycle». *Nucleic Acids Research* 20 (7): 1497–1501. https://doi.org/10.1093/nar/20.7.1497

Eyre-Walker, A., and L. D. Hurst. 2001. «The Evolution of Isochores». *Nature Reviews Genetics* 2 (7): 549–55. http://opus.bath.ac.uk/4362/

Falconer, D. S. 1952. «The Problem of Environment and Selection». *The American Naturalist* 86 (830): 293–98.

Falconer, D. S. 1960. *Introduction to Quantitative Genetics*. Oliver & Boyd, Edinburgh.

Falconer, D. S. 1983. *Introducción a la genética cuantitativa*. CIA. Editorial Continental, S.A. de C.V.

Falconer, D. S., and T. F. C. MacKay. 1996. *Introduction to Quantitative Genetics*. 4th ed. London, England: Longman.

Fang, L., Y. Li, Y. Zhang, D. Sun, L. Liu, Y. Zhang, and S. Zhang. 2013. «Identification of brachyspina syndrome carriers in Chinese Holstein cattle». *J Vet Diagn Invest* 25 (4): 508–10.

Feldman, M. W., and R. C. Lewontin. 1975. «The Heritability Hang-up: The Role of Variance Analysis in Human Genetics Is Discussed». *Science* 190 (4220): 1163–68. https://doi.org/10.1126/science.1198102

Felsenstein, J. 2019. *Theoretical Evolutionary Genetics*. University of Washington. https://felsenst.github.io/pgbook/pgbook.html

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc.

Finetti, B. de. 1926. «Considerazioni matematiche sull'ereditarietà mendeliana». *Metron* 6 (1): 3–41.

Fisher, R. A. 1930. *The Nature of Probability*. Oxford University Press.

Fisher, R. A. 1958. *«The Genetical Theory of Natural Selection». Centen. rev.* 2:261–74.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. «Whole-genome random sequencing and assembly of Haemophilus influenzae Rd». *Science* 269 (5223): 496–512.

Fonseca, D. M., N. Keyghobadi, C. A. Malcolm, C. Mehmet, F. Schaffner, M. Mogi, R. C. Fleischer, and R. C. Wilkerson. 2004. «Emerging Vectors in the *Culex pipiens* Complex». *Science* 303 (5663): 1535–38. https://doi.org/10.1126/science.1094247

Fox, J. M., and I. Erill. 2010. «Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression». *DNA Research, an International Journal for Rapid Publication of Reports on Genes and Genomes* 17 (3): 185–96. https://doi.org/10.1093/dnares/dsq012

Frioni, N. 2012. «Estimación de heredabilidad del intervalo entre partos y su correlación genética con producción de leche en ganado holando uruguayo». Facultad de Agronomía - Universidad de la República.

Frioni, N. 2016. «Estimación de parámetros genéticos en características reproductivas y productivas para la raza holando en Uruguay». Tesis de Maestría, Facultad de Agronomía - Universidad de la República.

Galton, F. 1886. «Regression Towards Mediocrity in Hereditary Stature». *Journal of the Anthropological Institute* 15: 246–63.

Garrow, A. G., A. Agnew, and D. R. Westhead. 2005. «TMB-Hunt: An Amino Acid Composition Based Method to Screen Proteomes for Beta-Barrel Transmembrane Proteins». *BMC Bioinformatics* 6 (1): 56. http://www.ncbi.nlm.nih.gov/pubmed/15769290

Gianola, D. 1982. «Theory and Analysis of Threshold Characters». *Journal of Animal Science* 54 (5): 1079–96.

Gianola, D., S. Qanbari, and H. Simianer. 2013. «An evaluation of a novel estimator of linkage disequilibrium». *Heredity (Edinb)* 111 (4): 275–85.

Gillespie, J. H. 2000. «Genetic drift in an infinite population. The pseudohitchhiking model». *Genetics* 155 (2): 909–19.

Gillespie, J. H. 2004. *Population Genetics. A Concise Guide*. The Johns Hopkins University Press.

