

APUNTES DE GENETICA CUANTITATIVA

A. Blasco

Capítulo 2

LA BIOMETRÍA DE LA VARIACIÓN CONTINUA

- 2.1. Componentes del valor Fenotípico. 11
 - 2.1.1. Genes mayores y genes menores 4
 - 2.1.2. El efecto del ambiente 2
 - 2.1.3. La descomposición del valor fenotípico 2
 - 2.1.4. La Interacción genotipo-medio 3
 - 2.1.5. Partición de la varianza fenotípica 2
- 2.2. Componentes del valor Genotípico
 - 2.2.1. Descomposición del valor genotípico 13
 - 2.2.1.1. Caso de un locus 7
 - 2.2.1.2. Caso de dos loci 3
 - 2.2.1.3. El modelo infinitesimal 3
 - 2.2.2. Descomposición de la varianza genotípica
- 2.3. Correlaciones entre parientes. 9
 - 2.3.1. Correlación genética entre parientes
 - 2.3.2. Correlación ambiental entre parientes
- 2.4. Parámetros genéticos de una población.
 - 2.4.1. Heredabilidad de un carácter
 - 2.4.2. Repetibilidad de un carácter
 - 2.4.3. Correlación genética entre caracteres
 - 2.4.4. Estimación de los parámetros genéticos

2.1. Componentes del valor fenotípico.

2.1.1. GENES MAYORES Y GENES MENORES

La genética se ha desarrollado hasta épocas recientes no observando los genes directamente, sino sus efectos. Mendel habla de “factores” o “caracteres” que producían efectos directamente observables, como el color verde o amarillo de los guisantes (la palabra “gen” para nombrar ese factor de herencia fue introducida por el biólogo danés Johannsen en 1909). La observación de las proporciones de distintos colores en las semillas o en flores de la progenie, según cómo fueran las plantas parentales, permitía inferir la existencia de un factor de herencia que lo causaba, por más que su naturaleza fuera un misterio. Las famosas leyes de Mendel explicaban cómo los caracteres eran transmitidos a la descendencia y permitían predecir los valores esperables en los hijos, conociendo los de los padres, para ciertos caracteres.

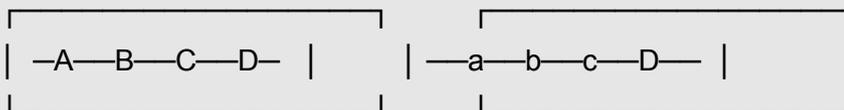
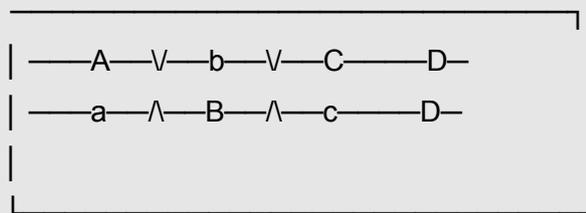
Como es sabido, los trabajos de Mendel, aunque conocidos en el ámbito de los productores de híbridos de plantas, fueron ignorados por los biólogos que buscaban las leyes de la herencia hasta 1900, año en que fueron redescubiertos por De Vries, Correns y Tschermak. La causa era que la mayor parte de los caracteres observados no parecía regirse por las leyes de Mendel, sino que mostraba una variación continua, algo de lo que Mendel era perfectamente consciente. En una carta al eminente botánico C. Nageli, Mendel dice “Me di cuenta de que los resultados obtenidos por mí no se compaginaban fácilmente con nuestro conocimiento científico contemporáneo” (Stern y E.R. Sherwood, CITA). La ciencia estaba en aquella época dominada por el Darwinismo, y el propio Darwin creía que los individuos deberían ser intermedios entre los valores de los padres. El problema de esta teoría es que, de ser cierta, rápidamente se extinguiría la variabilidad observada y todos los individuos serían intermedios, lo que contradecía la experiencia cotidiana.

Sin embargo, el propio Mendel hizo notar que cuando en un experimento de cruzamientos se observaban colores de flor intermedios, probablemente la causa era la combinación de varios de estos factores hereditarios (MENDEL, 1866). Aunque entonces no fue percibido, el hecho de que una variación continua pudiera provenir de muchos factores independientes permitía el mantenimiento de la variación, puesto que al fin y al cabo todos los factores que intervenían en el carácter volvían a aparecer en la generación siguiente. Los trabajos posteriores de YULE (1902), PEARSON (1904) y sobretodo de FISHER (1918), permitieron compaginar el hecho de que se observara una variación continua con la existencia de

genes¹. Así pues, se admitió finalmente un modelo explicativo de la transmisión de los caracteres según el cual coexisten *genes mayores* con un gran efecto en el carácter, con *genes menores* que tienen un efecto individual inapreciable pero que juntos producen una variación sustancial en el carácter. Cuando el número de genes que determinan un carácter aumenta, las posibilidades de nuevos genotipos aumentan también.

EJEMPLO 2.1

Consideremos un carácter dependiente de cuatro genes cada uno con dos alelos, y representemos con mayúsculas alelos favorables (mayor peso, por ejemplo) y con minúsculas alelos desfavorables. Consideremos el genotipo de un padre, que da lugar a dos gametos



En el ejemplo un padre intermedio ha producido un gameto con muchos alelos favorables y otro gameto con pocos. Así se explica un fenómeno conocido en ganadería, el que padres buenos den lugar a hijos regulares o malos.

¹ Es curioso que Pearson (1857-1936) no creyera en la existencia de los genes, cuando ya en 1904 propuso un modelo basándose en la acción independiente de muchos genes que permitía explicar bastante bien la variación continua de los caracteres. Su modelo sólo permitía dominancia mendeliana completa para cada gen, por lo que no se ajustaba lo bien que él hubiera deseado. Para una historia apasionante de las polémicas de principios del siglo XX en torno a los genes, ver Provine (1971)

Un carácter determinado por n genes con dos alelos cada uno puede dar lugar a 3^n genotipos diferentes. Así los caracteres métricos (peso, altura, etc.) que están determinados por más de 10 genes, pueden dar lugar a miles de genotipos diferentes. En realidad hay varios alelos posibles por gen y el número de genotipos posibles es aún mucho mayor. La figura 2.1 muestra cómo al aumentar el número de factores que controlan un carácter, la variación observada se va asemejando a una variación casi continua. Hoy en día la realidad se explica de forma más compleja, y se sabe que hay genes de efectos intermedios, o que genes mayores para un carácter pueden ser menores para otro carácter, pero los modelos utilizados siguen basándose esencialmente en este tipo de explicación genética.

Figura 2.1. Distribución de la segregación de dos alelos en el caso de un locus (a), seis (b) ó 24 loci (c). (Derivada de Falconer, 1996)

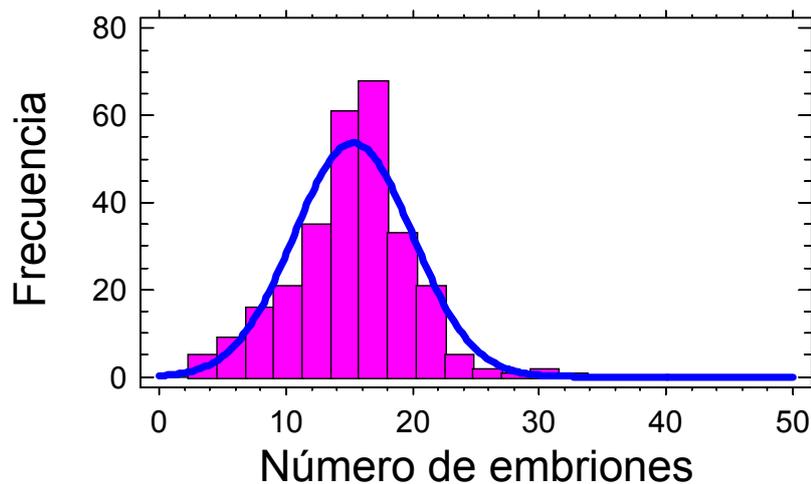
FIGURA 2.1. CERCA DE AQUI

2.1.2. EL EFECTO DEL AMBIENTE

Uno de los problemas principales para extraer conclusiones en experimentos en torno a la herencia es la existencia de factores ambientales que influyen sobre el carácter modificando su valor, lo que impide adscribir las observaciones o las medidas al efecto exclusivo de los genes. Mendel fue consciente de este hecho, y eligió para sus experimentos con cruzamientos de líneas puras aquellos caracteres que mostraban grandes diferencias entre líneas y que por tanto darían lugar a resultados con una influencia ambiental menor.

Estos factores ambientales son de dos tipos, los que denominamos *efectos sistemáticos*, llamados también *efectos fijos*, y que actúan aumentando o disminuyendo el valor de un carácter en todos los individuos (por ejemplo, el efecto del verano deprime el crecimiento de los animales y el invierno tiene un efecto favorable), y los *efectos aleatorios*, que están causados por infinidad de pequeños factores con un peso muy pequeño cada uno de ellos, y que inciden aleatoriamente sobre los individuos aumentando o disminuyendo el valor del carácter o caracteres en los que se está interesado². Los efectos ambientales aleatorios suavizan las diferencias que quedan entre las clases de los genotipos formados por varios genes, y la resultante es una distribución aproximadamente Normal (figura 2.2), lo que tiene ventajas prácticas, como comentaremos más adelante.

Figura 2.2. Distribución de frecuencias de un carácter métrico (número de embriones implantados en cerdo de una raza hiperprolífica) con una curva Normal superpuesta



2.1.3. LA DESCOMPOSICIÓN DEL VALOR FENOTÍPICO

El valor observado, llamado *valor fenotípico P* (del inglés 'Phenotype') tiene, pues, dos componentes, una *genética G* y una *ambiental E* (del inglés 'Environment'). La relación entre

² En el capítulo 3 se examinan con más detalle las diferencias entre efectos fijos y aleatorios

ambas es compleja, pero en una primera aproximación podemos suponer que el valor fenotípico es simplemente la suma de sus componentes³.

$$P = G + E$$

ó bien

$$P = m + G + E \quad (2.1)$$

si se desea referir estos efectos a la media (en ese caso la suma de los valores genéticos es cero, y también la suma de valores ambientales). La estimación de estos valores se podría hacer exponiendo individuos genéticamente idénticos (i.e.: líneas puras en plantas o clones en animales) a distintos ambientes, y resolviendo por mínimos cuadrados el modelo

$$P_{ijk} = m + G_i + E_j + e_{ijk}$$

donde P_{ijk} es el dato del individuo k de genotipo i en el ambiente j , y e_{ijk} es la componente residual que explica el que haya variación entre individuos del mismo genotipo en el mismo ambiente. Por motivos de coste (particularmente en el caso de animales) es infrecuente realizar esta estimación.

Es importante notar la naturaleza estadística de estas componentes: si nosotros comparamos ocho valores genéticos tendremos un resultado distinto de si comparamos cuatro, puesto que los resultados son relativos a cada experimento. Lo mismo puede decirse del efecto del ambiente.

Los efectos ambientales aleatorios, por estar causados por varios factores independientes de pequeño efecto cada uno de ellos, se distribuyen con arreglo a una ley Normal, de

³ Obsérvese que el valor fenotípico podría expresarse con otra ley, por ejemplo el producto de sus componentes. Motivos prácticos relacionados con la estimación y el análisis de la variación aconsejan usar un modelo aditivo como el que aquí expresamos. Es interesante hacer notar que Fisher y Cochran intentaron presentar un análisis de la varianza multiplicativo antes que aditivo. Dado que la varianza aumenta con la media (por ejemplo, hay más variabilidad si el tamaño de camada de una raza de corderos es 3 que si es 1.2) , este modelo parece más adecuado, puesto que expresa los efectos indicando en qué porcentaje modifican la media, pero dificultades operativas hicieron que ambos cambiaran de opinión (COCHRAN, 1977? CITA)

acuerdo al *Teorema Central del Límite*. Por la misma razón, si el carácter sólo está determinado por genes menores, los efectos genéticos se distribuyen de forma Normal, y el valor fenotípico también. Esta recurrencia a la ley Normal tiene muchas ventajas de cómputo e interpretación; por ejemplo, si dos variables se distribuyen conjuntamente de forma Normal, la regresión de una sobre la otra es lineal, y si su correlación es cero entonces son variables independientes⁴.

2.1.3. LA INTERACCIÓN GENOTIPO-MEDIO

Como en todo diseño factorial, puede haber interacciones entre los factores; esto es, podría ocurrir que la suma del valor genotípico y ambiental no diera lugar al valor fenotípico. Es conocido que los mejores genotipos en ciertos ambientes no necesariamente son los mejores en otros ambientes. Por ejemplo, las vacas Frisonas, mucho más productivas en climas templados que las razas locales del Sudeste asiático, producen por término medio menos que estas en climas húmedos y calurosos. Una preocupación permanente de los ganaderos que compran reproductores mejorados a núcleos de selección es si los descendientes de los mejores animales, que serán criados en sus granjas con ambientes menos cuidados que el ambiente de los núcleos, serán también los que mejor rindan. Para incluir esta interacción entre el genotipo y el medio, la descomposición del valor fenotípico pasa a ser

$$P = m + G + E + GxE \quad (2.2)$$

Donde el término GxE indica la interacción entre el genotipo y el medio (a veces se representa como $G * E$ para evitar confusiones con el signo x , que aquí no indica multiplicación).

En la práctica es difícil contar con la interacción en los programas de mejora genética, por lo que se tiende simplemente a evitar que exista dentro de lo posible, generando líneas o cruzamientos adaptados a medios concretos si es necesario. Por ejemplo, se han creado aves de cuello desnudo para producir en climas cálidos, o se han realizado cruzamiento entre razas productivas y razas locales en vacuno y ovino de carne .

⁴ Cosa que no ocurre si las variables no son Normales, algo que se olvida con frecuencia.

2.1.4. PARTICIÓN DE LA VARIANZA FENOTÍPICA

La variabilidad observada se puede descomponer en una parte atribuible a los genes y otra atribuible al ambiente. Si calculamos la varianza del valor fenotípico (*varianza fenotípica*) en la ecuación (2.1), tenemos

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + 2 \text{cov}(G,E)$$

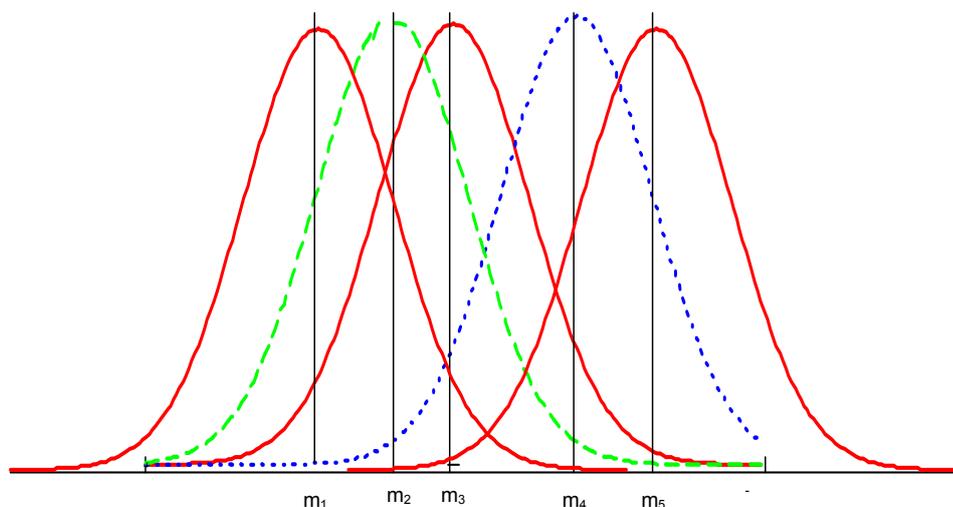
Donde σ_G^2 es la varianza debida a los genes (*varianza genética*), σ_E^2 es la debida al ambiente (*varianza ambiental*) y el término $\text{cov}(G,E)$ es la covarianza entre el valor genotípico y el ambiental. A la proporción de la varianza fenotípica que se debe a los genes se le llama *heredabilidad en sentido amplio*

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2} \quad (2.3)$$

y es poco utilizada debido a las dificultades que presenta su estimación, puesto que en una población no es posible separar la componente genética de la ambiental en cada individuo. Sí es posible, en plantas o en líneas altamente consanguíneas de animales experimentales (como la mosca *Drosophila melanogaster*, por ejemplo), separar la variabilidad debida al ambiente, y se pueden hacer cruzamientos de estas líneas para tener una estimación de la varianza genética, pero estos experimentos se basan en que la variabilidad debida al ambiente es la misma para cada genotipo, lo que no es necesariamente cierto.

En ocasiones, cuando se dispone de varias *líneas puras* en plantas (todos los individuos de una línea pura tienen en mismo genotipo), se calcula la varianza entre las medias de estas líneas y se le llama también varianza genética. Debido al alejamiento entre las medias de estas líneas, la varianza fenotípica global es muy similar a la varianza genética (Figura 2.3), lo que da lugar a expresiones de la heredabilidad en sentido amplio próximas a la unidad. Este uso de la heredabilidad en sentido amplio difiere del que aquí le estamos dando, puesto que nosotros nos referimos al valor genotípico de los individuos de una población, y estamos interesados en conocer la proporción en que la variabilidad observada es debida a estos valores genotípicos.

Figura 2.3. Distribución de los valores fenotípicos de cinco líneas puras.



El término $cov(G,E)$ es distinto de cero cuando hay una asociación entre los valores genotípicos y los ambientales. Por ejemplo, frecuentemente los granjeros de vacuno de leche que compran semen caro (de alto valor genético), crían a las hijas obtenidas con ese semen dándoles cuidados particulares, puesto que es un animal al que aprecian mucho. Así, los mejores valores ambientales van a parar a las hijas con el mejor valor genotípico, creándose una covarianza positiva que, al no ser tenida en cuenta al evaluar a los animales, genera una sobrevaloración de estas hijas (este fenómeno, conocido como *tratamiento preferencial no declarado*, es uno de los problemas más graves de la evaluación de reproductores en vacuno de leche). Esta covarianza se procura, pues, que desaparezca en la evaluación de reproductores⁵. Las estimas de la heredabilidad pueden también verse afectadas por la presencia de esta covarianza.

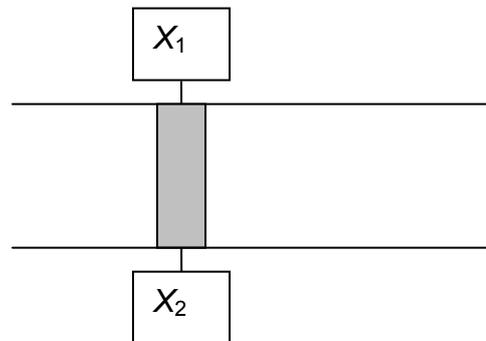
2.2. Componentes del valor genotípico

2.2.1. DESCOMPOSICIÓN DEL VALOR GENOTÍPICO

⁵ Por alguna razón se confunde a veces la covarianza genotipo medio con las interacciones genotipo medio. Son dos conceptos no relacionados entre sí, puede haber covarianza genotipo medio sin interacciones o viceversa. De hecho, utilizando el modelo completo de la ecuación 2.2, puede haber covarianzas entre el genotipo y la interacción o entre el medio y la interacción.