Goldberg, V., G. Ciappesoni, and I. Aguilar. 2012. «Genetic Parameters for Nematode Resistance in Periparturient Ewes and Post-Weaning Lambs in Uruguayan Merino Sheep». *Livestock Science* 147 (1): 181–87. https://doi.org/10.1016/j.livsci.2012.05.003

González, A. V., and E. Gianoli. 2004. «Morphological Plasticity in Response to Shading in Three Convolvulus Species of Different Ecological Breadth». *Acta Oecologica* 26 (June): 185–90.

Grandi, M. F., L. R. de Oliveira, S. L. Dans, and E. A. Crespo. 2012. «A Hunted Population in Recovery: Effective Population Size for South American Sea Lions from Patagonia». *Animal Biology* 62 (4): 433–50. https://doi.org/10.1163/157075612X642941

Grigoriev, A. 1998. «Analyzing Genomes with Cumulative Skew Diagrams». *Nucleic Acids Research* 26 (10): 2286–90. https://doi.org/10.1093/nar/26.10.2286

Gupta, A. P., and R. C. Lewontin. 1982. «A Study of Reaction Norms in Natural Populations of *Drosophila pseudoobscura*». *Evolution* 36 (5): 934–48.

Hadfield, J. D. 2010. «MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package». *Journal of Statistical Software* 33 (2): 1–22. https://www.jstatsoft.org/v33/i02/

Hamilton, M. B. 2009. *Population Genetics*. Wiley-Blackwell.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. 1998. «Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products». *Chem Biol* 5 (10): R245–249.

Hardy, G. H. 1908. «Mendelian Proportions in a Mixed Population». *Science* 28 (706): 49–50.

Hartl, D. L. 2020. *A Primer of Population Genetics and Genomics*. 4th ed. London, England: Oxford University Press.

Hartl, D. L, and A. G. Clark. 2007. *Principles of Population Genetics*. 4th ed. Sinauer.

Havenstein, G. B., P. R. Ferket, S. E. Scheideler, and B. T. Larson. 1994. «Growth, livability, and feed conversion of 1957 vs 1991 broilers when fed "typical" 1957 and 1991 broiler diets». *Poult Sci* 73 (12): 1785–94.

Henderson, C. R. 1949. «Estimation of Changes in Herd Environment». *Journal of Dairy Science* 32: 706.

Henderson, C. R. 1950. «Estimation of Genetic Parameters». *Ann Math Stat* 21: 309–10.

Henderson, C. R. 1959. «The Estimation of Environmental and Genetic Trends from Records Subject to Culling». *Biometrics* 15 (2): 192–218.

Henderson, C. R., and R. L. Quaas. 1976. «Multiple Trait Evaluation Using Relatives' Records». *Journal of Animal Science* 43 (6): 1188–97. https://doi.org/10.2527/jas1976.4361188x