2.2.1.1. Carácter determinado por un locus

Consideremos un carácter determinado por un locus autosómico (es decir, no ligado al sexo).



Según qué alelo se sitúe en el lugar X_1 afectará a un carácter de una u otra forma, y lo mismo podemos decir de los alelos que se sitúen en X_2 . Consideremos un carácter determinado por un gen con dos alelos (A, a). Un valor genotípico (por ejemplo, el del genotipo Aa) está determinado por los efectos individuales de los alelos A, a , y por su interacción. Si pudiéramos medir los valores genotípicos, podríamos estimar el valor de estos efectos y su interacción tal y como se hace en el análisis de varianza convencional con un diseño factorial con dos efectos. Podríamos considerar que X_1 es un *efecto* con dos *niveles* (A, a), o con más niveles si hubiera más alelos. X_2 es otro efecto con otros dos (o más) *niveles*, (A, a). Tendríamos entonces un modelo de dos efectos con interacción.

$$G = m + X_1 + X_2 + X_{12}$$

efectos de los alelos interacción entre alelos

- A la suma de efectos X_1 y X_2 , que son los efectos de los alelos del mismo locus, se le llama *valor aditivo A* del locus.

- A la interacción X_{12} entre los efectos de los alelos del mismo locus se les llama *valor dominante D* del locus, y no debe confundirse con la dominancia mendeliana.

El modelo se puede expresar, pues, como

$$G = m + A + D \tag{2.4}$$

En el caso de que el modelo sea equilibrado, podrían estimarse estos efectos simplemente como efectos medios. Si el modelo no está equilibrado; esto es, la población no está en equilibrio Hardy-Weinberg, Encuentre un error: en la parte de modelo infinitesimal (documento 1), la condición de independencia de los efectos de cada locus esta relacionado con la covarianza entre alelos de distintos loci en la misma gameta, o desequilibrio de ligamiento. No con Hardy-Weinberg, informalmente hablando (la covarianza entre alelos dentro de locus es del orden n , mientras que la covarianza entre alelos de distintos loci en la misma gameta, o desequilibrio de ligamiento, es del orden n^2). Aparte de esto, el equilibrio de H-W se obtiene inmediatamente, para todos los loci (ignorando el problema de poblaciones finitas), no así el otro. Es el desequilibrio de ligamiento que esta vinculado al Bulmer effect, y no el de H-W. entonces hay que estimar los efectos por mínimos cuadrados. Esta estimación forma parte de la teoría estadística que habitualmente se enseña en los manuales, y no la desarrollaremos aquí.

EJEMPLO 2.2

Enunciado: En la siguiente tabla figuran los valores genotípicos y las frecuencias de un gen con dos alelos. Hallar el valor aditivo y el dominante.

	AA	Aa	aa
Valor	0.75	2.00	-3.00
Frecuencia	0.16	0.48	0.36

Resolución: La población se encuentra en equilibrio, puesto que las frecuencias de los alelos son

$$p = 0.16 + \frac{1}{2} 0.48 = 0.4$$

$$q = 1 - 0.4 = 0.6$$

y las frecuencias genotípicas son las de Hardy Weinberg, $\text{frec}(AA) = p^2 = 0.16$, $\text{frec}(Aa) = 2pq = 0.48$, $\text{frec}(aa) = q^2 = 0.36$. Si disponemos la tabla como en un análisis de varianza factorial, con el efecto X_1 en columnas, con sus dos niveles (A, a), y el efecto X_2 en filas, con sus dos niveles (A, a), y con las frecuencias entre paréntesis, tenemos

A [X_1] a [X_1]

A [X ₂]	0.75 (0.16)	2.00 (0.24)	1.50 (0.40)
a [X ₂]	2.00 (0.24)	-3.00 (0.36)	-1.00 (0.60)
	1.50 (0.40)	-1.00 (0.60)	0

Nota: Recuérdese que las medias marginales se calculan como sigue:

$$A [X_1] = \frac{0.75 \times 0.16 + 2 \times 0.24}{0.16 + 0.24} = 1.5$$

$$a [X_1] = \frac{2 \times 0.24 - 3 \times 0.36}{0.24 + 0.36} = 0.6$$

y de la misma forma se calculan A [X₂] y a [X₂].

La media es

$$m = 1.5 \times 0.4 - 1 \times 0.6 = 0$$

Los valores genéticos *G*, aditivos *A*, y dominantes *D*, se refieren habitualmente a la media, que en este caso es cero. Por tanto estos valores son

$$G_{AA} = 0.75$$

$$A_{AA} = A [X_1] + A [X_2] = 1.5 + 1.5 = 3$$

$$D_{AA} = G_{AA} - A_{AA} = 0.75 - 3 = -2.25$$

$$G_{Aa} = 2.00$$

$$A_{Aa} = \frac{1}{2} \{ A [X_1] + a [X_2] \} + \frac{1}{2} \{ A [X_1] + a [X_2] \} = (1.5 - 1 + 1.5 - 1) / 2 = 0.5$$

$$D_{Aa} = G_{Aa} - A_{Aa} = 2 - 0.5 = 1.5$$

$$G_{aa} = -3$$

$$A_{aa} = a [X_1] + a [X_2] = -1 + (-1) = -2$$

$$D_{aa} = G_{aa} - A_{aa} = -3 - (-2) = -1$$

El valor aditivo de Aa es exactamente intermedio entre el de AA y el de aa. El valor aditivo de aa es -2, y si sustituimos un alelo 'a' por uno 'A' tenemos el valor aditivo del heterocigoto Aa, que es 0.5, pero si sustituimos el alelo 'a' restante por un alelo 'A', tenemos el valor aditivo del homocigoto, que es 3. En cada paso hemos añadido 1.5. A este valor se le llama *efecto de sustitución de un gen*.

De forma general, si asignamos los valores (a , d , $-a$) a los genotipos (AA,Aa,aa) respectivamente⁶, y si la población está en equilibrio,

	AA	Aa	aa
Valor genotípico	a	d	$-a$
Frecuencia	p^2	$2pq$	q^2

	A [X_1]	a [X_1]
A [X_2]	$a (p^2)$	$d (pq)$
a [X_2]	$d (pq)$	$-a (q^2)$

$$A[X_1] = A[X_2] = \frac{a \cdot p^2 + d \cdot pq}{p^2 + pq} = \frac{a \cdot p + d \cdot q}{p + q} = a \cdot p + d \cdot q$$

$$a[X_1] = a[X_2] = \frac{-a \cdot q^2 + d \cdot pq}{q^2 + pq} = \frac{-a \cdot q + d \cdot p}{p + q} = -a \cdot q + d \cdot p$$

El efecto de sustitución de un gen será

$$\alpha = A[X_1] - a[X_1] = ap + dq - (-aq + dp) = a(p + q) + d(q - p) = a + d(q - p) \quad (2.5)$$

En los modelos de análisis de la varianza, los efectos suelen referirse a la media de la población.

$$m = ap^2 + d \cdot 2qp - aq^2 = a(p - q) + 2pqd$$

El efecto del alelo A es $A[X_1] - m = A[X_2] - m$

$$\begin{aligned} A[X_1] - m &= ap + dq - a(p - q) + 2pqd = dq + aq + 2pqd = aq + dq(1 - 2p) = \\ &= q[(a + d(1 - p - p))] = q[(a + d(q - p))] = q \cdot \alpha \end{aligned}$$

análogamente se deduce el efecto del alelo a,

⁶ No confundir el alelo a con el valor genotípico a. La nomenclatura está tan extendida que hemos preferido no alterarla.

$$a[X_1] - m = -p \cdot \alpha$$

ahora ya podemos calcular los efectos genotípico, aditivo y dominante

$$\begin{aligned} G_{AA} &= a - m = a - m = a - a(p - q) - 2d pq = a(1 - p - q) - 2dpq = \\ &= -2aq - 2dpq = -2q(a - dp) = -2q(a - dp + dq - dq) = -2q(\alpha - dq) \end{aligned}$$

$$A_{AA} = (A[X_1] - m) + (A[X_2] - m) = 2q\alpha$$

$$D_{AA} = G_{AA} - A_{AA} = -2q(\alpha - dq) - 2q\alpha = -2q^2d$$

De análoga forma se deducen los efectos de los otros genotipos, con lo que finalmente tenemos

	AA	Aa	aa
G	$-2q(\alpha - dq)$	$(q - p)\alpha + 2pqd$	$-2p(\alpha - dp)$
A	$2q\alpha$	$(q - p)\alpha$	$-2p\alpha$
D	$-2q^2d$	$2pqd$	$-2p^2d$

OBSERVESE

1) Obsérvese la naturaleza estrictamente estadística de esta descomposición. Los valores aditivos son los efectos del modelo (2.7), y el valor dominante la interacción. Este valor dominante poco tiene que ver con la dominancia en sentido mendeliano: obsérvese que los tres genotipos tienen valor aditivo y valor dominante. No es imposible proponer modelos distintos al (2.7), por ejemplo modelos multiplicativos en los que el valor genotípico no se explica por la suma de efectos sino por el producto de efectos (ver, p. ej. CHEVALET, cita).

2) Obsérvese que los valores aditivos y dominante dependen de las frecuencias génicas de la población, no son, pues, características biológicas estrictamente hablando, sino características de tipo estadístico. El efecto de sustitución de un gen también depende de

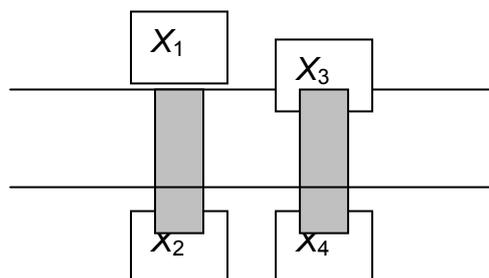
las frecuencias génicas, y (salvo cuando no hay dominancia mendeliana; $d=0$) es mayor en valor absoluto si las frecuencias son extremas.

3) Obsérvese que las interacciones, por término medio, no se heredan. Los gametos contienen X_1 o X_2 , pero no ambos alelos. Por tanto sólo se hereda, por término medio, el valor aditivo, razón por la cual se le conoce como *valor de mejora*.

4) El cálculo de estos efectos como valores medios de los niveles es sólo válido cuando la población está en equilibrio, en caso contrario hay que hacer la clásica deducción por mínimos cuadrados que figura en los manuales de estadística.

2.2.1.2. Carácter determinado por dos loci

Consideremos ahora un carácter determinado por dos loci⁷ autosómicos.



Según qué alelo se sitúe en el lugar X_1 afectará a un carácter de una u otra forma, y lo mismo podemos decir de los alelos que se sitúen en cada X_i . Un valor genotípico (por ejemplo, el del genotipo AaBB) está determinado por los valores de las variables efectos individuales de los alelos A, a, B y B y por sus interacciones. Si pudiéramos medir todos los posibles valores genotípicos, podríamos estimar el valor de estos efectos y sus interacciones tal y como se hace en el análisis de varianza convencional con un diseño factorial con cuatro efectos. Podríamos considerar que X_1 es un *efecto* con dos *niveles* (A, a), o con más niveles si hubiera más alelos. X_3 es otro efecto con otros dos (o más) *niveles*, (B, b), y lo mismo ocurre con X_2 y X_4 . Tendríamos entonces un modelo con cuatro efectos con interacciones dobles, triples y cuádruple.

⁷ El plural de *locus* en latín es *loci*, aunque en algunos textos, particularmente en los textos franceses, se usa *locus* para ambos.

$$\begin{aligned}
 G = & m & + X_1 + X_3 + X_2 + X_4 + & \text{efectos de los alelos} \\
 & & + X_{12} + X_{34} + & \text{interacciones entre alelos del un mismo locus} \\
 & & + X_{13} + X_{14} + X_{23} + X_{24} + & \text{interacciones entre alelos de distinto locus} \\
 & & + X_{123} + X_{124} + X_{134} + X_{234} + & \text{interacción entre alelo y locus} \\
 & & + X_{12\ 34} & \text{interacción entre los dos loci}
 \end{aligned}$$

▪ A la suma de efectos X_1 y X_2 , que son los efectos de los alelos del mismo locus, se le llama *valor aditivo* A_1 del primer locus y a la suma de X_3 y X_4 *valor aditivo* A_2 del segundo locus.

▪ A las interacciones dobles X_{12} y X_{34} entre los efectos de los alelos del mismo locus se les llama *valor dominante* D_1 del primer locus y D_2 del segundo locus respectivamente, y no debe confundirse con la dominancia mendeliana.

▪ A las interacciones dobles entre los efectos de los alelos de distinto locus se les llama *valor epistático aditivo por aditivo* AA

▪ A las interacciones entre efectos de un alelo y efecto dominante del otro locus se les denomina *valor epistático aditivo por dominante* AD

▪ Finalmente, a las interacciones entre efectos dominantes se les denomina *valor epistático dominante por dominante* DD , con lo que la expresión del valor genotípico es

$$G = A + D + AA + AD + DD \quad (2.6)$$

Esta notación se puede generalizar a más de dos loci, pero no lo haremos puesto que no es factible estimar valores epistáticos de orden tan elevado. De hecho rara vez se puede estimar la epistasia, como no sea en experimentos de laboratorio o usando líneas puras y sus cruzamientos en plantas. El cálculo de valores medios se realiza exactamente igual que en el ejemplo anterior.

2.2.1.3. El modelo infinitesimal

Fisher (1918) propuso un modelo según el cual los caracteres estaban determinados por infinitos loci de efecto infinitesimal cada uno de ellos ⁽⁸⁾. Así, recurriendo al Teorema central del límite, podemos sostener que el valor genético sigue una distribución Normal. Este teorema indica que la suma (o la media) de muchas variables independientes de pequeño efecto cada una, sigue una distribución Normal. Para asegurar la independencia, no debe haber efecto genético epistático y la población debe estar en equilibrio Hardy-Weinberg, y para que las variables sean de pequeño efecto no debe haber genes mayores. En este caso, la población tiene dos propiedades importantes derivadas de la normalidad de la distribución:

1) Todas las regresiones son lineales (lo que utilizaremos más adelante al estimar el valor aditivo de los individuos por regresión lineal).

2) La distribución de los errores de la regresión es Normal con varianza constante (lo que utilizaremos en el capítulo próximo para calcular el intervalo de confianza de la estimación del valor aditivo).

En este modelo los valores genéticos, aditivos o dominantes son la suma de los correspondientes valores en cada gen:

$$G = m + G_1 + G_2 + \dots + G_n = m + A + D \quad (2.7)$$

$$A = A_1 + A_2 + \dots + A_n$$

$$D = D_1 + D_2 + \dots + D_n$$

Una característica de este modelo es que los cambios en las frecuencias génicas son infinitesimales. Sin embargo la selección es posible porque cambios infinitesimales en las frecuencias génicas pueden producir cambios en la media de la población.

EJEMPLO 2.3

Supongamos genes de efecto estrictamente aditivo con las mismas frecuencias p y q . La media será

⁸ Este artículo, el más importante en la historia de la genética cuantitativa, fue presentado en 1916 para su publicación en los Proc. of the Royal Society, pero debido al informe negativo de los dos revisores fue rechazado para su publicación. Fisher logró publicarlo más tarde en los Proc. de la Royal Soc. de Edimburgo

$$m = 2n [a p^2 + 0 \cdot 2pq + (-a)q^2] = 2na (p-q)(p+q) = 2na (p-q) = 2na (p-1+p) = \\ = 2na (2p-1)$$

La diferencia entre medias en una generación y la siguiente debido a un cambio de frecuencias de p es

$$m' - m = 4na (p' - p)$$

Supongamos que el cambio en la media *no* es infinitesimal, sino que es apreciable. Pese a ello, $p' - p$ puede ser infinitesimal.

$$p' - p = \frac{m' - m}{4na}$$

para que $p' - p$ sea un infinitésimo, basta con que a sea de orden $1/\sqrt{n}$, puesto que $n/\sqrt{n} = \sqrt{(n^2/n)} = 1/\sqrt{n}$, con lo que

$$p' - p = \frac{m' - m}{4\sqrt{n}} \quad \text{y por tanto } p' - p \text{ es un infinitésimo.}$$

En el modelo con varios genes, si una pareja de reproductores produjera todos los gametos posibles y tuviera todos los hijos posibles, la media de los valores aditivos de todos ellos coincidiría con la media de los valores aditivos de sus padres. Cada uno de esos hijos tendría, sin embargo, un valor aditivo diferente, puesto que uno habría heredado de sus progenitores una combinación de alelos y otro habría heredado otra combinación, como indicábamos en el ejemplo 2.1. Cada valor aditivo se podría representar como la media de los valores de los padres (A_s y A_d) más una cantidad A_e que es lo que le falta o le sobra a cada hijo para ajustarse al valor aditivo que tiene realmente.

$$A_H = \frac{1}{2} A_s + \frac{1}{2} A_d + A_e \quad (2.8)$$

Cuando se trata de un modelo de infinitos genes, los valores aditivos son variables continuas y su distribución Normal, y A_e actúa como un residuo independiente de los valores aditivos de los padres. Los valores aditivos de los padres son independientes cuando los apareamientos se realizan al azar.

OBSERVESE

1) En el modelo infinitesimal no hay fijación o pérdida de genes, simplemente cambian sus frecuencias.

2.2.2. DESCOMPOSICION DE LA VARIANZA GENETICA

2.2.2.1. Un sólo locus

Obsérvese que, en el equilibrio, todas las covarianzas entre componentes son nulas porque los efectos aditivo y dominante provienen de un modelo estadístico factorial en el que estas componentes son independientes. Por tanto, a partir de la fórmula (2.4) podemos descomponer la varianza genética en

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$$

Tomemos el caso de un gen con dos alelos. La varianza aditiva será, por definición de varianza,

$$\begin{aligned} \sigma_A^2 &= (A_{AA})^2 p^2 + (A_{Aa})^2 2pq + (A_{aa})^2 q^2 = (2q\alpha)^2 \cdot p^2 + (q-p)^2 \alpha^2 \cdot 2pq + (-2p\alpha)^2 \cdot q^2 \\ &= 2pq \alpha^2 (2pq + q^2 - 2qp + p^2 + 2pq) = 2pq \alpha^2 \end{aligned} \quad (2.9)$$

Análogamente se deduce la varianza dominante

$$\sigma_D^2 = (2pq d)^2$$

La varianza genotípica es la suma de ambas (si no hay covarianza; esto es, en situaciones de equilibrio).

$$\sigma_G^2 = 2pq \alpha^2 + (2pq d)^2$$

OBSERVESE

- 1) En situación de aditividad mendeliana; esto es, cuando $d = 0$ y por tanto el genotipo Aa es intermedio entre el AA y el aa, la varianza dominante es nula y la varianza genética coincide con la aditiva. En estos casos la varianza es máxima para frecuencias génicas intermedias $p = q = 0.5$.
- 2) En caso de dominancia mendeliana; esto es, cuando $a = d$, la varianza aditiva máxima no se alcanza con frecuencias intermedias. Esto se debe a que α depende de la diferencia entre las frecuencias génicas cuando hay dominancia.
- 3) Si hubiera una ventaja completa del heterocigoto ($a = 0$), seguiría habiendo varianza aditiva salvo para frecuencias intermedias, en las que $\alpha = d(q - p)$ sería nulo. La confusión proviene de que la dominancia mendeliana no es la misma que la dominancia estadística que estamos tratando en este capítulo. La varianza aditiva no está necesariamente producida por genes con efecto aditivo (esto es, con $d=0$), sino que puede estar producida por genes con cualquier grado de dominancia.