Hill, W. G. 1972. «Estimation of realised heritabilities from selection experiments. I. Divergent selection». *Biometrics* 28 (3): 747–65.
Holsinger, K. E., and B. S. Weir. 2009. «Genetics in geographically structured populations: defining, estimating and interpreting F(ST)». *Nat Rev Genet* 10 (9): 639–50.
Hoque, M. N., A. Istiaq, R. A. Clement, K. M. Gibson, O. Saha, O. K. Islam, R. A. Abir, et al. 2020. «Insights Into the Resistome of Bovine Clinical Mastitis Microbiome, a Key Factor in Disease Complication». *Front Microbiol* 11: 860.
Houchmandzadeh, B. 2014. «An Alternative to the Breeder's and Lande's Equations». *G3 (Bethesda)* 4 (1): 97–108.
Houle, D. 1992. «Comparing evolvability and variability of quantitative traits». *Genetics* 130 (1): 195–204.
IMERL, Facultad de Ingeniería. n.d.a. «Geometría y Álgebra Lineal I». https://eva.fing.edu.uy/pluginfile.php/304733/mod_resource/content/2/manual.pdf.
IMERL, Facultad de Ingeniería. n.d.b. «Geometría y Álgebra Lineal II». https://eva.fing.edu.uy/pluginfile.php/103625/mod_resource/content/1/libro%20rojo.pdf
IMERL, Facultad de Ingeniería. n.d.c. «Integrales». https://eva.fing.edu.uy/course/view.php?id=1442&section=6#tabs-tree-start
IMERL, Facultad de Ingeniería. n.d.d. «Notas del curso de Probabilidad y Estadística». https://eva.fing.edu.uy/mod/folder/view.php?id=116419
Iraola, G., L. Spangenberg, B. Lopes Bastos, M. Graña, L. Vasconcelos, Á. Almeida, G. Greif, C. Robello, P. Ristow, and H. Naya. 2016. «Transcriptome Sequencing Reveals Wide Expression Reprogramming of Basal and Unknown Genes in *Leptospira biflexa* Biofilms». *mSphere* 1 (2).
Johannsen, W. L. 1903. *Über Erblichkeit in Populationen und in Reinen Linien: Ein Beitrag Zur Beleuchtung Schwebender Selektionsfragen*. Fischer (Jena).
Jong, G. de. 1990. «Quantitative Genetics of Reaction Norms». *J. Evol. Biol*: 447–68.
Kagawa, Y, H. Nojima, N. Nukiwa, M. Ishizuka, T. Nakajima, T. Yasuhara, T. Tanaka, and T. Oshima. 1984. «High Guanine Plus Cytosine Content in the Third Letter of Codons of an Extreme Thermophile. DNA Sequence of the Isopropylmalate Dehydrogenase of *Thermus thermophilus*». *The Journal of Biological Chemistry* 259 (5): 2956–60.
Kawashima, S., and M. Kanehisa. 2000. «AAindex: Amino Acid Index Database». *Nucleic Acids Research* 27 (1): 374. https://eva.fing.edu.uy/course/view.php?id=1442&section=6#tabs-tree-start
Kellenberger, R. T., K. J. R. P. Byers, R. M. De Brito Francisco, Y. M. Staedler, A. M. LaFountain, J. Schönenberger, F. P. Schiestl, and P. M. Schlüter. 2019. «Emergence of a Floral Colour Polymorphism by Pollinator-Mediated Overdominance». *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-018-07936-x
Keller, Lukas F., and Donald M. Waller. 2002. «Inbreeding Effects in Wild Populations». *Trends in Ecology and Evolution* 17 (5): 230–41.
Kimura, M. 1962. «On the probability of fixation of mutant genes in a population». *Genetics* 47 (June): 713–19.
Kimura, M., and T. Ohta. 1969. «The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population». *Genetics* 61 (3): 763–71.
Kingman, J. F. C. 1982a. «On the Genealogy of Large Populations». *Journal of Applied Probability* 19A: 27–43.
Kingman, J. F. C. 1982b. «The Coalescent». *Stochastic Processes and Their Applications* 13: 235–48.

Kipp, S., D. Segelke, S. Schierenbeck, F. Reinhardt, R. Reents, C. Wurmser, H. Pausch, et al. 2016. «Identification of a haplotype associated with cholesterol deficiency and increased juvenile mortality in Holstein cattle». *J Dairy Sci* 99 (11): 8915–31.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, et al. 2002. «A high-resolution recombination map of the human genome». *Nat Genet* 31 (3): 241–47.

Kong, A., G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, et al. 2010. «Fine-scale recombination rate differences between sexes, populations and individuals». *Nature* 467 (7319): 1099–103.

Lande, R. 1979. «Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry». *Evolution* 33 (1 Part. 2): 402–16.

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. 1985. «Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses». *Proc Natl Acad Sci U S A* 82 (20): 6955–59.

Li, C. C. 1955. *Population Genetics*. First edition. Chicago Press.

Loáisiga, C. H., O. Rocha, A. K. Brantestam, B. Salomon, and A. Merker. 2012. «Genetic Diversity and Gene Flow in Six Accessions of Meso-America Teosintes». *Genet Resour Crop Evol* 59: 95–111.