2.2.2.2. Varios loci. El modelo infinitesimal

Con varios loci la fórmula es más compleja. Por ejemplo, con dos loci, a partir de (2.6) tenemos

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2$$

Estas varianzas son difíciles de estimar salvo en los casos de líneas puras en plantas o en animales experimentales.

En el modelo infinitesimal, a partir de las fórmulas (2.7), la varianza genética se puede expresar como

$$\sigma_G^2 = \sigma_{G_1}^2 + \sigma_{G_2}^2 + \dots + \sigma_{G_n}^2$$

ya que las covarianzas entre loci son nulas si la población está en equilibrio Hardy-Weinberg. Se puede demostrar (aunque la demostración no es sencilla, ver CROW &

KIMURA, 1970) que la varianza genética no cambia en el modelo infinitesimal. Los cambios infinitesimales de las frecuencias génicas no conducen a cambios apreciables en la varianza genética.

EJEMPLO 2.3

Tomemos el modelo simple que utilizamos anteriormente, de genes de tipo aditivo ($d=0$) con las mismas frecuencias. La varianza aditiva es, según (2.9) y (2.5),

$$\sigma_A^2 = 2n p q \alpha^2 = 2npq a^2 = 2np(1-p) a^2 = 2n(p-p^2) a^2$$

Un cambio de frecuencias infinitesimal de p a p' produce un cambio en la varianza de

$$\begin{aligned} \Delta\sigma_A^2 &= 2n(p'-p^2)a^2 - 2n(p-p^2)a^2 = 2n a^2 [p' - p - (p^2 - p^2)] = \\ &= 2n a^2 [p' - p - (p' - p)(p' + p)] = 2n a^2 (p' - p)(1 - p' - p) \end{aligned}$$

y como a es un infinitésimo de orden $1/n$, y $p' - p$ es un infinitésimo de orden $1/\sqrt{n}$, el cambio en la varianza es un infinitésimo de orden $(1/n) (1/\sqrt{n}) = n^{-3/2}$

Hay, sin embargo, una fuente adicional de variación. A partir de la fórmula 2.8, y teniendo en cuenta que la varianza aditiva de los hijos es igual a la de los padres y a la de las madres si los apareamientos son al azar; esto es, $\sigma_A^2 = \sigma_{A_H}^2 = \sigma_{A_s}^2 = \sigma_{A_d}^2$, tenemos

$$\sigma_A^2 = \sigma_{A_H}^2 = (1/4)\sigma_{A_s}^2 + (1/4)\sigma_{A_d}^2 + \sigma_{A_e}^2 = (1/4)\sigma_A^2 + (1/4)\sigma_A^2 + \sigma_{A_e}^2$$

$$\sigma_{A_e}^2 = (1/2)\sigma_A^2$$

Sin embargo, si hacemos selección, los valores aditivos de los padres serán más semejantes entre sí que si fueran tomados al azar en la población, y por tanto la varianza aditiva de los padres será menor que la varianza aditiva de la población. La varianza aditiva tras hacer selección será entonces menor que antes de hacer selección. La varianza aditiva se irá reduciendo con el tiempo, lo que contradice la afirmación hecha en el ejemplo 2.3.

Al hacer selección se ha producido un desequilibrio en las frecuencias de los genes, la población ya no está en equilibrio Hardy-Weinberg y los efectos genéticos de cada locus en

la fórmula (2.7) ya no son independientes, aparecen covarianzas entre loci. Por ejemplo, para la varianza aditiva

$$\sigma_A^2 = \sigma_{A_1}^2 + \sigma_{A_2}^2 + \dots + \sigma_{A_n}^2 + 2\text{cov}(A_1, A_2) + 2\text{cov}(A_1, A_3) + \dots$$

y estas covarianzas⁹ tienen que ser necesariamente negativas para que la varianza se reduzca. La selección produce, pues, un cambio de frecuencias génicas, un alejamiento del equilibrio Hardy-Weinberg, y una consecuente reducción de la varianza aditiva incluso en el modelo infinitesimal. Bulmer (1971) llamó la atención sobre este hecho, por lo que se le conoce desde entonces como “efecto Bulmer”.

2.3. Correlaciones entre parientes

Tomemos dos individuos X e Y emparentados en una población en equilibrio Hardy-Weinberg. Su covarianza fenotípica será

$$\begin{aligned} \text{cov}(P_X, P_Y) &= \text{cov}(G_X + E_X, G_Y + E_Y) = \\ &= \text{cov}(G_X, G_Y) + \text{cov}(G_X, E_Y) + \text{cov}(E_X, G_Y) + \text{cov}(E_X, E_Y) \end{aligned}$$

consideraremos que las covarianzas genotipo-medio son nulas, o procuraremos que así lo sean. En ese caso, la covarianza fenotípica es simplemente la suma de la covarianza genética y la ambiental

$$\text{cov}(P_X, P_Y) = \text{cov}(G_X, G_Y) + \text{cov}(E_X, E_Y)$$

Examinaremos a continuación ambas covarianzas

2.3.1. COVARIANZA GENÉTICA ENTRE PARIENTES

⁹ Nótese que estas covarianzas no lo son entre loci del mismo individuo, sino entre loci distintos de los individuos de la población

Tomemos dos individuos X e Y emparentados en una población en equilibrio Hardy-Weinberg, y expresemos el valor genético de cada uno de ellos como

$$G_X = A_X + D_X + I_X \qquad G_Y = A_Y + D_Y + I_Y$$

donde I recoge los efectos epistáticos. La covarianza entre ellos será

$$\text{cov}(G_X, G_Y) = \text{cov}(A_X, A_Y) + \text{cov}(D_X, D_Y) + \text{cov}(I_X, I_Y)$$

ya que los efectos aditivos, dominante y epistático no están relacionados en poblaciones en equilibrio. Como vimos en (2.7), el valor aditivo de uno de estos hijos puede expresarse como la media del valor aditivo de los padres, más un término de aleatorio debido a que no todos los hijos tienen los mismos genes. Expresaremos el valor aditivo de estos individuos como

$$A_X = \frac{1}{2} A_{sX} + \frac{1}{2} A_{dX} + A_{eX} \qquad A_Y = \frac{1}{2} A_{sY} + \frac{1}{2} A_{dY} + A_{eY}$$

donde A_s es el valor aditivo del padre¹⁰, A_d es el de la madre y A_e es el factor aleatorio incorrelacionado con los otros dos e incorrelacionado entre individuos. Consideremos que en los apareamientos se evita cruzar parientes para evitar los problemas que causa la consanguinidad; esto es, consideremos que

$$\text{cov}(A_{sX}, A_{dX}) = \text{cov}(A_{sY}, A_{dY}) = 0$$

la covarianza entre X e Y será,

$$\text{cov}(A_X, A_Y) = \frac{1}{4} \text{cov}(A_{sX}, A_{dY}) + \frac{1}{4} \text{cov}(A_{sX}, A_{dY})$$

En el caso de que X e Y sean medios hermanos (lo que denotaremos¹¹ por HS), como es frecuente en animales de granja, tienen el mismo padre pero no la misma madre, por lo que

$$\text{cov}_{\text{HS}}(A_X, A_Y) = \frac{1}{4} \text{cov}(A_{sX}, A_{sY}) = \frac{1}{4} \text{cov}(A_s, A_s) = \frac{1}{4} \sigma_A^2 \qquad (2.10)$$

ya que en una población que se aparea al azar, la varianza aditiva de los padres es la misma que la de los hijos o la de cualquier conjunto de individuos de la población. Si los

¹⁰ La notación proviene del inglés, s = sire, d= dam.

¹¹ Del inglés half-sibs

individuos fueran hermanos completos (denotado por FS, *full sibs*) , comparten padre y madre

$$\text{cov}_{\text{FS}} (A_X, A_Y) = \frac{1}{4} \text{cov} (A_s, A_d) + \frac{1}{4} \text{cov} (A_d, A_d) = \frac{1}{4} \sigma_A^2 + \frac{1}{4} \sigma_A^2 = \frac{1}{2} \sigma_A^2$$

La covarianza entre padre e hijo es

$$\text{cov} (A_X, A_{sX}) = \text{cov} (\frac{1}{2} A_s, A_s) = \frac{1}{2} \sigma_A^2 \quad (2.11)$$

La covarianza entre nieto y abuelo se construye de la misma forma. El padre tiene un valor aditivo que se puede expresar como

$$A_s = \frac{1}{2} (A_s)_s + \frac{1}{2} (A_s)_d + (A_s)_e$$

y el individuo

$$A_X = \frac{1}{2} A_s + \frac{1}{2} A_d + A_e = \frac{1}{2} [\frac{1}{2} (A_s)_s + \frac{1}{2} (A_s)_d + (A_s)_e] + \frac{1}{2} A_d + A_e$$

$$\text{cov} [A_X, (A_s)_s] = \frac{1}{4} \text{var} [(A_s)_s] = \frac{1}{4} \sigma_A^2$$

De forma similar pueden hallarse otras covarianzas entre parientes. La fórmula general es

$$\text{Cov} (A_X, A_Y) = 2 r_{XY} \sigma_A^2 \quad (2.12)$$

donde r_{XY} es el coeficiente de parentesco entre los individuos; esto es, la probabilidad de que tengan alguno de los alelos de un locus idénticos por descendencia (ver Apéndice 1)¹².

De forma similar se deduce la covarianza entre efectos dominantes. Para que haya dominancia debe haber una interacción entre dos alelos del mismo locus. Para que esta interacción sea la misma en dos parientes, estos alelos tienen que ser los mismos en ambos parientes, y tienen que haber llegado no por azar sino debido precisamente a tener antepasados comunes. La covarianza dominante será, pues

$$\text{cov} (D_X, D_Y) = u_{XY} \sigma_D^2$$

¹² En EE.UU. es corriente usar el coeficiente de correlación genética en lugar del de parentesco: $r_g = \text{cov}(A, A_Y) / V_A$ con lo que la fórmula pasa a ser $\text{cov}(A, A_Y) = r_g V_A$

donde u_{XY} es la probabilidad de que ambos alelos de un locus sean idénticos por descendencia en el individuo X e Y . En nuestros ejemplos anteriores sólo los hermanos completos pueden haber heredado los mismos alelos en el mismo locus.

La covarianza epistática es más difícil de calcular, porque depende de a qué interacciones nos refiramos, las probabilidades de que los alelos estén en un individuo y en un pariente son unas u otras. No siendo importante esta interacción, o más bien siendo ignorada por motivos prácticos debido a las dificultades que se encuentran para su estimación, consideraremos que la deducción de estas interacciones quedan fuera de los límites de este libro.

2.3.2. COVARIANZA AMBIENTAL ENTRE PARIENTES

Normalmente no debe producirse una covarianza ambiental entre parientes, puesto que sus caracteres productivos no son medidos al mismo tiempo, e incluso en el caso de medios hermanos no debe producirse un tratamiento ambiental preferente hacia ciertas familias de medios hermanos. Puede haber covarianza madre-hija si los granjeros dan a las madres muy productivas un tratamiento preferente y también a sus hijas, pero en experimentos bien diseñados estas covarianzas se pueden evitar. Hay sin embargo una notoria excepción, que es el caso de los datos de hermanos, particularmente los que provienen de la misma camada. Los hermanos han compartido el ambiente uterino de la hembra y frecuentemente han compartido también los cuidados maternos y se han alimentado con la leche producida por una misma hembra. A los efectos comunes que producen sobre los caracteres medidos el hecho de haber compartido el mismo ambiente materno se les conoce como "efectos maternos", y generan un parecido entre parientes cuyo origen no es genético; esto es, aparece una covarianza positiva de tipo ambiental entre hermanos. Algunos caracteres serán muy sensibles a los efectos maternos (por ejemplo el peso al destete) y otros no lo serán (por ejemplo, el pH muscular en la canal a la edad de sacrificio). En el caso de que efectivamente el carácter esté afectado por efectos maternos, las covarianzas entre hermanos tienen una componente ambiental común, con lo que la fórmula (2.1) pasa a ser

$$P = m + G + E_M + E_0$$

donde E_M es el efecto del ambiente materno y E_0 el efecto ambiental ejercido sobre el carácter y no dependiente del efecto materno (al que denominaremos “efecto directo”). Si una hembra produce más leche que otra, el peso al destete de sus gazapos será, para el mismo tamaño de camada, mayor; esto es, habrá ejercido un efecto materno positivo sobre sus gazapos. El hecho de que una hembra produzca más leche tiene unas causas genéticas, y también ambientales, por lo que podría decirse que las hembras que tienen genes tales que les hacen producir mucha leche ejercen un efecto materno sobre el peso al destete de sus hijos que tiene unas causas genéticas, *aunque para el hijo sea un efecto estrictamente ambiental sobre el peso al destete*. La fórmula (2.1) pasaría a ser, entonces,

$$E_M = G_1 + E_1$$

$$P = m + G_0 + G_1 + E_1 + E_0$$

donde G_1 es la parte del fenotipo (por ejemplo $G_1 = 20$ gramos en el carácter peso al destete en conejos), que se debe al hecho de que la hembra tenga un buen genotipo para ser buena lechera, tener un buen ambiente uterino, etc; es decir, la parte del efecto materno atribuible a los genes de la madre. Del mismo modo, E_1 es la parte del fenotipo debida a que sobre la madre se ejerció un ambiente que influyó luego sobre el carácter medido. Podría ocurrir que parte de los genes que hacen que la madre sea buena lechera, por ejemplo, tuvieran un efecto pleiotrópico y fueran alelos desfavorables para el peso al destete en los hijos. En ese caso tendríamos una covarianza negativa entre G_1 y G_0 . La varianza fenotípica sería entonces

$$\sigma_P^2 = \sigma_{G_0}^2 + \sigma_{G_1}^2 + 2\text{cov}(G_0, G_1) + \sigma_{E_1}^2 + \sigma_{E_0}^2$$

naturalmente, la descomposición puede seguir, y E_1 ser descompuesta en efectos genéticos y maternos de abuela, etc. Blasco et al. (1982) dan una fórmula general para el parecido entre parientes teniendo en cuenta todo tipo de efectos maternos, pero en la práctica los programas que usan efectos maternos, y que salvo por motivos experimentales son los de vacuno (tanto de carne como de leche), sólo consideran la descomposición en el primer estrato, hasta efectos maternos solamente. Esto ya complica los programas notoriamente, puesto que hacen falta grandes bases de datos para estimar con precisión estas componentes de varianza, y además los métodos de estimación requieren altos costes de computación.

Otro tipo de correlación entre hermanos se presenta, particularmente en plantas, cuando familias de hermanos deben competir por recursos limitados. En esos casos la covarianza entre hermanos puede ser negativa para caracteres como la tasa de crecimiento, puesto que si un individuo acapara más recursos, sus hermanos tendrán menos. Esto ocurre también cuando familias de peces son criadas en el mismo estanque y la cantidad de alimento disponible es constante. En estos casos las correlaciones entre hermanos son de escasa utilidad, puesto que las causas de su parecido o disimilitud no son de tipo genético exclusivamente.

2.3.3. COVARIANZA ENTRE VARIAS MEDIDAS DEL MISMO INDIVIDUO

Hay caracteres que pueden ser registrados más de una vez, por ejemplo el tamaño de camada o los caracteres de producción y calidad de leche en vacuno y ovino lechero. En esos casos el parecido entre varias medidas de un mismo individuo se debe por una parte a que los mismos genes controlan esas medidas, y por otra parte a que hay un ambiente que afecta de forma permanente a esas medidas. Por ejemplo, si una hembra tiene quistes en el útero, todos sus tamaños de camada serán reducidos, y por una causa no genética. La descomposición del valor fenotípico de la ecuación (2.1) pasa a ser

$$P = m + G + E_p + E_e$$

donde E_p es la parte del fenotipo atribuible al *ambiente permanente* sobre el carácter, y E_e es el efecto general del ambiente restante. Dos medidas sobre un individuo se representarían como

$$P_i = m + G + E_p + E_{ei}$$

$$P_k = m + G + E_p + E_{ek}$$

puesto que ambas medidas comparten el mismo efecto genético y el ambiente permanente. La correlación entre varias medidas tomadas sobre el mismo individuo es, teniendo en cuenta la independencia entre efectos genéticos y ambientales, y entre efectos ambientales permanentes y generales,

$$r = \frac{\text{cov}(P_i, P_k)}{\sigma_P \sigma_P} = \frac{\sigma_G^2 + \sigma_{E_p}^2}{\sigma_P^2} = H^2 + \frac{\sigma_{E_p}^2}{\sigma_P^2}$$

y se le conoce como *repetibilidad* de un carácter. No hay que confundirla con una medida de la precisión de la estimación del valor aditivo usada en vacuno de leche, y que tiene el mismo nombre.

2.5. Parámetros genéticos de una población

2.5.1. HEREDABILIDAD DE UN CARÁCTER

Ya hemos visto la definición de heredabilidad de un carácter en sentido amplio en el apartado 2.1.4 (ecuación 2.3). Con miras a la selección, una definición más útil de la heredabilidad debería referirse a la parte de la variación observada que se debe a los valores aditivos, que son al fin y al cabo los que pasan a la descendencia, puesto que las interacciones genéticas no se heredan como ya dijimos antes. Definimos la heredabilidad en sentido estricto como el cociente entre la varianza aditiva y la fenotípica:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}$$

Si un carácter tiene una heredabilidad elevada, la variación que se observa en la población tiene causas genéticas y el ambiente influye poco en el carácter, por lo que un individuo con valores observados elevados será también, por término medio, un individuo con valores genéticos elevados, mientras que si la heredabilidad es baja, un individuo con valores genéticos elevados será difícil de detectar, puesto que en este caso la variación entre los distintos individuos que se observa se debe esencialmente a causas ambientales (figura 2.4).

Figura 2.4. Variabilidad aditiva y ambiental en el caso de heredabilidad alta (a) o baja (b)

OBSERVESE

1) Que la heredabilidad depende del ambiente, luego reduciendo la varianza ambiental aumenta la heredabilidad (por ejemplo, en el caso de animales no haciendo cambios de alimentación, poniendo calefacción en invierno, etc.). Por eso las granjas de mejora son granjas muy bien acondicionadas, y caras de instalar.