Lobo, I. 2008. «Pleiotropy: One Gene Can Affect Multiple Traits». *Nature Education* 1 (1): 10.

Loehlin, J. C., and A. A. Beaujean. 2017. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. Fifth edition. Routledge (Taylor; Francis Group).

López-Correa, R. D. 2013. «Variabilidad genética del desempeño de caballos criollos en pruebas de resistencia». Tesis de Maestría, Facultad de Agronomía - Universidad de la República.

López-Correa, R. D., F. Peñagaricano, G. Rovere, and J. I. Urioste. 2018. «Genetic Parameter Estimation for Long Endurance Trials in the Uruguayan Criollo Horse». *Journal of Animal Breeding* 135 (3): 186–93.

Lstibůrek, M., V. Bittner, G. R. Hodge, J. Picek, and T. F. C. Mackay. 2018. «Estimating Realized Heritability in Panmictic Populations». *Genetics* 208 (1): 89–95.

Lynch, M. 1991. «Methods for the Analysis of Comparative Data in Evolutionary Biology». *Evolution* 45 (5): 1065–80.

Lynch, M., and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates. Oxford University Press.

Ma, L., T. S. Sonstegard, J. B. Cole, C. P. VanTassell, G. R. Wiggans, B. A. Crooker, C. Tan, D. Prakapenka, G. E. Liu, and Y. Da. 2019. «Genome Changes Due to Artificial Selection in U.S. Holstein Cattle». *BMC Genomics* 20 (1). https://doi.org/10.1186/s12864-019-5459-x

Marco de Lima, B., E. P. Cappa, O. B. Silva-Junior, C. Garcia, S. D. Mansfield, and D. Grattapaglia. 2019. «Quantitative genetic parameters for growth and wood properties in Eucalyptus "urograndis" hybrid using near-infrared phenotyping and genome-wide SNP-based relationships». *PLoS One* 14 (6): e0218747.

Maynard Smith, J., and J. Haigh. 1974. «The hitch-hiking effect of a favourable gene». *Genet Res* 23 (1): 23–35.

McEwan, C. E., D. Gatherer, and N. R. McEwan. 1998. «Nitrogen-Fixing Aerobic Bacteria Have Higher Genomic GC Content Than Non-Fixing Species Within the Same Genus». *Hereditas* 128 (2): 173–78. http://www.ncbi.nlm.nih.gov/pubmed/9687237

McParland, S., J. F. Kearney, M. Rath, and D. P. Berry. 2007. «Inbreeding effects on milk production, calving performance, fertility, and conformation in Irish Holstein-Friesians». *J Dairy Sci* 90 (9): 4411–19.

Medugorac, I., D. Seichter, A. Graf, I. Russ, H. Blum, K. H. Göpel, S. Rothammer, M. Förster, and S. Krebs. 2012. «Bovine polledness–an autosomal dominant trait with allelic heterogeneity». *PLoS ONE* 7 (6): e39477.

Miglior, F., B. L. Muir, and B. J. Van Doormaal. 2005. «Selection Indices in Holstein Cattle of Various Countries». *Journal of Dairy Science* 88 (3): 1255–63. https://doi.org/10.3168/jds.s0022-0302(05)72792-2

Miglior, F., B. Szkotnicki, and E. B. Burnside. 1992. «Analysis of levels of inbreeding and inbreeding depression in Jersey cattle». *J Dairy Sci* 75 (4): 1112–18.

Moldovan, M. A. 2019. «[Prokaryotic and Mitochondrial Linear Genomes: Their Genesis, Evolutionary Significance, and the Problem of Replicating Chromosome Ends]». *Mol Biol (Mosk)* 53 (2): 218–24.

Mrode, R. 2014. *Linear Models for the Prediction of Animal Breeding Values*. 3rd ed. CABI.

Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, and G. Bernardi. 2004. «Correlations Between Genomic GC Levels and Optimal Growth Temperatures in Prokaryotes». *FEBS Letters* 573 (1-3): 73–77. http://www.ncbi.nlm.nih.gov/pubmed/15327978

Nagahata, H. 2004. «Bovine leukocyte adhesion deficiency (BLAD): a review». *J Vet Med Sci* 66 (12): 1475–82.