2) Que cada población tiene su heredabilidad y sus valores aditivos. En la práctica, sin embargo, las heredabilidades de un mismo carácter no difieren mucho de población a población, en parte porque las instalaciones de las empresas de mejora son buenas y relativamente similares, ya que intentan minimizar en lo posible la variabilidad ambiental.

3) Que usualmente los errores de estimación de la heredabilidad son grandes -salvo usando muchos datos-, por lo que en ocasiones se recurre a usar estimaciones de la heredabilidad proveniente de otras poblaciones. Se corre el riesgo evidente de que la influencia ambiental y la situación genética de sean distintas, aunque como hemos dicho esto no es frecuente, y en cualquier caso éste es un dilema difícil de resolver, puesto que una estima imprecisa puede ocasionar problemas en los programas de selección.

En general, los caracteres reproductivos suelen tener heredabilidades bajas (menos de 0.1), los de crecimiento moderadas (0.2 a 0.4) y los de contenido en carne de la canal altas (0.5 a 0.6).

2.5.2. CORRELACIÓN ENTRE CARACTERES

2.5.2.1. Correlación genética

En ocasiones un mismo grupo de genes influyen sobre dos caracteres simultáneamente. Por ejemplo, los genes que controlan la hormona de crecimiento influyen sobre la altura y el peso de los individuos simultáneamente. A este efecto, conocido como *pleiotropía* se debe el que varios caracteres estén *correlacionados genéticamente*. Por supuesto, cualquier factor ambiental puede producir, además, correlaciones entre caracteres; por ejemplo, variaciones en la alimentación producirán individuos de más o menos peso y también simultáneamente más o menos altos. A esta correlación entre efectos ambientales se le conoce como *correlación ambiental*. Por los mismos motivos prácticos expuestos al hablar de la heredabilidad, estamos más interesados en conocer la correlación entre los efectos genéticos aditivos. Definimos, pues, la correlación genética aditiva entre caracteres como

$$r_A = \frac{\text{cov}(A_1, A_2)}{\sigma_{A_1} \cdot \sigma_{A_2}}$$

donde $\text{cov}(A_1, A_2)$ es la covarianza entre los valores aditivos de cada carácter y σ_{A_1} y σ_{A_2} las correspondientes desviaciones típicas aditivas de los caracteres. El conocimiento de la dirección y de la cuantía de estas correlaciones es importante desde el punto de vista de la selección, pues podemos predecir lo que ocurrirá en otros caracteres como consecuencia de efectuar selección sobre uno de ellos, y podemos integrar la información de varios caracteres para plantear esquemas más eficaces económicamente.

2.5.2.2. Correlación ambiental

La correlación ambiental se define como

$$r_E = \frac{\text{cov}(E_1, E_2)}{\sigma_{E_1} \cdot \sigma_{E_2}}$$

donde $\text{cov}(E_1, E_2)$ es la covarianza entre los valores ambientales de cada carácter y σ_{E_1} y σ_{E_2} las correspondientes desviaciones típicas ambientales de los caracteres.

2.5.2.3. Correlación fenotípica

Finalmente, es también útil conocer las correlaciones entre los valores observados de los caracteres, y se define a la correlación fenotípica como

$$r_P = \frac{\text{cov}(P_1, P_2)}{\sigma_{P_1} \cdot \sigma_{P_2}}$$

donde $\text{cov}(P_1, P_2)$ es la covarianza entre los valores fenotípicos de cada carácter y σ_{P_1} y σ_{P_2} las correspondientes desviaciones típicas ambientales de los caracteres. Aunque es frecuente que el signo de estas tres correlaciones sea el mismo, y también que el valor de la correlación genética sea similar al valor de la correlación fenotípica (GIBSON), no siempre ocurre así, por lo que hay que ser cauto en el caso de caracteres de los que no se disponga de información previa.

Obsérvese que la correlación fenotípica no es la suma de la genética y la ambiental sino que

$$\begin{aligned} r_P &= \frac{\text{cov}(P_X, P_Y)}{\sigma_{P_X} \cdot \sigma_{P_Y}} = \frac{\text{cov}(A_X, A_Y)}{\sigma_{P_X} \cdot \sigma_{P_Y}} + \frac{\text{cov}(E_X, E_Y)}{\sigma_{P_X} \cdot \sigma_{P_Y}} = \\ &= \frac{\text{cov}(A_X, A_Y)}{\sigma_{A_X} \cdot \sigma_{A_Y}} \cdot \frac{\sigma_{A_X} \cdot \sigma_{A_Y}}{\sigma_{P_X} \cdot \sigma_{P_Y}} + \frac{\text{cov}(E_X, E_Y)}{\sigma_{E_X} \cdot \sigma_{E_Y}} \cdot \frac{\sigma_{E_X} \cdot \sigma_{E_Y}}{\sigma_{P_X} \cdot \sigma_{P_Y}} = r_A \cdot h_X \cdot h_Y + r_E \sqrt{(1-h_X^2)(1-h_Y^2)} \end{aligned}$$

OBSERVESE

- 1) No podemos mejorar la correlación genética entre caracteres modificando el ambiente, es un parámetro estrictamente dependiente de la estructura genética de la población
- 2) La suma de la correlación genética y ambiental NO da la correlación fenotípica.
- 3) La estimación de la correlación genética requiere muchos datos, más aún que en el caso de la heredabilidad, por lo que en programas de mejora genética es frecuente usar estimas de otras poblaciones.
- 4) La precisión de la estimación de la correlación genética no depende sólo del tamaño de la muestra, sino de la propia determinación genética del carácter. Si una de las varianzas aditivas es muy pequeña, errores en la estimación de esta varianza pueden producir cambios formidables en la correlación, al estar situadas en el denominador del cociente. Por ejemplo, si

una varianza aditiva vale 0.01 y por error se estima 0.02, la correlación pasa a valer la mitad. Por otra parte, el error de estimación de la correlación genética depende del propio valor de la correlación. Para detectar correlaciones pequeñas hace falta muestras muy grandes, puesto que hay que dirimir si la ligera asociación que se observa se debe al azar o no, cosa que en el caso de correlaciones fuertes es mucho más obvia y se necesitan por tanto muestras mucho más reducidas para detectarla.

2.5.3. ESTIMACIÓN DE LOS PARÁMETROS GENÉTICOS.

2.5.3.1. Estimación de la heredabilidad

Dado que no podemos observar los genes, la única forma de estimar los parámetros genéticos es calcular el parecido entre parientes. Si, por ejemplo, representamos en un gráfico los valores de las madres y de sus hijas para el carácter tamaño de camada en conejos, obtendremos una figura como la Fig. 2.5., en la que cada punto es la pareja formada por el valor del tamaño de camada de una madre y de una de sus hijas, a los que denominaremos (P_M , P_H)

Figura 2.5. Relación entre los valores de madre e hija para tamaño de camada

FIGURA 2.5. CERCA DE AQUI

La pendiente de esa recta es positiva, por lo que parece que por término medio las mejores madres dan lugar a mejores hijas, aunque la pendiente es tan suave que no parece que los valores elevados de las madres se vayan a traducir necesariamente en valores elevados en las hijas. La pendiente de esa recta es

$$b = \frac{\text{cov}(P_M, P_H)}{\sigma_{P_M}^2} = \frac{(1/2)\sigma_A^2}{\sigma_P^2} = \frac{h^2}{2}$$

la covarianza entre los valores fenotípicos de madre e hija es la mitad de la varianza aditiva, como vimos en la fórmula (2.11). Por otra parte, si la población no está seleccionada, la varianza fenotípica de las madres es la varianza fenotípica de la población, puesto que las madres son una muestra aleatoria de la misma.

Lo mismo podríamos haber hecho realizando la regresión con cualquier otro tipo de parientes. Podríamos también haber usado la correlación entre parientes en lugar de la regresión; por ejemplo, la correlación entre medios hermanos, usando la fórmula (2.10)

$$r_{HS} = \frac{\text{COV}_{HS}}{\sigma_P^2} = \frac{(1/4)\sigma_A^2}{\sigma_P^2} = \frac{h^2}{4}$$

Podría haber servido también, por ejemplo, la correlación entre hermanos si no hubiera efectos maternos. Los métodos modernos de estimación de los parámetros genéticos utilizan todas las relaciones entre parientes existentes en la población, ponderándolas adecuadamente y corrigiendo simultáneamente los datos por efectos ambientales que puedan haber perturbado los resultados (por ejemplo, ciertos individuos pueden estar medidos en verano y ciertos en invierno, con lo que su crecimiento se ve afectado por la estación).

2.5.3.2. Estimación de las correlaciones genéticas

Para estimar las correlaciones genéticas entre dos caracteres, dispondremos a los parientes de la misma forma, pero midiendo un carácter en un pariente y el otro carácter en el otro pariente. Por ejemplo, dispongamos en el eje X los valores de cantidad de leche producida por un conjunto de vacas, y en eje Y los valores de porcentaje de grasa producido por medias hermanas de esas vacas. Cada punto de la figura 2.6. es el valor observado para leche de una vaca y el de porcentaje de grasa de una media hermana suya (P_L , P_G).

Figura 2.6. Relación entre la producción de leche de unas vacas y la de grasa de sus medias hermanas

FIGURA 2.6. CERCA DE AQUI

Como ocurría en la ecuación (2.4), lo único en común que comparten dos medias hermanas es la mitad del valor aditivo del padre, por lo que

$$\text{cov}(P_L, P_G) = \text{cov}\left(\frac{1}{2}A_L, \frac{1}{2}A_G\right) = \frac{1}{4}\text{cov}(A_L, A_G)$$

con lo que la pendiente de la recta será

$$b = \frac{\text{cov}(P_L, P_G)}{\sigma_{P_G}^2} = \frac{(1/4)\text{cov}(A_L, A_G)}{\sigma_{A_L}\sigma_{A_G}} \cdot \frac{\sigma_{A_L}\sigma_{A_G}}{\sigma_{P_G}^2} = r_A \cdot \frac{\sigma_{A_L}\sigma_{A_G}}{4\sigma_{P_G}^2}$$

dado que podemos estimar la varianza fenotípica de la cantidad de grasa con los datos que tenemos, y que podemos estimar las desviaciones típicas aditivas a partir de las varianzas aditivas como explicamos en el apartado anterior, disponemos de una estimación de la correlación genética aditiva entre grasa y leche.

Como en el caso de la heredabilidad, podríamos haber utilizado cualquier otra pareja de parientes para estimar la correlación genética, y como en el caso de la heredabilidad también, los métodos modernos de estimación de correlaciones genéticas utilizan todas las relaciones de parentesco existentes en la población ponderándolas adecuadamente y corrigiendo los datos simultáneamente por efectos ambientales sistemáticos que pueden sesgar los resultados.

La estimación de parámetros genéticos requiere ciertas hipótesis que no siempre se cumplen, por ejemplo el estado de equilibrio de la población, o el que la población sea cerrada (hoy en día casi ninguna población de una empresa de mejora genética lo es). Si la población que se analiza está seleccionada, aparece desequilibrio gamético que conduce a una reducción de la varianza. Si la población no está cerrada, la entrada de nuevos genes produce no sólo este desequilibrio, sino la dificultad de referir los parámetros genéticos a una población determinada. En la práctica es difícil encontrar poblaciones de plantas o animales de interés comercial que no hayan estado sometidas a selección, o en las que no migren genes para hacerlas más productivas.

2.5.3.3. Uso de modelos lineales en la estimación

Los datos están sometidos a efectos ambientales sistemáticos que alteran grupos de valores. Por ejemplo, en invierno los animales crecen más que en verano porque el calor disminuye el apetito. Las camadas de primer parto suelen tener por término medio menos individuos. Hay granjas que cuidan mejor a los animales y estos producen más. Los datos pueden precoregirse antes de efectuar los análisis, y referir los valores de los animales a la media de la granja o de la estación, pero podría ocurrir, por ejemplo, que las mejores granjas invirtieran más en semen de calidad, con lo que al precoregir los datos y referirlos a la media de la granja, estaríamos descontando también un efecto genético. Además sería conveniente tener en cuenta la totalidad de los datos y de las relaciones parentales al hacer la estimación, no sólo las correlaciones entre hermanos o la regresión padre-hijo. La forma correcta de considerar este problema la exponemos a continuación.

La descomposición del valor fenotípico de la fórmula (2.1)

$$P = m + G + E$$

puede representarse como

$$y = m + g + e$$

donde y es el vector de fenotipos, m es el vector de medias de los datos, g es el efecto genético, e es el efecto del ambiente de tipo aleatorio, con media cero. Para recoger los efectos ambientales de tipo sistemático, varios grupos de datos tienen la misma media (por ejemplo, los que nacieron en la misma estación, o los de la misma granja). El modelo se representa entonces como

$$y = Xb + Zu + e$$

donde b contiene las medias comunes a varios individuos (los efectos de estación, por ejemplo), u los efectos genéticos (nos interesamos, como antes sólo en los de tipo aditivo), y X y Z son matrices de incidencia; esto es, de unos y ceros indicando la presencia o ausencia de un efecto para un individuo concreto. En el caso de que haya covariables, una columna de X contiene los valores de la covariable para cada individuo. A los efectos contenidos en b se les llama “fijos” y a los contenidos en u “aleatorios”, una distinción que algunos autores como Fisher o Yates consideran innecesaria. En el próximo capítulo abordaremos la diferente forma de estimar estos efectos.

EJEMPLO 3.12

En la tabla siguiente se indican los datos de tamaño de camada obtenidos por dos conejas en dos estaciones distintas, y el peso de ambas

	CONEJA 1		CONEJA 2	
	INVIERNO	PRIMAVERA	INVIERNO	PRIMAVERA
PARTO 1	12		9	
PARTO 2		7		11
PARTO 3				8

Llamando E_1 y E_2 a los efectos de estación, P_1 , P_2 , P_3 a los efectos de parto,

$$\left. \begin{array}{l} 12 = E_1 + P_1 + u_1 + e_1 \\ 7 = E_2 + P_2 + u_1 + e_2 \\ 9 = E_1 + P_1 + u_2 + e_3 \\ 11 = E_2 + P_2 + u_2 + e_4 \\ 8 = E_2 + P_3 + u_2 + e_5 \end{array} \right\} \begin{array}{l} \left[\begin{array}{c} 12 \\ 7 \\ 9 \\ 11 \\ 8 \end{array} \right] = \left[\begin{array}{ccccc} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array} \right] \cdot \left[\begin{array}{c} E_1 \\ E_2 \\ P_1 \\ P_2 \\ P_3 \end{array} \right] + \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array} \right] \cdot \left[\begin{array}{c} u_1 \\ u_2 \end{array} \right] + \left[\begin{array}{c} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \right] \\ \mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{Z} \mathbf{u} + \mathbf{e} \end{array}$$

En este modelo, por conveniencia de cálculo, los efectos aleatorios se refieren a la media; esto es, la media de los efectos genéticos aditivos es cero.

$$E(\mathbf{y}) = \mathbf{X}\mathbf{b} \quad ; \quad E(\mathbf{u}) = \mathbf{0} \quad ; \quad \text{var}(\mathbf{u}) = \mathbf{G} \quad ; \quad \text{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}' \text{var}(\mathbf{u}) \mathbf{Z} + \text{var}(\mathbf{e}) = \mathbf{Z}' \mathbf{G} \mathbf{Z} + \mathbf{I}\sigma_e^2$$

Si ignoramos la dominancia y la epistasia, la matriz \mathbf{G} es la matriz de correlaciones genéticas entre todos los valores aditivos, y se calcula utilizando la fórmula (2.12). Esta matriz suele expresarse como $\mathbf{G} = \mathbf{A}\sigma_A^2$ donde \mathbf{A} recoge, de acuerdo a la fórmula (2.12), el doble de los coeficientes de parentesco entre los individuos. La varianza se representa entonces como

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}' \mathbf{A} \mathbf{Z} \sigma_A^2 + \mathbf{I}\sigma_e^2$$

y el objetivo es entonces estimar las varianzas σ_A^2 y σ_e^2 . Frecuentemente se usa σ_u^2 en estos modelos para representar a la varianza aditiva. Modelos análogos se utilizan para la

estimación de covarianzas genéticas entre caracteres, y no entraremos en el detalle de los mismos. Del mismo modo, estos modelos pueden ampliarse para contener efectos maternos o efectos ambientales permanentes si se dispone de varios datos de un individuo. El lector interesado puede consultar textos como los de **RICO (2000) o MRODE (FECHA)**.

2.5.3.4. La estimación por máxima verosimilitud y máxima verosimilitud residual

La mayor parte de poblaciones comerciales están seleccionadas, por lo que no se cumplen las condiciones de equilibrio requeridas en los modelos anteriores, y la mayor parte de procedimientos da lugar a estimas sesgadas de las componentes de varianza. Sin embargo, si se usan los procedimientos de máxima verosimilitud o los procedimientos bayesianos que expondremos a continuación, la selección puede ser ignorada siempre que

- 1) Se incluyan en el análisis todos los datos usados para seleccionar
- 2) Se incluya la matriz de parentesco completa hasta la generación base

En ese caso se puede proceder como si la selección no hubiera actuado. Las razones de ello se exponen en el apéndice IV. Seguidamente pasaremos a exponer los métodos de máxima verosimilitud y los métodos bayesianos.

El concepto de verosimilitud y el método de máxima verosimilitud fueron formalizados por Fisher entre 1912 y 1922. En 1912 la teoría de la estimación aún no estaba desarrollada, por lo que no se le podían atribuir al método propiedades particularmente interesantes y el artículo quedó prácticamente ignorado. En 1922 Fisher publica un artículo en el que se definen las propiedades que debe reunir un buen estimador, y se encuentra que el método de máxima verosimilitud produce estimadores con buenas propiedades, por lo que es entonces cuando se produce su aceptación en la comunidad científica (FISHER 1912, 1922). En el apéndice II se explica con detalle el concepto de verosimilitud y el método de máxima verosimilitud.

El método de máxima verosimilitud ha encontrado su aplicación más frecuente en mejora genética animal en la estimación de componentes de varianza, singularmente debido a una variante del procedimiento, la máxima verosimilitud residual o restringida (REML). HARTLEY Y RAO (1967) propusieron obtener estimas de componentes de varianza utilizando el método de máxima verosimilitud (ML). El procedimiento da lugar a ecuaciones bastante complicadas que deben resolverse de forma iterativa, por lo que dificultades de cómputo impidieron su utilización rutinaria en programas de mejora. Surgió, además, alguna dificultad conceptual: al resolver modelos mixtos, ML estima componentes de varianza como si la estima de los efectos fijos se hubiera realizado sin error; esto es, sin tener en cuenta los grados de libertad perdidos al

estimar los efectos fijos, lo que podría ser preocupante cuando se estiman muchos efectos fijos, como es el caso del vacuno de leche. Por estas razón se propuso realizar un cambio de sistema de coordenadas y proyectar los datos en un subespacio en el que no hubieran efectos fijos, maximizando la verosimilitud en este subespacio. A este procedimiento se le llamó máxima verosimilitud restringida (REML) y fue generalizado por PATTERSON Y THOMPSON en 1971. El REML presenta, además, la ventaja -más bien de tipo psicológico- de que sus estimas coinciden con las del ANOVA cuando los datos estaban equilibrados. Las estimas ML en diseños equilibrados dan resultados diferentes a las del ANOVA -que es óptimo para esos diseños-, lo que resulta un tanto inquietante. Además, como veremos a continuación, las estimas ML no tienen en cuenta pérdidas de grados de libertad debido a la estimación de los efectos fijos, lo que resulta también inquietante.

Para examinar mejor las diferencias entre máxima verosimilitud y REML, estimaremos la varianza por ambos métodos utilizando un modelo sencillo.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{e} = \mathbf{1}\mu + \mathbf{e}$$

donde $\mathbf{X} = \mathbf{1}$ es un vector de unos. La matriz de varianzas-covarianzas de los errores es $\mathbf{V} = I\sigma^2$, y su determinante $|\mathbf{V}| = (\sigma^2)^n$. La función de verosimilitud, en el caso de que el carácter se distribuya de forma normal $N(\boldsymbol{\mu}, \sigma^2)$ y la muestra tenga n datos, es

$$L(\sigma^2 | \mathbf{y}) = \text{cte} \cdot \sigma^{-2(n/2)} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{1}\mu)'(\mathbf{y} - \mathbf{1}\mu)\right]$$

Si derivamos la expresión (o su logaritmo, que resulta más sencillo) e igualamos a cero, obtendremos el valor que hace máximo a $L(\sigma^2 | \mathbf{y})$, y que es

$$\hat{\sigma}^2 = (1/n)(\mathbf{y} - \mathbf{1}\mu)'(\mathbf{y} - \mathbf{1}\mu) = (1/n)\sum (y_i - \mu)^2$$

Obsérvese que debemos conocer μ para obtener la estima de la varianza. Como no conocemos μ lo sustituimos por la estima máximo verosímil de μ

$$\hat{\mu} = (1/n)\sum y_i$$

con lo que la estima ML de la varianza es

$$\hat{\sigma}_{ML}^2 = (1/n)(\mathbf{y} - \mathbf{1}\hat{\mu})'(\mathbf{y} - \mathbf{1}\hat{\mu})$$

Esta estima, a pesar de ser función de otra, sigue teniendo las mismas buenas propiedades asintóticas que todas las estimas de máxima verosimilitud, y no hay razón formal para rechazarla.

Para calcular las estimas REML se proyectan los datos en un subespacio sin efectos fijos. Si la matriz de proyección es \mathbf{K} , el método consiste en hacer

$$\mathbf{K}'\mathbf{y} = \mathbf{K}'\mathbf{1}\mu + \mathbf{K}'\mathbf{e} = \mathbf{K}'\mathbf{e}$$

de forma que $\mathbf{K}'\mathbf{1} = \mathbf{0}$. Por ejemplo, la matriz

$$\mathbf{K}' = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & \dots & 0 \\ 1 & 0 & 0 & -1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & \dots & -1 \end{bmatrix}$$

cumple la condición.

La varianza de $\mathbf{K}'\mathbf{y}$ es $\mathbf{K}'\mathbf{V}\mathbf{K} = \mathbf{K}'\mathbf{K}\sigma^2$. La verosimilitud es, ahora

$$L(\sigma^2 | \mathbf{K}'\mathbf{y}) = \text{cte} \cdot |\mathbf{K}'\mathbf{K}\sigma^2|^{1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{K}'\mathbf{y})'(\mathbf{K}'\mathbf{y})\right]$$

Se deduce inmediatamente que

$$|\mathbf{K}'\mathbf{K}\sigma^2| = n(\sigma^2)^{n-1}$$

con lo que se puede obtener como antes el valor que maximiza la verosimilitud, y que ahora es

$$\hat{\sigma}_{REML}^2 = [1/(n-1)] \mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}$$

es sencillo ver en nuestro ejemplo que

$$\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{y} = \left[\mathbf{y} - \mathbf{1}\left(\frac{1}{n}\sum y_i\right) \right]' \left[\mathbf{y} - \mathbf{1}\left(\frac{1}{n}\sum y_i\right) \right] = (\mathbf{y} - \mathbf{1}\hat{\mu})'(\mathbf{y} - \mathbf{1}\hat{\mu})$$

y la estima REML de la varianza es

$$\hat{\sigma}_{REML}^2 = \left[1/(n-1) \right] (\mathbf{y} - \mathbf{1}\hat{\mu})' (\mathbf{y} - \mathbf{1}\hat{\mu})$$

que es idéntica a la estima ML, pero dividiendo por $(n-1)$ en vez de por n . A pesar de la similitud de las fórmulas hay que hacer notar que, al contrario que en la estima ML, no se sustituye el valor verdadero de μ por el estimado, sino que al deducir la fórmula del estimador de la varianza aparece una expresión, $(1/n) \sum y_i$, que coincide con la estima máximo verosímil de μ . El hecho de que se ha tenido en cuenta el grado de libertad perdido al estimar μ , se refleja en que se divide por $(n-1)$ en lugar de por n . Cuando hay muchos efectos fijos esta distinción es notable, puesto que la estima máximo verosímil es

$$\hat{\sigma}_{ML}^2 = \left(\frac{1}{n} \right) (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})' (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

mientras que la estima REML es

$$\hat{\sigma}_{REML}^2 = \left(\frac{1}{n - \text{rg}(\mathbf{X})} \right) (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})' (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

donde $\text{rg}(\mathbf{X})$ es el rango de \mathbf{X} y $\hat{\mathbf{b}}$ la estima máximo verosímil de \mathbf{b} .

Con la matriz \mathbf{K} propuesta, el análisis de la varianza se hace sobre un vector

$$(\mathbf{K}'\mathbf{y})' = [y_1 - y_2, y_1 - y_3, \dots, y_1 - y_n]$$

la idea es que una forma de hacer desaparecer la media es hacer el análisis sobre diferencias de datos en lugar de sobre los propios datos, puesto que

$$y_i - y_j = (\mu + e_i) - (\mu + e_j) = e_i - e_j$$

donde $\mathbf{1}$ ha desaparecido. Hay varias matrices que cumplen la condición $\mathbf{K}'\mathbf{1} = \mathbf{0}$. El análisis se podría hacer también sobre

$$(\mathbf{K}'\mathbf{y})' = [y_1 - y_2, y_2 - y_3, \dots, y_{n-1} - y_n]$$

con el mismo resultado. Esto no quiere decir que cualquier \mathbf{K} valga. Una matriz con la mitad de sus filas compuestas por ceros también cumple $\mathbf{K}'\mathbf{1} = \mathbf{0}$, y también lo cumple la matriz $\mathbf{K} = \mathbf{0}$. Se trata de encontrar matrices que no hagan perder información relativa a la dispersión, lo que se consigue introduciendo en \mathbf{K} el máximo número de contrastes lineales independientes. De hecho no importa la \mathbf{K} concreta, siempre que se utilice el máximo número de contrastes lineales independientes¹³. Obsérvese que \mathbf{K} es de dimensiones $(n-1) \times n$ puesto que sólo hay $n-1$ parejas de diferencias: perdemos un grado de libertad, o lo que es lo mismo, utilizamos la información algo resumida: el nuevo vector de datos $\mathbf{K}'\mathbf{y}$ es un vector de $n-1$ elementos. Si tuviéramos que hacer una representación geométrica de nuestra muestra, necesitaríamos un espacio de n dimensiones, sin embargo, al utilizar REML deberíamos representar nuestros datos en un espacio de $n-1$ dimensiones, nos movemos en un espacio más restringido, hemos perdido un *grado de libertad*. Cuando se estiman muchos efectos fijos, esta pérdida es más notable.

No hay un argumento claro para preferir REML a ML. En el ejemplo anterior la estima REML es insesgada pero tiene un riesgo mayor que la ML para todos los valores posibles de la varianza, pero esto puede no ocurrir en casos más complejos. En otras situaciones el riesgo de la estimación REML será mayor o menor que la de ML dependiendo de los valores verdaderos de las componentes de varianza y de la estructura de los datos. La elección que en mejora animal se ha hecho hacia el método REML está relacionada más con argumentos indirectos como los expuestos antes que con una razón basada en el riesgo del estimador o sus propiedades. El argumento de que las estimas REML coinciden con las de ANOVA en diseños equilibrados es escasamente convincente en mejora animal, en donde los diseños están siempre sometidos a fuertes desequilibrios.

2.5.3.5. La estimación Bayesiana

¹³ Se puede demostrar con carácter general que

$$\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y} = [\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}]' \mathbf{V}^{-1} [\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}]$$

con lo que \mathbf{K} no aparece ya en la fórmula

En los últimos años se ha empezado a popularizar la estimación bayesiana de parámetros genéticos, debido a ciertas ventajas que la estadística bayesiana tiene sobre la estadística clásica. Queda fuera del alcance del presente texto el hacer una exposición detallada de las ventajas e inconvenientes de ambos paradigmas estadísticos; el lector interesado en una visión comparativa de la aplicación de ambos paradigmas a la mejora genética puede consultar la revisión de BLASCO (2001). En el apéndice III se ofrecen los principios de la teoría bayesiana. Aquí recogemos su aplicación a la estimación de componentes de varianza. Llamaremos σ_u^2 a la varianza aditiva, como es uso habitual en la literatura de modelos lineales.

La forma bayesiana de abordar el problema consiste en estimar las funciones de densidad de probabilidad posterior $f(\sigma_u^2 | \mathbf{y})$ y $f(\sigma_e^2 | \mathbf{y})$, o también $f(h^2 | \mathbf{y})$. Estas funciones se estiman a partir de la función de densidad de probabilidad conjunta $f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y})$ integrando las partes que no interesan; por definición de función de densidad de probabilidad, tenemos que

$$f(\sigma_u^2 | \mathbf{y}) = \int \cdots \int f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y}) d\mathbf{b} d\mathbf{u} d\sigma_e^2$$

donde, por el teorema de Bayes,

$$f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y}) = f(\mathbf{y} | \mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2) f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2) / f(\mathbf{y})$$

En principio conocemos $f(\mathbf{y} | \mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2)$, puesto que los datos se distribuyen de forma Normal, dados los valores de $\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2$. El primer problema aparece al considerar las densidades de probabilidad *a priori* $f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2)$. Hay una considerable discusión en torno a cómo representarlás, pero es frecuente considerar que estos parámetros son independientes, por lo que

$$f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2) = f(\mathbf{b}) \cdot f(\mathbf{u}) \cdot f(\sigma_u^2) \cdot f(\sigma_e^2)$$

para apelar seguidamente al “principio de indiferencia” y considerar $f(\mathbf{b}) = \text{constante}$, y a menudo también $f(\sigma_u^2) = \text{constante}$ y $f(\sigma_e^2) = \text{constante}$. En cuanto a $f(\mathbf{u})$ se le considera distribuida de forma Normal con media cero y varianza $\mathbf{A}\sigma_u^2$ como ya vimos. Finalmente, $f(\mathbf{y})$

es la probabilidad de la muestra, una constante que se obtiene integrando todos los parámetros.

El siguiente problema es resolver las integrales. Hasta hace pocos años este era uno de los puntos débiles más notorios de la aproximación bayesiana: era teóricamente posible, pero prácticamente no era factible resolver estas integrales. La aparición de técnicas de muestreo aleatorio en Cadenas de Markov (lo que se conoce como MCMC: Monte Carlo Markov Chains) permitió aproximar estas integrales, de forma que hoy en día es factible extraer muestras aleatorias de esas funciones de densidad. Mediante estas técnicas se puede obtener una matriz

$$\begin{bmatrix} b_{11} & b_{21} & \dots & u_{11} & u_{21} & \dots & \sigma_{u1}^2 & \sigma_{e1}^2 \\ \dots & \dots \\ b_{1i} & b_{2i} & \dots & u_{1i} & u_{2i} & \dots & \sigma_{ui}^2 & \sigma_{ei}^2 \\ \dots & \dots \end{bmatrix}$$

cuyas filas son puntos extraídos aleatoriamente de la función multivariante $f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y})$. Por tanto, la columna $(b_{11}, \dots, b_{1i}, \dots)$ es un conjunto de puntos de $f(b_1 | \mathbf{y})$, y lo mismo se puede decir de las otras columnas; así, la columna $(\sigma_{u1}^2, \dots, \sigma_{ui}^2, \dots)$ es una muestra de la función de densidad de probabilidad $f(\sigma_u^2 | \mathbf{y})$, y puede ser usada para hacer inferencias sobre σ_u^2 . Para hacer inferencias sobre la heredabilidad, creamos una nueva columna a partir de los valores de σ_u^2 y σ_e^2 que hay en cada fila. El conjunto de puntos $[\sigma_{u1}^2 / (\sigma_{u1}^2 + \sigma_{e1}^2), \dots, \sigma_{ui}^2 / (\sigma_{ui}^2 + \sigma_{ei}^2), \dots]$ es una muestra aleatoria de la función de densidad de probabilidad $f(h^2 | \mathbf{y})$. Como el número de puntos que muestreamos aleatoriamente es arbitrario, se pueden obtener histogramas muy precisos de las funciones de densidad de probabilidad que queremos estimar.

Las inferencias se hacen a partir de los puntos muestreados de la densidad posterior multivariante. Por ejemplo, supongamos que hemos muestreado 5.000 puntos de una densidad posterior conjunta y dispongo, pues, de 5.000 puntos de la función de densidad de probabilidad de la heredabilidad. Hallando la media de valores de heredabilidad obtenidos, tengo una estima de la media de la densidad posterior; creando un histograma o dibujando la función de densidad a partir de esos 5.000 valores puedo obtener una estima de la moda, y ordenándolos puedo obtener la mediana con facilidad

2. 5.3.6. Los errores de estimación

Se puede obtener, además, una medida de la precisión de la estimación derivada de la propia función de verosimilitud. Una medida de la precisión se obtiene examinando la forma de la curva de verosimilitud; si esta es apuntada, hay razones para considerar a su máximo como un valor fiable; si la función es más bien aplanada, su valor máximo es tan *verosímil* como otros valores que difieren notablemente de él. Para medir la precisión Fisher introdujo el concepto de *cantidad de información*. Si θ es el parámetro que deseamos estimar, $d \log(L(\theta|y))/d\theta$ es la pendiente de la curva del logaritmo de su verosimilitud. Este valor, positivo o negativo, será elevado si la función de verosimilitud es apuntada, y será pequeño si es plana. Elevando al cuadrado para evitar problemas de signo y hallando la media de todos estos valores, esta media será un valor elevado si la función es apuntada y bajo en caso de que sea plana. En el entorno del máximo, si la función es apuntada ello quiere decir que el grado de credibilidad de los puntos alejados del máximo es pequeño; si por el contrario la función es muy plana, puntos distantes tendrán un grado de credibilidad similar. Esta cantidad fue llamada por Fisher cantidad de información intrínseca a los datos, probablemente porque cuanto mayor sea la cantidad de información, más apuntada es la curva de verosimilitud y menos verosímiles son las estimas alejadas del máximo. En muestras pequeñas, además, la función de verosimilitud puede presentar máximos locales o ser asimétrica en torno al máximo, con lo que el "grado de creencia racional" que la verosimilitud proporciona no es el mismo a un lado del máximo que al otro lado. El motivo de usar logaritmos es facilitar el que la cantidad de información sea aditiva. Si los n individuos de una muestra son independientes, la verosimilitud es

$$L(\theta|x_1, \dots, x_n) = L(\theta|x_1) \cdot L(\theta|x_2) \cdot \dots \cdot L(\theta|x_n),$$

$$\text{por tanto } \log L(\theta|x_1, \dots, x_n) = \log L(\theta|x_1) + \log L(\theta|x_2) + \dots + \log L(\theta|x_n)$$

La escuela frecuentista ha reducido el problema a encontrar el estimador máximo verosímil debido a que éste tiene buenas propiedades desde el punto de vista frecuentista, la principal de las cuales es que converge a una distribución Normal cuya media es el valor verdadero del parámetro que se quiere estimar, y su varianza la inversa de la cantidad de información aplicada en el punto máximo (ver, p. ej., Stuart y Ord, 1991, pp. 659 y 660). El principal problema de esta medida de la precisión reside en que depende del tamaño de la muestra, puesto que se basa en una aplicación del teorema central del límite, por lo que no está claro qué sucede en muestras pequeñas, ni tampoco cuál es el tamaño muestral a partir del cual estos problemas desaparecen.

Desde un punto de vista bayesiano, como se dispone de la función de densidad posterior, puede calcularse cualquier intervalo de confianza (los bayesianos prefieren denominarlos intervalos de credibilidad); por ejemplo, si disponemos de las muestras producidas por un proceso MCMC puede verse qué proporción de puntos dan valores mayores que un cierto valor, qué proporción de puntos están entre dos valores dados, o bien cuál es el menor intervalo que contiene al 95% de los puntos, a partir de qué valor se encuentra el 95% superior de los puntos, etc. Todas estas proporciones dan lugar a inferencias: probabilidad de que el valor de la heredabilidad sea mayor que un cierto valor, probabilidad de que la heredabilidad se encuentre entre dos valores dados, la heredabilidad se encuentra entre estos dos valores con un 95% de probabilidad, la heredabilidad es igual o mayor que este valor con un 95% de probabilidad, etc.

SOFTWARE DISPONIBLE

Hay varios programas públicos que permiten estimar componentes de varianza por métodos de máxima verosimilitud y REML. Los más conocidos son el VCE de GROENEVELD (CITA, Fecha), el DFREML de MEYER (CITA, Fecha) y el ASREML de McGUIRCK (CITA, fecha)

AÑADIR DIRECCIONES WEB

Bibliografía recomendada

FALCONER

BULMER

RICO

OLLIVIER

Referencias

Cuestiones

Cuestiones

1. Se dispone de datos peso del fruto de una línea pura de tomate. Estas líneas puras son completamente homocigotas, tienen sus dos cromátidas idénticas, por lo tanto los descendientes son idénticos a los padres. ¿Cuál es la heredabilidad, en sentido estricto, del peso del fruto en esa población?

A) 1 ; B) 0 ; C) No tiene heredabilidad ; D) 0.5 ; E) No se puede contestar a esa pregunta sin conocer los datos.

2. ¿Es posible obtener una estima negativa de la heredabilidad de un carácter?

3. Si el carácter A está correlacionado positivamente con el B, y el B lo está con el C, ¿implica esto que necesariamente el A lo está con el C?

4. ¿Es posible que dos caracteres presenten una correlación genética positiva y sin embargo la correlación ambiental sea negativa?

5. ¿Es posible aumentar la heredabilidad de un carácter?

7. Definición de heredabilidad: Una heredabilidad de 0.20 significa que

A) Hay un 20% de probabilidad de heredar el carácter

B) El valor aditivo es el 20% del valor fenotípico

C) Un 20% del carácter es heredable

D) La varianza aditiva es un 20% de la varianza fenotípica

E) En un 20% de los casos la varianza aditiva pasará a la descendencia.