Naya, H., D. Gianola, H. Romero, J. I. Urioste, and H. Musto. 2006. «Inferring parameters shaping amino acid usage in prokaryotic genomes via Bayesian MCMC methods». *Mol Biol Evol* 23 (1): 203–11.

Naya, H., F. Peñagaricano, and J. I. Urioste. 2017. «Modelling female fertility traits in beef cattle using linear and non-linear models». *J Anim Breed Genet* 134 (3): 202–12.

Naya, H., H. Romero, A. Zavala, B. Álvarez, and H. Musto. 2002. «Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes». *Journal of Molecular Evolution* 55 (3): 260–64. http://www.ncbi.nlm.nih.gov/pubmed/12187379

Naya, H., J. I. Urioste, and J. Franco. 2002. *Identification of Production Environments and Presence of GxE Interactions in Uruguay Using Holstein Herds Records.* Vol. 2002. Session 1.

Nonnemann, B., U. Lyhs, L. Svennesen, K. A. Kristensen, I. C. Klaas, and K. Pedersen. 2019. «Bovine mastitis bacteria resolved by MALDI-TOF mass spectrometry». *J Dairy Sci* 102 (3): 2515–24.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory, J. Hagenblad, M. Kreitman, et al. 2002. «The extent of linkage disequilibrium in Arabidopsis thaliana». *Nat Genet* 30 (2): 190–93.

Novembre, J. A. 2002. «Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias». *Molecular Biology and Evolution* 19 (8): 1390–94. http://mbe.oxfordjournals.org/content/19/8/1390.long

Novozhilov, A. S., Y. I. Wolf, and E. V. Koonin. 2007. «Evolution of the Genetic Code: Partial Optimization of a Random Code for Robustness to Translation Error in a Rugged Fitness Landscape». *Biology Direct* 2 (1): 24. http://www.ncbi.nlm.nih.gov/pubmed/17956616

Otto, S. P., and T. Day. 2007. *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press.

Pielou, E. C. 1969. *An Introduction to Mathematical Ecology*. Wiley-Interscience.

Price, G. R. 1970. «Selection and covariance». *Nature* 227 (5257): 520–21.
Price, G. R. 1972. «Extension of covariance selection mathematics». *Ann Hum Genet* 35 (4): 485–90.
R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
Reeve, E. C. R., and F. W. Robertson. 1953. «Studies in Quantitative Inheritance: II. Analysis of a Strain of *Drosophila melanogaster* Selected for Long Wings». *Journal of Genetics* 51 (2): 276–316. https://doi.org/10.1007/bf03023299
Rice, Sean H. 2004. *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates. Oxford University Press.
Robertson, A. 1959. «The Sampling Variance of the Genetic Correlation Coefficient». *Biometrics* 15 (3): 469–85.
Robertson, A. 1966. «A Mathematical Model of the Culling Process in Dairy Cattle». *Animal Production* 8: 95–108.
Robinson, D. A., D. Falush, and E. J. Feil. 2010. *Bacterial Population Genetics in Infectious Disease*. Wiley-Blackwell.
Rocha, E. P. C. 2004. «Codon Usage Bias from tRNA's Point of View: Redundancy, Specialization, and Efficient Decoding for Translation Optimization». *Genome Research* 14 (11): 2279–86. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=525687&tool=pmcentrez&rendertype=abstract
Rocha, E. P. C. 2018. «Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics». *Mol Biol Evol* 35 (6): 1338–47.
Rocha, E. P. C., and Antoine Danchin. 2002. «Base Composition Bias Might Result from Competition for Metabolic Resources». *Trends in Genetics* 18 (6): 291–94. https://doi.org/10.1016/s0168-9525(02)02690-2
Rocha, E. P. C., and Edward J. Feil. 2010. «Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria?» Edited by Michael WEditor Nachman. *PLoS Genetics* 6 (9): 4. http://dx.doi.org/10.1371/journal.pgen.1001104
Rogers, A. R., J. C. Dunne, C. Romay, M. Bohn, E. S. Buckler, I. A. Ciampitti, J. Edwards, et al. 2021. «The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment». *G3 Genes|Genomes|Genetics* 11 (2). https://doi.org/10.1093/g3journal/jkaa050
Romero, H., E. Pereira, H. Naya, and H. Musto. 2009. «Oxygen and Guanine-Cytosine Profiles in Marine Environments». *Journal of Molecular Evolution* 69 (2): 203–6.
Rosenberg, N. A., and M. Nordborg. 2002. «Genealogical trees, coalescent theory and the analysis of genetic polymorphisms». *Nat Rev Genet* 3 (5): 380–90.
Roymondal, U., S. Das, and S. Sahoo. 2009. «Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to Escherichia Coli Genome». *DNA Research An International Journal for Rapid Publication of Reports on Genes and Genomes* 16 (1): 13–30. http://www.ncbi.nlm.nih.gov/pubmed/19131380
Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum, R. H. Waterston, R. K. Wilson, and D. C. Page. 2003. «Abundant gene conversion between arms of palindromes in human and ape Y chromosomes». *Nature* 423 (6942): 873–76.
Saccone, S., A. De Sario, J. Wiegant, A. K. Raap, G. Della Valle, and G. Bernardi. 1993. «Correlations Between Isochores and Chromosomal Bands in the Human Genome». *Proceedings of the National Academy of Sciences of the United States of America* 90 (24): 11929–33. https://doi.org/10.1073/pnas.90.24.11929