8. Valor aditivo: El toro Y tiene la siguiente valoración del valor aditivo de él mismo y de sus antecesores, en kg de leche al año: BISABUELO: +100, ABUELO: +100, PADRE: +100, TORO Y: +200.

El toro X tiene la siguiente valoración del valor aditivo de él mismo y de sus antecesores, en kg de leche al año: BISABUELO: +200, ABUELO: +200, PADRE: +200, TORO Y: +100.

¿Qué toro es el más adecuado como reproductor?

A) el X

B) el Y

C) La información suministrada no es útil para tomar la decisión

D) Ambos están igualmente valorados como reproductores

E) Ninguno de los dos es adecuado

9. Covarianza genotipo-medio: ¿En qué consiste la covarianza genotipo-medio?

A) Los mejores genotipos (por ejemplo, las mejores vacas) reciben el mejor ambiente (por ejemplo, la mejor alimentación).

B) Los mejores genotipos en un ambiente (por ejemplo, clima templado) no son los mejores genotipos en otro ambiente (por ejemplo, clima caluroso).

- C) El medio (por ejemplo, una buena alimentación) produce individuos que dan lugar a hijos genéticamente mejores
- D) El genotipo determina las condiciones del medio
- E) En animales no hay interacciones genotipo-medio

10. Valor aditivo: El valor aditivo de un toro que tiene cinco hijas probadas es

- A) La media de la producción de las cinco hijas probadas
- B) El valor de los genes que ha transmitido por igual a cada una de sus hijas
- C) La adición de los valores económicos de cada carácter objeto de selección
- D) La media de la producción de todas su hijas posibles
- E) El conjunto de los genes del animal que se heredan

Referencias

MENDEL G. 1865. Experimentos en Híbridos de plantas. Artículo aparecido originalmente en *Verhandlungen des Naturforschenden Vereines*. Brünn, 4. Abhand., pp. 3-47. Reimpreso en *The origin of genetics* C. Stern y E.R. Sherwood, Eds. Freeman & Co. San Francisco, 1966. Traducción española en *El origen de la genética*. Alhambra. Madrid, 1973.

STERN C., SHERWOOD E.R. (Eds.) 1966. *The origin of genetics*. Freeman & Co. San Francisco, 1966. Traducción española en *El origen de la genética*. Alhambra. Madrid, 1973.

Apéndice I. Coeficiente de parentesco

Para estudiar los apareamientos en los que intervienen individuos emparentados es conveniente disponer de alguna medida del parentesco. Una forma de medirlo puede ser comparar un alelo de cada uno de los dos individuos que se aparean y ver si son iguales por haber sido transmitidos por un antecesor común. Cuando los individuos están muy emparentados, es más probable que los dos alelos sean iguales.

Algunos de los hijos tendrán sus dos alelos iguales debido a que provienen ambos de un antecesor común -el abuelo o la abuela-, y otros hijos los tendrán iguales por azar. Para medir el parentesco nos interesan los alelos que provienen de un antecesor común, puesto que cuanto más probable sea encontrarlos esto indicará que los individuos están más emparentados.

Pondremos un ejemplo de cómo se calculan los coeficientes de consanguinidad y parentesco de forma recurrente. Si tenemos dos individuos X e Y, llamaremos a los alelos de uno de los locus de X (A,a), y a los alelos del mismo locus de Y (A',a'). Al escoger un alelo al azar de X pudiera éste ser **A**, y al escoger uno de Y pudiera ser **a'**, y pudiera ocurrir que **A** fuera idéntico -por descendencia, por supuesto- a **a'**. También pudiera ocurrir que el alelo escogido de Y fuera **A'**; en ese caso habría que examinar la probabilidad de que **A=A'**. Si consideramos todos los casos posibles, la probabilidad de que al escoger un alelo de X y uno de Y sean iguales es

$$r_{XX} = \frac{1}{4} [P(\mathbf{A}=\mathbf{A}') + P(\mathbf{A}=\mathbf{a}') + P(\mathbf{a}=\mathbf{A}') + P(\mathbf{a}=\mathbf{a}')]$$

donde P significa probabilidad. Esto es, por definición, el coeficiente de parentesco r_{XY} entre los individuos X e Y, o el coeficiente de consanguinidad **F** de cualquiera de sus hijos.

Finalmente, si sabemos que el individuo X es hijo de A y de B, y que el individuo Y es hijo de C y de D, es fácil demostrar que

$$r_{XY} = r_{AxB,CxD} = \frac{1}{4} (r_{AC} + r_{AD} + r_{BC} + r_{BD})$$

lo que permite calcular coeficientes de parentesco y de consanguinidad disponiendo del árbol genealógico de un individuo.

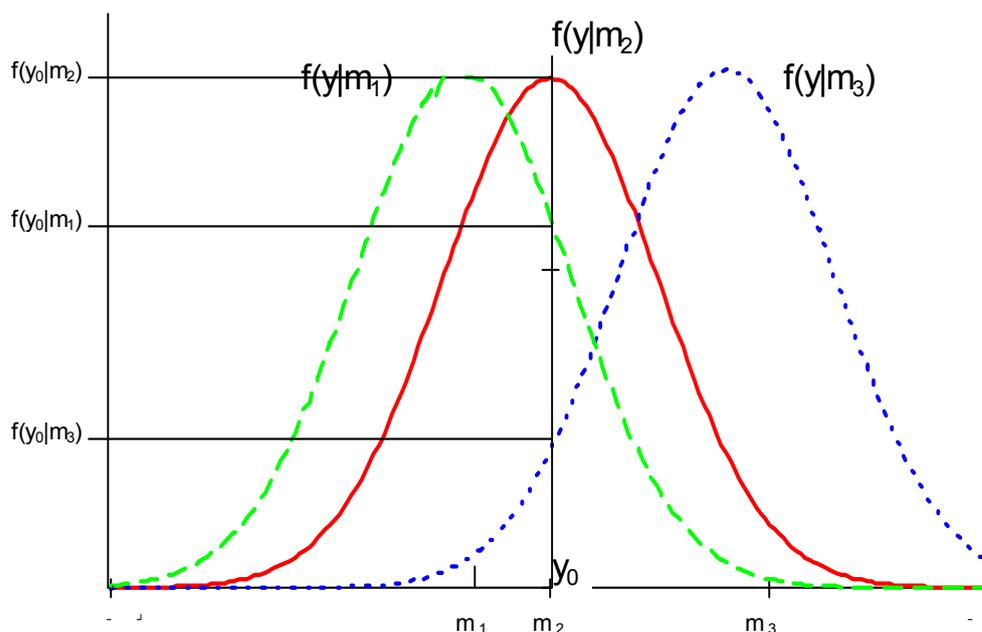
Apéndice II. La verosimilitud

Supongamos que conocemos cómo se distribuyen las muestras en el espacio muestral (esto es, cómo se agruparían distintas muestras si repitiéramos infinitas veces el experimento). Si llamamos $f(\mathbf{y}|u)$ a la función de densidad de la muestra para un valor de u determinado, el método de máxima verosimilitud consiste en maximizar $f(\mathbf{y}|u)$. Se supone que $f(\mathbf{y}|u)$ es función de u para cada muestra \mathbf{y} . Para explicarlo recurriremos a un ejemplo sencillo.

Supongamos que quiero averiguar la media del peso de conejos de una determinada raza a las ocho semanas de edad, y tomo una muestra. Para hacer el ejemplo más sencillo supongamos que la muestra es de un sólo conejo que pesa $y_0=1.60$ Kg. En la figura 1 se representan varias de las posibles poblaciones de las que puede haber sido extraído el conejo, con medias $m_1 = 1.50$, $m_2= 1.60$, $m_3= 1.80$.

Obsérvese que las densidades de probabilidad de la primera o la tercera población $f(y_0|m_1)$ y $f(y_0|m_3)$, son inferiores a la de la segunda población, $f(y_0|m_2)$. Debo reconocer que de todas las poblaciones de conejos posible, la que tiene una media como el peso de mi conejo es la que hace más verosímil que al muestrear haya sacado yo a ese conejo en concreto. Obsérvese que la función de verosimilitud está compuesta por valores $f(y_0|m_1)$, $f(y_0|m_2)$, etc. que forman una curva que no tiene por qué tener ni siquiera el aspecto de una función conocida.

Figura 1. Valores de verosimilitud para distintas medias



Obsérvese también que si represento los valores $f(y_0|m_1)$, $f(y_0|m_2)$,... obtengo una función en la que lo que varía son los valores de m mientras que lo que está fijado es el valor de la muestra, y_0 . En general esto es equivalente a suponer que disponemos de un número muy grande de valores $f(y|m_1)$, $f(y|m_2)$, $f(y|m_3)$,.... Cada uno representa una "probabilidad instantánea"⁽¹⁴⁾, en el sentido de que $f(y|m_1)$ es, por supuesto, una función de densidad de la muestra y *en unas circunstancias determinadas* (esto es, si y proviene de una población cuya media es m_1), y la probabilidad de que y esté entre dos valores y_0 e y_1 *en esas circunstancias concretas* viene dada por el área entre $f(y_0|m_1)$ y $f(y_1|m_1)$, pero al mismo tiempo es obvio que no se pueden sumar las probabilidades definidas por áreas en el entorno de $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$, ..., porque están tomadas de poblaciones diferentes, y el conjunto de estos valores no obedece a las leyes de la probabilidad -no suman 1, por ejemplo-.

Fisher (1912) propone tomar el valor m_i que dé lugar a un valor mayor de $f(y|m_i)$, porque le parece intuitivo que de todas las poblaciones a las que da lugar la serie $f(y|m_1)$, $f(y|m_2)$, $f(y|m_3)$, la que tiene mayor valor de $f(y|m_i)$ es la que hace que la muestra que se ha tomado parezca más *probable*. Como aquí la palabra *probable* puede inducir a confusión, puesto que ya hemos dicho que esas "probabilidades instantáneas" no pueden ser tomadas como probabilidades en

¹⁴ Con este nombre es como en el artículo de 1912 aparece la verosimilitud. Fisher no utilizará el término *verosimilitud* hasta mucho después.

conjunto, posteriormente dirá que hay que tomar el valor m_i que hace más *verosímil* el que cuando hemos muestreado haya salido esa muestra. Obsérvese que el método de máxima verosimilitud no es, por tanto, el que procura el estimador *más probable* dada la muestra. Literalmente, el método de máxima verosimilitud provee el valor del parámetro que, *de ser verdadero*, haría más probable a la muestra observada

Cuando Fisher propuso el método no quedaba claro que este estimador fuera particularmente bueno. Fisher lo propuso porque le parecía que la verosimilitud suministraba un *grado de creencia racional* que, aunque no gozara de las propiedades de una probabilidad, le permitía expresar la incertidumbre de una forma no muy diferente. Fisher proponía en realidad usar toda la curva de verosimilitud y no sólo su máximo, lo que hoy en día es frecuente en algunos problemas de marcadores genéticos. En esos casos se corre el riesgo de acabar interpretando la verosimilitud como si fuera una probabilidad. En realidad el estadístico frecuentista no usa la curva de verosimilitud sino sólo su máximo. El método ha sido aceptado por sus buenas propiedades frecuentistas -es asintóticamente insesgado, suficiente cuando hay estimadores suficientes, eficiente, óptimo asintóticamente normal, etc.- , pero estas propiedades son asintóticas, y si las muestras son pequeñas no tienen por qué conducir a un estimador particularmente bueno. Por otra parte no es necesariamente el estimador que minimiza el Riesgo. El método presenta, sin embargo, una ventaja aparte de sus propiedades como estimador frecuentista: Cualquier reparametrización conduce a la misma estima; por ejemplo, la estima máximo verosímil de la varianza es el cuadrado de la estima máximo verosímil de la desviación típica, y un estimador máximo verosímil función de otros que también lo sean, es a su vez máximo verosímil.

Desde un punto de vista práctico el método de máxima verosimilitud es una herramienta importante para el investigador aplicado. La escuela frecuentista ha determinado una serie de propiedades interesantes que deben reunir los estimadores, pero no indica cómo hacerse con estimadores que las posean. El método de máxima verosimilitud da al investigador y al técnico una forma de obtener estimadores razonables, si bien de forma asintótica.

Apéndice III. La inferencia bayesiana

La escuela bayesiana fue fundada por Laplace por medio de varios trabajos publicados de 1774 a 1812, y durante el siglo XIX ocupó un papel preponderante en la inferencia científica (Stigler, 1986). Antes que Laplace, y sin que al parecer éste tuviera conocimiento, se había presentado en la Royal Society de Londres un trabajo póstumo atribuido a un oscuro clérigo, el

reverendo Thomas Bayes (quien no publicó trabajos matemáticos en vida), formalizando el mismo principio de inferencia. Al parecer este principio había sido formulado anteriormente, y Stigler (1983) lo atribuye a Sauderson, un profesor de óptica ciego, autor de numerosos trabajos en diversos campos de la matemática. Los trabajos sobre verosimilitud de Fisher en los años 20 y los de la escuela frecuentista en los 30 y 40 hicieron casi desaparecer a la escuela bayesiana, hasta que comenzó un "revival" a mediados de los 50 que dura *in crescendo* hasta nuestros días. En mejora genética animal el bayesianismo fue introducido por Daniel Gianola, primero en trabajos sobre caracteres umbral en colaboración con J.L. Foulley, y posteriormente en artículos en los que se desarrollan aplicaciones a prácticamente todos los campos de la mejora genética animal (ver Blasco, 2001, para una revisión sobre los métodos bayesianos en mejora genética).

La forma esencial de trabajar de la escuela bayesiana consiste en, dados los datos observados en el experimento, describir toda la incertidumbre que puede existir en torno a un parámetro, usando como medida natural de la incertidumbre la probabilidad de que el parámetro tome determinados valores. Por ejemplo, en el caso de la heredabilidad se obtendría la función de densidad de probabilidad $f(h^2|\mathbf{y})$ siendo \mathbf{y} el vector de valores observados. Una vez obtenida esa distribución se pueden hacer inferencias de múltiples maneras: por ejemplo, se puede desear averiguar entre qué valores se encuentra h^2 con una probabilidad del 95%, o qué probabilidad tiene el que h^2 esté entre tal y tal valor. En los casos en los que es necesaria una estimación puntual de h^2 , por ejemplo para un índice de selección, hay varios parámetros de la función de densidad $f(h^2|\mathbf{y})$ que pueden ser usados como estimación puntual, y cuyo uso depende de la preferencia del investigador. Por ejemplo, la *moda*, que es el valor más probable de h^2 dada la muestra \mathbf{y} ; la *mediana*, cuyo valor hace tan probable que el valor verdadero sea superior como inferior a esta estima y minimiza el riesgo de estimación cuando la función de pérdidas es $|h^2 - \hat{h}^2|$; o la *media*, que es el estimador que minimiza el riesgo mínimo cuadrático $E(h^2 - \hat{h}^2)^2$.

Para poder hacer todas estas inferencias es menester disponer de la función de densidad de probabilidad $f(h^2|\mathbf{y})$. De acuerdo con las leyes de la probabilidad, la probabilidad $P(A,B)$ de que se presenten dos sucesos simultáneamente es

$$P(A,B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

con lo que

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

En nuestro caso,

$$f(h^2|y) = f(y|h^2) f(h^2) / f(y) = cte \cdot f(y|h^2) f(h^2)$$

donde f significa “función de densidad”, pero no es necesariamente la misma para $y|h^2$ que para h^2 . Obsérvese que $f(h^2|y)$ es una función de h^2 , pero no de y , que está fijada; por tanto $f(y|h^2)$ es aquí función de h^2 , pero no de y , que es exactamente la definición de verosimilitud. Por la misma razón $f(y)$ es una constante, ya que no depende de h^2 e y está fijado. Finalmente, $f(h^2)$ es la densidad de probabilidad de h^2 al margen de nuestro experimento.

Las críticas al bayesianismo tienen que ver con esta última probabilidad llamada *a priori* porque no depende de los datos, es previa al experimento. En ocasiones esta información está claramente determinada; por ejemplo, la probabilidad *a priori* de obtener un individuo recesivo en el cruce de dos heterocigotos es 1/4, al margen del experimento, pero en el caso de la heredabilidad no está claro qué se quiere decir con esta probabilidad previa. En muchas ocasiones es difícil cuantificar la información *a priori* de una forma tan objetiva como la de los ejemplos que acabamos de citar. En esos casos los estadísticos no bayesianos consideran que no es posible aplicar el teorema de Bayes y el problema no tiene solución por la vía de las probabilidades. Dentro del campo bayesiano se ha intentado dar respuesta a esta dificultad de varias formas, bien definiendo la probabilidad como un estado de creencias del investigador, quien define $f(h^2)$ según su opinión y los experimentos realizados previamente o consultados en la literatura, o bien eliminando en la práctica la influencia de la probabilidad *a priori* a base de aumentar el tamaño muestral. Si se dispusiera de suficientes datos, la probabilidad *a priori* no influiría en la distribución de la densidad posterior de probabilidades, por tanto se deben hacer experimentos con un número de datos suficiente como para que la función *a priori* carezca de relevancia. En ese caso la función de densidad de probabilidades *a priori* se busca de forma relativamente arbitraria (se procura que coincida en lo posible con una opinión defendible; p. ej, que no sea muy probable que la heredabilidad tenga un valor de 0.95), y habitualmente se procura que facilite los cálculos de la función posterior y que no conduzca a paradojas o a resultados inadmisibles. Es frecuente en ese caso probar varias funciones *a priori* diferentes y alguna función de referencia (p. ej., un *a priori* plano en el que todos los valores presentan la misma probabilidad) para comprobar que el resultado final (la función posterior) apenas se altera. Cuando no hay información *a priori*, o cuando se desea actuar como si no la hubiera, el bayesianismo se enfrenta a la dificultad de que es imposible realizar inferencias, puesto que la probabilidad *a priori* es necesaria para poder aplicar el teorema de Bayes, y cualquier forma

que tenga esa probabilidad es de alguna manera informativa. Se ha sugerido suponer que cuando no hay información sobre los distintos sucesos posibles, hay que asignarles a todos la misma probabilidad *a priori*. En el caso de variables continuas esto implica representarlas como una recta paralela al eje de las X en un intervalo concreto, por ejemplo al intervalo [0,1] para el caso de la heredabilidad, por lo que se les conoce también como *a priori planos* o *no informativos*, siendo este último nombre inapropiado, puesto que sí que son informativos (no es lo mismo decir que se ignora la probabilidad de los distintos sucesos que decir que todos tienen la misma probabilidad). Estos *a priori planos* son frecuentes en la literatura como funciones de referencia. Otras soluciones más complejas aunque escasamente aplicadas en el campo de la mejora genética son discutidas por Blasco (2001).