Sánchez, A. L., J. I. Urioste, F. Peñagaricano, K. Neimaur, I. Sienra, H. Naya, and R. Kremer. 2016. «Genetic parameters of objectionable fibers and of their associations with fleece traits in Corriedale sheep1». *Journal of Animal Science* 94 (1): 13–20. https://doi.org/10.2527/jas.2015-9619

Sardell, J. M., and M. Kirkpatrick. 2020. «Sex Differences in the Recombination Landscape». *Am Nat* 195 (2): 361–79.

Schmidt, P., J. Hartung, J. Rath, and H. P. Piepho. 2019. «Estimating Broad-Sense Heritability with Unbalanced Data from Agricultural Cultivar Trials». *Crop Science* 59: 525–36.

Searle, Shayle R. 1961. «Phenotypic, Genetic and Environmental Correlations». *Biometrics* 17 (3): 474–80.

Sharp, P. M., and W. H. Li. 1987. «The Codon Adaptation Index–a Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications». *Nucleic Acids Research* 15 (3): 1281–95. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3547335

Sharp, P. M., T. M. Tuohy, and K. R. Mosurski. 1986. «Codon Usage in Yeast: Cluster Analysis Clearly Differentiates Highly and Lowly Expressed Genes». *Nucleic Acids Research* 14 (13): 5125–43. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=311530&tool=pmcentrez&rendertype=abstract

Soppa, J. 2014. «Polyploidy in archaea and bacteria: about desiccation resistance, giant cell size, long-term survival, enforcement by a eukaryotic host and additional aspects». *J Mol Microbiol Biotechnol* 24 (5-6): 409–19.

Sorensen, D., and D. Gianola. 2002. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag.

Spangenberg, L., A. Correa, B. Dallagiovanna, and H. Naya. 2013. «Role of alternative polyadenylation during adipogenic differentiation: an in silico approach». *PLoS One* 8 (10): e75578.

Spangenberg, L., F. Battke, M. Graña, K. Nieselt, and H. Naya. 2011. «Identifying Associations Between Amino Acid Changes and Meta Information in Alignments». *Bioinformatics* 27 (20): 2782–89. http://www.ncbi.nlm.nih.gov/pubmed/21846738

Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong. 1996. «Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon». *J Bacteriol* 178 (3): 591–99.

Stojiljkovic, M., M. R. Foulquié-Moreno, and J. M. Thevelein. 2020. «Polygenic Analysis of Very High Acetic Acid Tolerance in the Yeast Saccharomyces Cerevisiae Reveals a Complex Genetic Background and Several New Causative Alleles». *Biotechnology for Biofuels* 13 (1). https://doi.org/10.1186/s13068-020-01761-5.

Sueoka, N. 1962. «On the Genetic Basis of Variation and Heterogeneity of DNA Base Composition». *Proceedings of the National Academy of Sciences of the United States of America* 48 (4): 582–92. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC220819/

Tagliani-Ribeiro, A., M. Oliveira, A. K. Sassi, M. R. Rodrigues, M. Zagonel-Oliveira, G. Steinman, U. Matte, N. J. Fagundes, and L. Schuler-Faccini. 2011. «Twin Town in South Brazil: a Nazi's experiment or a genetic founder effect?» *PLoS One* 6 (6): e20328.

Tajima, F. 1983. «Evolutionary Relationship of DNA Sequences in Finite Populations». *Genetics* 105 (2): 437–60.

Tallis, G. M. 1959. «Sampling Errors of Genetic Correlation Coefficients Calculated from Analyses of Variance and Covariance». *Australian Journal of Statistics* 1 (2): 35–43.

Thomsen, B., P. Horn, F. Panitz, E. Bendixen, A. H. Petersen, L. E. Holm, V. H. Nielsen, J. S. Agerholm, J. Arnbjerg, and C. Bendixen. 2006. «A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation». *Genome Res* 16 (1): 97–105.
Thurman, T. J., and R. D. Barrett. 2016. «The genetic consequences of selection in natural populations». *Mol Ecol* 25 (7): 1429–48.
Urioste, J. I. 2014. *Manual de Prácticos del Curso de Zootecnia*. Facultad de Agronomía, Udelar.
Van Inghelandt, D., J. C. Reif, B. S. Dhillon, P. Flament, and A. E. Melchinger. 2011. «Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm». *Theor Appl Genet* 123 (1): 11–20.
VanRaden, P. M. 2008. «Efficient methods to compute genomic predictions». *J Dairy Sci* 91 (11): 4414–23.
Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. «Heritability in the genomics era: concepts and misconceptions». *Nat Rev Genet* 9 (4): 255–66.
Vries, H. de. 1900. «Sur la loi de disjonction des Hybrides». *Comptes Rendus de l'Académie Des Sciences* 130 (-): 845–47.
Waddington, C. H. 1942. «Canalization of Development and the Inheritance of Acquired Characters». *Nature* 150 (3811): 563–65.
Wahlund, S. 1928. «Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus Betrachtet». *Hereditas* 11 (1): 65–106.
Walsh, B., and M. Lynch. 2018. *Evolution and Selection of Quantitative Traits*. Oxford University Press.
Wang, H., Y. Fang, L. Wang, W. Zhu, H. Ji, H. Wang, S. Xu, and Y. Sima. 2015. «Heterosis and Differential Gene Expression in Hybrids and Parents in *Bombyx mori* by Digital Gene Expression Profiling». *Scientific Reports* 5 (1). https://doi.org/10.1038/srep08750
Weinberg, W. 1908. «Über den Nachweis der Vererbung beim Menschen». *Jahreshefte Des Vereins für Vaterländische Naturkunde in Württemberg*. 64: 369–82.
Werf, J. van der. 2000. «Genetic Evaluation and Breeding Program Design». https://jvanderw.une.edu.au/gene422.htm
Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. «Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya». *Proc Natl Acad Sci U S A* 87 (12): 4576–79.
Woltereck, R. 1909. «Weitere Experimentelle Untersuchungen über das Wesen quantitativer Artunterschiede bei Daphniden». *Verhandlungen Der Deutschen Zoologischen Gesellschaft Leipzig, 19 110-173* 19: 110–73.
Wooldridge, J. M. 2013. *Introductory Econometrics: A Modern Approach*. 5th ed. South-Western, Cengage Learning.
Wright, D. 2015. «The Genetic Architecture of Domestication in Animals». *Bioinform Biol Insights* 9 (Suppl 4): 11–20.
Wright, F. 1990. «The 'Effective Number of Codons' Used in a Gene». *Gene* 87 (1): 23–29. http://www.ncbi.nlm.nih.gov/pubmed/2110097
Wright, S. 1918. «On the Nature of Size Factors». *Genetics* 3 (4): 367–74.
Wright, S. 1921. «Systems of Mating. I. the Biometric Relations between Parent and Offspring». *Genetics* 6 (2): 111–23.
Wright, S. 1922. «Coefficients of inbreeding and relationship». *American Naturalist* 56: 330–38.