Apéndice IV. Análisis de datos seleccionados

Si llamamos \mathbf{d} al vector de efectos y parámetros que se desea estimar, la inferencia bayesiana se basa en el examen de las funciones de densidad de probabilidad marginales de los efectos y parámetros a estimar, para lo que se necesita la distribución conjunta $P(\mathbf{d} | \mathbf{y}) = P(\mathbf{y} | \mathbf{d}) P(\mathbf{d})$, como vimos.

El problema reside en que esa distribución puede no ser la misma si los datos están seleccionados, puesto que entonces \mathbf{y} pertenece a un subconjunto del espacio muestral, no es un conjunto de datos tomados al azar. Supongamos un proceso de selección en el que en cada generación se selecciona un subconjunto de animales basándonos en sus datos, de forma que disponemos tanto de los datos de los individuos seleccionados como de los individuos eliminados. Disponemos de los datos \mathbf{y}_0 de la generación base, en la que los individuos están tomados al azar de la población. En esa generación fueron seleccionados un conjunto de individuos en base a sus datos; estos individuos dieron lugar a la primera generación de selección, de la que disponemos todos los datos \mathbf{y}_1 , tanto de los datos de los individuos seleccionados como de los desechados. Nuestro objetivo es inferir los valores de \mathbf{d} a partir del total de los datos disponibles.

$$P_S(\mathbf{d} | \mathbf{y}) = P_S(\mathbf{y} | \mathbf{d}) P(\mathbf{d}) / P(\mathbf{y})$$

$$P_s(\mathbf{y} | \mathbf{d}) = P_s(\mathbf{y}_0, \mathbf{y}_1 | \mathbf{d}) = P_s(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{d}) P_s(\mathbf{y}_0 | \mathbf{d})$$

Ahora bien, como \mathbf{y}_0 es una muestra aleatoria de la población base, al no ser datos seleccionados,

$$P_s(\mathbf{y}_0 | \mathbf{d}) = P(\mathbf{y}_0 | \mathbf{d})$$

Los datos de la generación 1 no son una muestra aleatoria de la generación base porque son hijos de animales seleccionados. Deberían ser, por tanto, “mejores” que los datos de la población base. Pero si condicionamos a los datos de la población base, la distribución de esos datos “dados los datos de la población base” es la misma con o sin selección, puesto que no estamos examinando la distribución completa de estos datos sino la distribución que no depende de la selección, la condicionada a los datos usados en la selección. Por tanto,

$$P_s(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{d}) = P(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{d})$$

así que al fin tenemos que

$$P_s(\mathbf{y} | \mathbf{d}) = P(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{d}) P(\mathbf{y}_0 | \mathbf{d}) = P(\mathbf{y} | \mathbf{d})$$

por tanto,

$$P_s(\mathbf{d} | \mathbf{y}) = P_s(\mathbf{y} | \mathbf{d}) P(\mathbf{d}) / P(\mathbf{y}) = P(\mathbf{y} | \mathbf{d}) P(\mathbf{d}) / P(\mathbf{y}) = P(\mathbf{d} | \mathbf{y}) / P(\mathbf{y})$$

Este resultado es crucial, puesto que permite realizar inferencias en poblaciones seleccionadas ignorando el proceso de selección. Para ello es menester que en \mathbf{y} estén contenidos todos los datos usados al seleccionar, y las relaciones entre ellos que hacen posible condicionar como acabamos de hacer. Obsérvese que

$$P_s(\mathbf{y}_1 | \mathbf{d}) \neq P(\mathbf{y}_1 | \mathbf{d})$$

aunque

$$P_s(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{d}) = P(\mathbf{y}_1 | \mathbf{y}_0, \mathbf{d})$$

de ahí la necesidad de disponer de todos los datos usados en la selección y de las relaciones usadas al seleccionar. Lo mismo podemos hacer con los datos correlacionados. Estos datos están sometidos a un proceso de selección indirecta a través de la correlación con el carácter (o los caracteres) objeto de selección.

Por tanto, pueden construirse las distribuciones posteriores ignorando el hecho de que haya habido selección. Para ello es necesario que sea posible la condicionalización que hemos expuesto; esto es, que estén incluidos todos los datos del proceso de selección así como las relaciones entre ellos que permiten la condicionalización. Las relaciones entre datos están en las matrices **A** y **G** del modelo, por lo que son necesarias para la inferencia. Esta argumentación sirve asimismo para los razonamientos basados en la verosimilitud. Una consideración detallada de estos temas se encuentra en Gianola et al.(1989) y en Im et al. (1989).

Capítulo 3

SELECCIÓN

3.1. Parámetros de la selección

3.1.1. Diferencial de selección y Respuesta

3.1.2. Intervalo generacional

3.1.3. Criterio de selección

3.2. Métodos de selección.

3.2.1. Selección directa

Selección individual

Selección indirecta

3.2.2. Selección parientes

3.3. Índices de selección para un carácter y para varios caracteres.

3.3.1. Índices para un carácter

3.3.2. Índices para varios caracteres

Observaciones sobre los índices

3.4. Modelos lineales. El BLUP.

3.4.1. BLUP

3.4.2. Las ecuaciones del modelo mixto. El modelo animal/planta

3.4.3. Otros modelos

3.4.4. Otras interpretaciones del BLUP

3.5. Cálculo de los valores de mejora. Software disponible.

3.1. Parámetros de la selección

3.1.1. DIFERENCIAL DE SELECCION Y RESPUESTA

Si un carácter es heredable, al representar los valores de padres e hijos en unos ejes de coordenadas, observaremos que por término medio el grupo de mejores padres da lugar al grupo de mejores hijos. En la **figura 5** se representa la relación entre padres e hijos para el carácter peso a las 9 semanas de edad en conejo. Cada punto representa una pareja (x =media del valor en los padres, y =media del valor en los hijos) suponiendo que la población no está seleccionada y la media global de padres y de hijos es la misma.

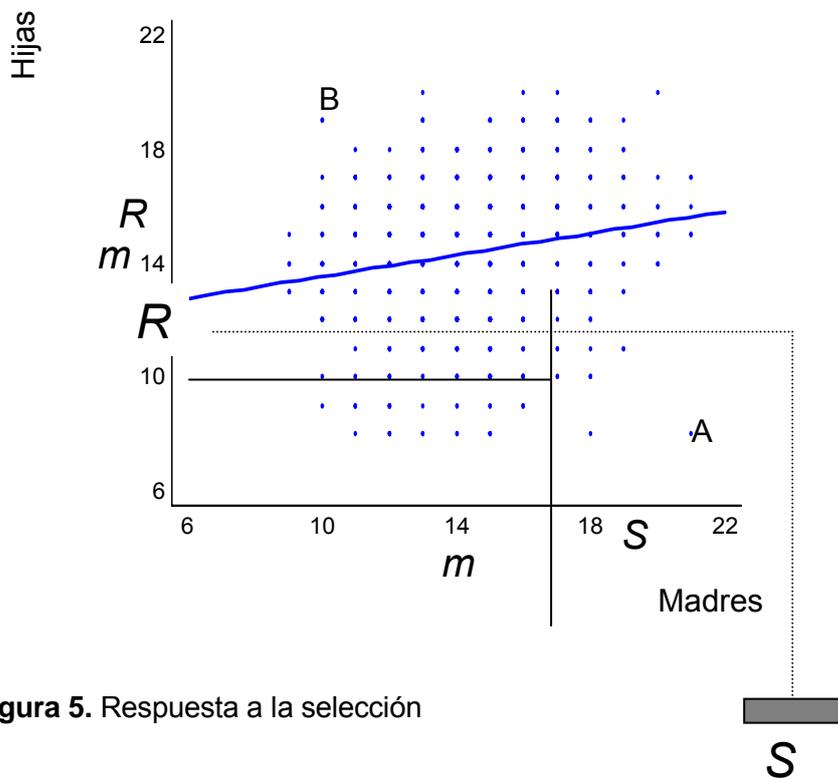


Figura 5. Respuesta a la selección

Si trazamos una recta de regresión para predecir la media de las hijas a partir de la media de los padres, tendremos la ecuación, ya establecida en el capítulo anterior

$$(H - m) = h^2 (P - m) \quad (3.1)$$

Vemos que padres con los mismos valores pueden dar lugar a hijos de valores muy distintos, hay una gran dispersión de puntos a lo largo de la recta. Vemos también que aunque en algún caso particular a buenos padres corresponden malos hijos -caso A- y al contrario -caso B-, si seleccionamos un conjunto de los mejores padres la media de todos sus hijos será algo mayor que la media de la población. Superioridad que no se traduce sino en una pequeña parte en la descendencia, debido a que h^2 suele ser pequeña.

A la media de padres seleccionados (área sombreada) se le denomina *diferencial de selección* (S), y a la media de los hijos de padres seleccionados (área sombreada) se le llama *respuesta a la selección* (R). Aplicando la ecuación 3.1 a estos valores obtenemos

$$R = h^2 S \quad (3.2)$$

La Respuesta depende no sólo de la heredabilidad del carácter sino de las posibilidades que tiene la población para permitir que la superioridad de los padres seleccionados S sea grande. En una población en la que todos los individuos se parecen, en donde no haya una variabilidad notable, no será posible encontrar grandes valores de S . Esto se pone de manifiesto si tipificamos el diferencial de selección, con lo que obtenemos la fórmula

$$R = h^2 \frac{S}{\sigma_P} = h^2 \cdot i \cdot \sigma_P \quad (3.3)$$

en la que se observa que no sólo es necesaria una heredabilidad elevada para obtener respuestas elevadas, sino que es menester que el carácter tenga una variabilidad suficiente. Así, caracteres como el tamaño de camada en especies prolíficas, cuya heredabilidad suele estar entre 0.05 y 0.10, son susceptibles de selección debido a la elevada variabilidad fenotípica que presentan (un coeficiente de variación en torno a 0.30, en comparación con un valor de 0.10 para, por ejemplo, caracteres de crecimiento).

Al diferencial de selección tipificado i se le conoce como *intensidad de selección*. Su utilidad deriva de que, al ser un parámetro adimensional de media cero y varianza la unidad, permite comparar la intensidad con la que se abordan distintos procesos de selección. Esta intensidad suele estar determinada por los medios con los que el experimentador cuenta. Por ejemplo, si se dispone de un número determinado de jaulas, no es posible guardar más animales que jaulas disponibles. A veces el problema es de tipo biológico: por ejemplo, si una vaca tiene cuatro partos a lo largo de su vida productiva no se dispone más que de 2 novillas por vaca, por término medio, para hacer la reposición, lo que hace que la selección vía madres sea muy poco eficaz. Resultaría, pues, interesante, expresar la intensidad de selección en función de la proporción de individuos seleccionados. Suponiendo que el carácter se distribuya de forma Normal (figura 6), y que los individuos son seleccionados cuando superan un cierto umbral c , si trabajamos con los caracteres tipificados, $m=0$, con lo que la media de los individuos seleccionados es igual a la intensidad de selección. Llamando p a la proporción de individuos seleccionados, la *media* de estos individuos es

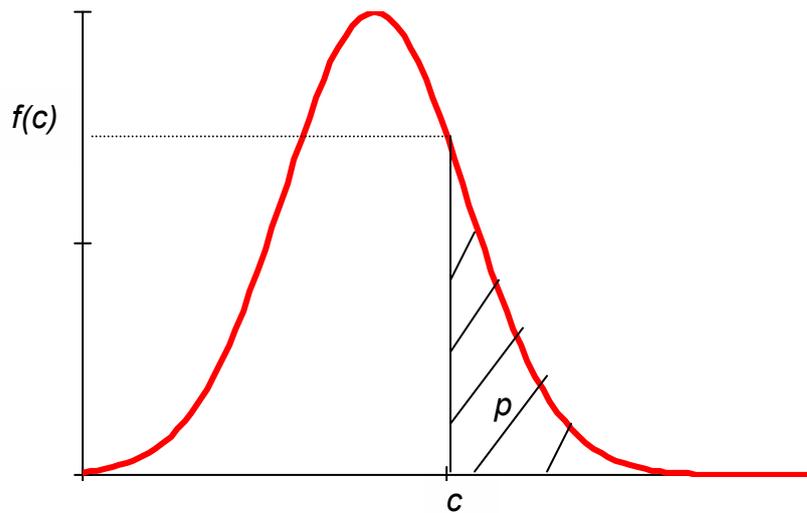


Figura 6. Presión de selección

$$i = \frac{1}{p} \int_c^{\infty} x \cdot f(x) dx = \frac{1}{p} \int_c^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

pero obsérvese que $\frac{df(x)}{dx} = -x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) = -x \cdot f(x)$, por lo que cambiando el signo dentro de la integral, tenemos a la propia derivada de la función, entonces

$$\begin{aligned} i &= -\frac{1}{p} \int_c^{\infty} -x \cdot f(x) dx = -\frac{1}{p} \int_c^{\infty} \frac{df(x)}{dx} \cdot dx = -\frac{1}{p} \int_c^{\infty} df(x) = -\frac{1}{p} [f(x)]_c^{\infty} = -\frac{1}{p} \cdot \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \right]_c^{\infty} = \\ &= \frac{1}{p} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{c^2}{2}\right) = \frac{1}{p} \cdot f(c) \end{aligned} \quad (3.4)$$

con lo que la expresión de la Respuesta pasa a ser

$$R = h^2 \cdot \frac{f(c)}{p} \cdot \sigma_P \quad (3.5)$$

donde se puede ver que aumentando la *presión de selección*; esto es, disminuyendo la proporción de individuos seleccionados, la respuesta aumenta.

Si la presión de selección es distinta en machos que en hembras, el diferencial de selección es la media de los diferenciales de los machos y las hembras, y la intensidad de selección es también la media aritmética de las intensidades.

EJEMPLO 3.1

Enunciado: En una población de conejos compuesta de 100 hembras y 20 machos se decide efectuar selección por ganancia media diaria de peso. Decidir el umbral a partir del cual se seleccionará cada individuo. Estimar la respuesta esperada en una generación de selección. Para reducir el intervalo generacional, las conejas tienen sólo un parto, con cuatro hijos útiles (que llegan a la edad de reproducción) por parto. La media de la población crece a razón de 35 g/día, y la varianza fenotípica del carácter es 78 g². La heredabilidad del carácter es, en esa población, $h^2 = 0.32$.

Resolución: El primer parto produce cuatrocientos conejos, de los que 200 son machos y 200 son hembras, aproximadamente. Como tras la selección hay que reconstituir la población original de 100 hembras y 20 machos, la presión de selección es, en hembras p_H y machos p_M respectivamente:

$$p_H = 100 / 200 = 50\% \quad ; \quad p_M = 20 / 200 = 10\%$$

Recurriendo a las tablas de la Normal tipificada, el valor de los umbrales para las presiones de 50 y 10% son $c_H = 0$ y $c_M = 1.28$ respectivamente, y los valores de la función son $f(c_H) = 0.40$; $f(c_M) = 0.18$.

Para calcular los umbrales con los datos sin tipificar, hacemos:

$$\frac{x - m}{\sigma_P} = c \quad \rightarrow \quad x = m + c \cdot \sigma_P \quad \rightarrow \quad \begin{cases} x_H = 35 + 0 \cdot \sqrt{78} = 35 \\ x_M = 35 + 1.28 \cdot \sqrt{78} = 46.3 \end{cases}$$

Por tanto seleccionaremos a los machos que superen 46.3 g/día y a las hembras que superen 35 g/día. Para calcular la respuesta esperada necesitamos las intensidades de selección (ecuación 3.4). Las intensidades de selección serán

$$i_H = 0.40 / 0.50 = 0.8 \quad ; \quad i_M = 0.18 / 0.10 = 1.8$$

La respuesta esperada y la media en la próxima generación serán (ecuación 3.3)

$$R = 0.32 \cdot \frac{0.8 + 1.8}{2} \cdot \sqrt{78} = 3.7 \quad ; \quad m_1 = 35 + 3.7 = 38.7$$

Cuando la heredabilidad se calcula como

$$h^2 = R / S$$

se le llama *heredabilidad realizada*, porque es la que efectivamente interviene en el proceso de selección. Otras estimas de la heredabilidad, como la que se obtiene a través de las correlaciones entre hermanos, puede estar sesgada debido a interacciones no consideradas, a covarianzas ambientales no consideradas (efectos maternos) o a otras causas, lo que no ocurre cuando ya se ha efectuado el proceso selectivo.

3.1.2. EL INTERVALO GENERACIONAL

La eficacia de un proceso de selección viene determinada por la respuesta obtenida por unidad de tiempo. El deseo de disminuir la proporción de individuos seleccionados puede conducir a mayores respuestas globales, pero menores respuestas por unidad de tiempo. Por ejemplo, si las cerdas de una línea tienen un tamaño medio de camada de 10 lechones, y los candidatos a la selección provienen de cuatro partos, dispondremos de 40 candidatos entre machos y hembras para efectuar la reposición, pero si provienen de los dos primeros partos sólo tendremos 20. La respuesta será mayor en el primer caso, pero el tiempo necesario en alcanzarla también, por lo que es necesario calcular bien el *intervalo entre generaciones* (L) en ambos casos para estimar la respuesta por unidad de tiempo.

Como para un ambiente dado la heredabilidad es una característica de la línea, **la acción fundamental del seleccionador consiste en actuar sobre el cociente i / L** , donde, en caso de distintas intensidades de selección, i sería la intensidad media. Si la selección no se realiza mediante generaciones discretas sino con generaciones solapadas y hay distintos intervalos generacionales para los dos sexos, L sería el intervalo generacional medio.

EJEMPLO 3.2

Enunciado: En el caso del problema anterior, en el ejemplo 3.1, calcular la respuesta por unidad de tiempo en el caso de que se seleccionaran individuos provenientes de dos partos en lugar de uno. El intervalo generacional con un parto es de 6 meses, y con dos partos de 7.5 meses.

Resolución: Con dos partos se dispone del doble de individuos para seleccionar, por lo que las presiones de selección se reducen a la mitad, $p_H=25\%$ y $p_M=5\%$. Recurriendo a las tablas de la Normal tipificada, los valores de la función para esas presiones son $f(c_H) = 0.32$; $f(c_M) = 0.10$. Si repitiéramos los cálculos del ejemplo anterior, seleccionaríamos hembras a partir de 41 g/día y machos a partir de 49.6 g/día. Las intensidades de selección serán (ecuación 3.4)

$$i_H = 0.32 / 0.25 = 1.28 \quad ; \quad i_M = 0.10 / 0.05 = 2$$

La respuesta esperada y las respuestas por unidad de tiempo en la próxima generación serán (ecuación 3.5)

$$R = 0.32 \cdot \frac{1.28 + 2}{2} \cdot \sqrt{78} = 4.6 \quad \left\{ \begin{array}{l} R_1 = \frac{3.7}{6} \cdot 12 = 7.4 \text{ g/año} \\ R_2 = \frac{4.6}{8} \cdot 12 = 6.9 \text{ g/año} \end{array} \right.$$

como se ve, aunque la respuesta usando dos partos sea superior, no compensa al alargarse el intervalo generacional. En la práctica intervienen otros criterios (por ejemplo, una presión de selección dentro de machos elevada aumenta excesivamente la consanguinidad, por lo que se suele seleccionar machos dentro de familia de macho), algunos de ellos no genético (por ejemplo, el elevado coste de eliminar a las hembras con solo un parto).