Wright, S. 1932. «The Roles of Mutation, Inbreeding, Cross-breeding and Selection in Evolution». *Proceedings of the Sixth Annual Congress of Genetics* 1 (-): 356–66.
Wright, S. 1934a. «An Analysis of Variability in Number of Digits in an Inbred Strain of Guinea Pigs». *Genetics* 19 (6): 506–36.
Wright, S. 1934b. «The Method of Path Coefficients». *Ann. Math. Statist.* 5 (3): 161–215.
Wright, S. 1937. «The Distribution of Gene Frequencies in Populations». *Proc Natl Acad Sci U S A* 23 (6): 307–20.
Wright, S. 1943. «Genetics Isolation by Distance». *Genetics* 28 (2): 114–38.
Wright, S. 1951. «Ann Eugen. The Genetical Structure of Populations». *Ann Eugen* 15 (4): 323–54.
Wright, S. 1978. *Evolution and the Genetics of Populations*. Chicago Press.
Yakovchuk, P., E. Protozanova, and M. D. Frank-Kamenetskii. 2006. «Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix». *Nucleic Acids Research* 34 (2): 564–74. http://www.ncbi.nlm.nih.gov/pubmed/16449200
Yasuda, N. 1984. «A note on gene frequency estimation in the ABO and ABO-like system». *Jinrui Idengaku Zasshi* 29 (3): 371–80.
Yengo, L., N. R. Wray, and P. M. Visscher. 2021. «Extreme Inbreeding in a European Ancestry Sample from the Contemporary UK Population». *Nature Communications* 10 (3719).
Yue, X. P., C. Dechow, and W. S. Liu. 2015. «A limited number of Y chromosome lineages is present in North American Holsteins». *J Dairy Sci* 98 (4): 2738–45.
Yule, G. U. 1902. «Mendel's Laws and their Probable Relations to Intra-Racial Heredity». *New Phytologist* 1: 193–207, 222–38.
Zerial, M., J. Salinas, J. Filipski, and G. Bernardi. 1986. «Gene Distribution and Nucleotide Sequence Organization in the Human Genome». *The Federation of European Biochemical Societies Journal* 160 (3): 479–85. http://www.ncbi.nlm.nih.gov/pubmed/3780716
Zwald, N. R., K. A. Weigel, W. F. Fikse, and R. Rekaya. 2003. «Identification of factors that cause genotype by environment interaction between herds of Holstein cattle in seventeen countries». *J Dairy Sci* 86 (3): 1009–18.
Zwietering, M. H., I. Jongenburger, F. M. Rombouts, and K. van 't Riet. 1990. «Modeling of the bacterial growth curve». *Appl Environ Microbiol* 56 (6): 1875–81.

Esta publicación cuenta con el apoyo de la Comisión Sectorial de Enseñanza y Educación Permanente de la Universidad de la República. Forma parte de la serie Manuales Didácticos, que tiene como objetivo mejorar las condiciones de aprendizaje de los estudiantes y, al mismo tiempo, propiciar la autoformación del docente mediante la reflexión sobre sus prácticas y sobre el estado del arte en su disciplina. Secundariamente, esta publicación pretende colaborar en la constitución de tradiciones disciplinares y culturas educativas nacionales.

ISBN: 978-9974-0-2365-9

UNIVERSIDAD DE LA REPÚBLICA URUGUAY
Prorrectorado de **Enseñanza**