3.1.3. CRITERIO DE SELECCION

En la figura 5 la selección se ha hecho basándose en la media del valor fenotípico de los padres, ese ha sido el *criterio de selección* utilizado. Para seleccionar se intenta estimar el valor aditivo de los individuos, y en general quedarse como progenitores de la generación siguiente a los individuos cuya estima sea la mayor. En el caso de la selección individual, el *criterio de selección* o estimador del valor aditivo es el valor fenotípico del individuo, pero no siempre es así; como veremos más adelante, en ocasiones se utiliza la media de los individuos de la familia, o la información de algunos parientes. En ese caso la figura es la misma, pero en eje X está representado el *criterio C* que se ha usado en lugar del valor fenotípico. La ecuación se construye de forma análoga a la ecuación 3.1,

$$H - m = b (C - m) \quad (3.6)$$

sólo que aquí el coeficiente b ya no coincide con la heredabilidad, sino que es

$$b = \frac{\text{cov}(A,C)}{\sigma_C^2} \quad (3.7)$$

La predicción de la respuesta toma la forma

$$R = b \cdot S = \frac{\text{cov}(A,C)}{\sigma_C^2} \cdot i \cdot \sigma_C = \rho_{A,C} \cdot i \cdot \sigma_A \quad (3.8)$$

donde

$$\rho_{A,C} = \frac{\text{cov}(A,C)}{\sigma_A \sigma_C} \quad (3.9)$$

es el coeficiente de correlación entre el valor aditivo y el criterio de selección, e indica la *precisión del criterio* a la hora de estimar el valor aditivo del individuo. En esta expresión se observa claramente cómo el éxito de la selección depende por un lado de la existencia de variabilidad genética y por otro de la existencia de un criterio que estime adecuadamente el valor aditivo.

La estimación del valor aditivo de los individuos se puede realizar por regresión y representar como

$$A - m = b(C - m) + e \quad \rightarrow \quad \hat{A} = m + b(C - m) \quad (3.10)$$

donde b es el coeficiente de regresión, y es igual al de la ecuación 3.7.

La precisión de esta estimación viene determinada por la desviación típica del error, que teniendo en cuenta que el criterio y el error son independientes, es

$$\begin{aligned} \sigma_A^2 &= b^2 \sigma_C^2 + \sigma_e^2 \\ \sigma_e^2 &= \sigma_A^2 - b^2 \sigma_C^2 = \sigma_A^2 - \frac{\text{cov}^2(A,C)}{\sigma_C^2} = \sigma_A^2 (1 - \rho_{A,C}^2) = h^2 \sigma_P^2 (1 - \rho_{A,C}^2) \end{aligned} \quad (3.11)$$

No todos los individuos que tienen el mismo valor aditivo estimado \hat{A} tienen el mismo valor aditivo real A . Como los errores se distribuyen de forma Normal, se puede calcular la probabilidad de que el valor aditivo verdadero se encuentre en un cierto intervalo. En la distribución Normal el 95% del área se encuentra aproximadamente en un intervalo entre dos desviaciones típicas, por lo que la probabilidad de que el valor aditivo verdadero se encuentre entre más menos dos desviaciones típicas del error es del 95%. Esto se expresa

$$P(A \in \hat{A} \pm 2 \sigma_e) = 0.95 \quad (3.12)$$

PARAMETROS DE LA SELECCION

Respuesta: Es la diferencia entre la media de la población en la generación seleccionada (hijos) y la generación base.

Diferencial de Selección: Es la diferencia entre la media de los padres y la media de la población antes de seleccionar.

Intensidad de selección: Es el diferencial de selección tipificado.

Presión de selección: Es la proporción de individuos seleccionados.

Intervalo generacional: Es el intervalo de tiempo entre dos estados de reproducción análogos (por ejemplo pubertad o primer parto) de la generación seleccionada y la generación parental.

Criterio de selección: Es el criterio utilizado para estimar el valor aditivo de los individuos y seleccionar en consecuencia.

Intervalo de confianza al 95%: Es un intervalo en el que sostenemos que se encuentra el valor aditivo, con una probabilidad de error del 5%.

Precisión del criterio: Es el coeficiente de correlación entre el valor aditivo y el criterio.

Heredabilidad realizada: Es la heredabilidad estimada como el cociente entre la respuesta y el diferencial de selección cuando la selección es individual.

3.2. Métodos de selección

3.2.1. SELECCIÓN INDIVIDUAL

Selección directa

Ya hemos visto este tipo de selección en el caso anterior. Aquí el *criterio de selección* C es el valor fenotípico del individuo, $C = P$. Teniendo en cuenta que

$$\text{cov}(A, P) = \text{cov}(A, A+E) = \sigma_A^2$$

tenemos que (ecuaciones 3.7, 3.9 y 3.11)

$$b = h^2$$

$$\rho_{A,P} = h$$

$$\sigma_e^2 = \sigma_P^2 \cdot h^2(1-h^2)$$

por tanto cuanto mayor sea la heredabilidad, mayor es la precisión con que el valor fenotípico estima el valor aditivo. La predicción de la respuesta es (ecuaciones 3.3 y 3.8)

$$R = \frac{\text{cov}(A, P)}{\sigma_P^2} \cdot i \cdot \sigma_P = h^2 \cdot i \cdot \sigma_P = h \cdot i \cdot \sigma_A \quad (3.13)$$

EJEMPLO 3.3

Enunciado: En el ejemplo anterior 3.1, calcular la precisión del método de selección y la desviación típica del error de estimación. Estimar el valor aditivo de un individuo cuyo valor fenotípico es de 50 g/día.

Resolución: La precisión es

$$\rho_{A,P} = \sqrt{0.32} = 0.56$$

Como se ve, y a pesar de que la heredabilidad es relativamente elevada, el criterio selección individual no es excesivamente preciso. La desviación típica del error es:

$$\sigma_e^2 = 78 \cdot 0.32 \cdot (1 - 0.32) = 17$$

lo que da lugar a una $\sigma_e = 4.1$. La predicción del valor aditivo del individuo es

$$\hat{A} = 35 + 0.32(50 - 35) = 39.8$$

con un intervalo de probabilidad del 95% (ecuación 3.12)

$$39.8 \pm 2 \cdot 4.1 = 39.8 \pm 8.2$$

supondremos, pues, que el verdadero valor aditivo está en el intervalo 39.8 ± 8.2 , es decir en el intervalo $[31.6, 48.0]$ con un riesgo de error del 5%.

Obsérvese que un individuo que se encuentra entre el 5% de los mejores, con un valor fenotípico elevado, tiene un valor aditivo estimado próximo a la media de la población. Obsérvese también que la imprecisión de la estimación es notoria.

Selección indirecta

Un caso particular de selección individual es la selección indirecta. Aquí se estima el valor aditivo de un carácter midiendo otro carácter. Es frecuente cuando el carácter de interés es difícil o caro de medir; por ejemplo, en el caso del índice de conversión individual, que es un carácter económicamente importante pero caro de medir (ya que hay que medir el consumo de pienso todos los días), y que es seleccionado habitualmente de forma indirecta a través de la velocidad de crecimiento (fácil de medir, puesto que sólo hay que pesar al individuo una o dos veces). Aquí se estima el valor aditivo del carácter seleccionado indirectamente (índice de conversión) \hat{A}_I mediante el valor fenotípico del carácter seleccionado directamente (velocidad de crecimiento) P_D . Teniendo en cuenta que

$$\text{cov}(A_I, P_D) = \text{cov}(A_I, A_D + E_D) = \text{cov}(A_I, A_D)$$

de las ecuaciones 3.7 y 3.9 obtenemos

$$b = \frac{\text{cov}(A_I, P_D)}{\text{var}(P_D)}$$

$$\rho_{A_I, P_D} = \frac{\text{cov}(A_I, A_D)}{\sigma_{A_I} \sigma_{P_D}} = \rho_g \cdot h_D$$

donde ρ_g es el coeficiente de correlación genética entre los caracteres y h_D la raíz cuadrada de la heredabilidad del carácter seleccionado directamente. Esto indica que la precisión depende tanto de que el carácter seleccionado directamente esté fuertemente correlacionado con el

carácter objeto de la selección, como de que efectivamente la heredabilidad del carácter que se selecciona directamente sea elevada.

La varianza del error de estimación es (ecuación 3.11)

$$\sigma_e^2 = \sigma_{A_i}^2 (1 - \rho_{A_i, P_D}^2) = \sigma_{P_i}^2 \cdot h_i^2 (1 - \rho_g^2 h_D^2)$$

fórmula notoriamente similar a la de la selección directa. La predicción de la respuesta a la selección indirecta es

$$R_{\text{correlacionada}} = \rho_{A_i, P_D} \cdot i \cdot \sigma_{A_i} = i \cdot \rho_g \cdot h_D \cdot \sigma_{A_i}$$

Si la selección se hubiera hecho directamente por el carácter objeto de selección, la respuesta hubiera sido (ecuación 3.13)

$$R = i \cdot h_i \cdot \sigma_{A_i}$$

por lo que la eficacia relativa de los métodos es

$$\frac{R_{\text{correlacionada}}}{R} = \rho_g \cdot \frac{h_D}{h_i}$$

EJEMPLO 3.4

Enunciado: Se desea mejorar el índice de conversión de pienso en carne de una población de conejos, pero como es muy caro de medir se decide seleccionar este carácter a través de la velocidad de crecimiento, que tiene una correlación genética de -0.7 con el índice de conversión ⁽¹⁵⁾. La heredabilidad del índice de conversión es de 0.25 , y la de la velocidad de crecimiento 0.32 . La desviación típica del índice de conversión es de 0.3 unidades. Averiguar la eficacia relativa de la selección de índice de conversión a través de la velocidad de crecimiento.

Resolución:

¹⁵ Obsérvese que se desea disminuir el índice de conversión (usar menos pienso para producir un kg de conejo), por lo que la correlación conveniente es, efectivamente, negativa.

$$\frac{R_{\text{correlacionada}}}{R} = -0.7 \cdot \sqrt{\frac{0.32}{0.20}} = -0.89$$

Lo que indica que el índice de conversión decrecería un 89% de lo que podría decrecer en caso de que la selección hubiera sido directamente por éste carácter. Obsérvese que si la heredabilidad del índice de conversión fuera 0.15 ó inferior, la respuesta correlacionada al seleccionar por velocidad de crecimiento superaría a la de la selección directamente por índice de conversión. La desviación típica del error es

$$\sigma_e = \sqrt{0.20 \cdot 0.3^2 (1 - 0.7^2 \cdot 0.32)} = 0.123$$

Si la selección se hiciera directamente, la desviación típica del error sería 0.120, algo menor. En ambos casos el intervalo de probabilidad del 95% es grande.

$$2 \cdot 0.123 = \pm 0.25$$

3.3.2. SELECCION POR PARIENTES

En ocasiones bien porque el carácter sólo se expresa en un sexo (por ejemplo la producción de leche o el tamaño de camada), bien porque resulta excesivamente caro el identificar a cada individuo y es más cómodo y económico medir una media familiar (caso corriente en piscicultura), la selección se realiza midiendo la media de producción de un grupo familiar, en el que ocasionalmente puede estar incluido el propio candidato a la selección.

Selección por la media de la descendencia

Un ejemplo típico es la selección por producción de leche en vacuno, que hace algunos años se realizaba midiendo la media de producción de las hijas de un toro. Aunque el BLUP ha dejado obsoleto este método en vacuno de leche, es interesante estudiarlo para hacer algunas consideraciones sobre la precisión de la estimación y sobre la diferencia entre efectos fijos y aleatorios. Aquí el *criterio de selección* C es el valor medio de la producción lechera de n hijas, $C = \bar{D}$, con lo que la estima del valor aditivo será (ecuación 3.10)

$$\hat{A} = m + b \cdot (\bar{D} - m) = m + \frac{\text{cov}(A, \bar{D})}{\sigma_{\bar{D}}^2} \cdot (\bar{D} - m)$$

Obsérvese que la variabilidad de la media de las hijas depende del número de hijas. Lógicamente los valores de sólo una hija, o de la media de dos hijas, son mucho más variables que los de la media de cien hijas. Como la varianza de la media de las hijas está en el denominador, esto implica que cuanto mayor sea el número de hijas menor será la varianza del valor medio de todas ellas, y consecuentemente mayor será la *estimación* de su valor aditivo (no será mayor su valor aditivo real, desconocido, que depende de los genes concretos de ese toro, sino su estimación). Formalizaremos a continuación estas ideas. Teniendo en cuenta que las hijas son medio hermanas y que por tanto la covarianza entre dos de ellas es $cov(D_i, D_j) = \sigma_A^2 / 4$, y que la covarianza entre el valor aditivo del padre y una de sus hijas i es $cov(A, A_i) = \sigma_A^2 / 2$, tenemos que

$$cov(A, \bar{D}) = cov\left(A, \frac{D_1 + D_2 + \dots + D_n}{n}\right) = \frac{1}{n} \cdot n \cdot cov(A, D_i) = cov(A, A_i + E_i) = \frac{1}{2} \sigma_A^2$$

$$\begin{aligned} \sigma_{\bar{D}}^2 &= var\left(\frac{D_1 + D_2 + \dots + D_n}{n}\right) = \frac{1}{n^2} [n \cdot var(D_i) + n(n-1)cov(D_i, D_j)] = \\ &= \frac{1}{n} \left[\sigma_P^2 + (n-1) \frac{\sigma_A^2}{4} \right] = \frac{\sigma_P^2}{n} \left(1 + (n-1) \frac{h^2}{4} \right) \end{aligned} \quad (3.14)$$

$$b = \frac{(n/2) \cdot h^2}{1 + (n-1)(h^2/4)} = \frac{2n}{(n-1) + (4/h^2)}$$

EJEMPLO 3.5

Enunciado: Supongamos que disponemos de dos toros, el toro X con 10 hijas evaluadas y el toro Y con 100 hijas evaluadas. La media de producción de sus hijas es exactamente la misma, 8.000 kg de leche al año. La media de la población es de 6.000 kg/año. Tomando un valor de heredabilidad para la producción de leche de 0.25, estimar sus valores aditivos.

Resolución:

$$b = \frac{2n}{(n-1) + 16}$$

Sus valores aditivos *estimados* serán:

$$\hat{A}_x = 6000 + \frac{2 \cdot 10}{9 + 16} \cdot (8000 - 6000) = 7600$$

$$\hat{A}_y = 6000 + \frac{2 \cdot 100}{99 + 16} \cdot (8000 - 6000) = 9478$$

como se ve, hay un efecto de “regresión”, de forma que las estimaciones de los valores aditivos son menores cuando la información es escasa. Disponiendo de infinitas hijas, la estimación del valor aditivo sería el doble de la media de las hijas (el toro tiene dos cromosomas y pasa la mitad de la dotación a sus espermatozoides. El valor aditivo de una hija depende también del valor aditivo de la madre).

Este efecto de “regresión” a la media de la población que se pone de manifiesto en el ejemplo es importante porque si sólo se evaluara a los toros por la media de sus hijas, solamente se seleccionarían toros con poca información. No es imposible que un toro evaluado con una sola hija tenga un valor de 20.000 kg de leche, pero es prácticamente imposible que un toro evaluado con 100 hijas tenga un valor medio de producción de sus hijas de 20.000 kg. Cuando se estima el valor de un toro simplemente con la media de sus hijas, se está considerando que el valor de cada toro es un efecto *fijo*, y no se tienen en cuenta las relaciones de parentesco ni la heredabilidad del carácter. Cuando se estima el valor del toro como lo hemos hecho nosotros, estamos considerando que el valor de cada toro es un efecto *aleatorio*, y que por tanto presenta una variabilidad y los distintos efectos están relacionados a través del parentesco.

La varianza del error de estimación es (ecuación 3.11)

$$\sigma_e^2 = \sigma_P^2 \cdot h^2 (1 - \rho_{A,D}^2)$$

obsérvese que conforme aumenta el número de datos, σ_D^2 disminuye, el cociente aumenta y la varianza del error de estimación disminuye. Sin embargo, la varianza del estimador del valor aditivo aumenta al aumentar la cantidad de información:

$$\text{var}(\hat{A}) = b^2 \sigma_D^2 = \frac{(1/4)\sigma_A^4}{\sigma_D^2}$$

obsérvese que al aumentar la cantidad de datos, σ_D^2 disminuye y por tanto aumenta $\text{var}(\hat{A})$. Cuando n tiende a infinito, de la ecuación 3.14, tenemos

$$\sigma_D^2 = \frac{\sigma_P^2}{n} \left(1 + (n-1) \frac{h^2}{4} \right) = \frac{\sigma_P^2}{n} + \frac{\sigma_P^2 h^2}{4} - \frac{\sigma_P^2 h^2}{4n} \approx \frac{\sigma_P^2 h^2}{4} = \frac{\sigma_A^2}{4}$$

$$\text{var}(\hat{A}) = \approx \frac{\sigma_A^4 / 4}{\sigma_A^2 / 4} = \sigma_A^2$$

$$\sigma_e^2 = \sigma_A^2 - \text{var}(\hat{A}) \approx \sigma_A^2 - \sigma_A^2 = 0$$

Cuando estimamos el valor aditivo del toro como la media de las hijas, el error de estimación es el error de estimación de una media; esto es, si $\hat{A} = \bar{D}$ entonces $\sigma_e^2 = \text{var}(\hat{A}) = \sigma_P^2 / n$, con lo que la aumentar el número de datos disminuyen simultáneamente $\text{var}(\hat{A})$ y la varianza del error. Esto ha sido tradicionalmente una fuente de confusión, y el origen reside en las diferencias entre efectos fijos y aleatorios (ver Apéndice I). En el caso de los efectos fijos, la varianza del error de estimación coincide con la varianza del estimador, cosa que no ocurre, como acabamos de ver, en el caso de que se estime el valor aditivo como un efecto aleatorio. En ambos casos intentamos minimizar la varianza del error de estimación, no la varianza del estimador, aunque por la costumbre se habla de esta última en el caso de efectos fijos.

La precisión del método es (ecuación 3.9)

$$\rho_{A,\bar{D}} = \frac{\text{cov}(A, \bar{D})}{\sigma_A \sigma_{\bar{D}}} = \frac{(1/2)\sigma_A^2}{\sigma_A \sigma_P \sqrt{\frac{1}{n} + \frac{(n-1)h^2}{4n}}} = \sqrt{\frac{n}{n-1 + \frac{4}{h^2}}}$$

conforme aumenta el número de hijas controladas, la precisión se acerca a 1. La predicción de la respuesta es (ecuación 3.8)

$$R = \rho_{A,\bar{D}} \cdot i \cdot \sigma_A$$

EJEMPLO 3.6

Enunciado: Calcular las respuestas esperadas relativamente si se seleccionan toros en base a la media de 10 hijas o de 100 hijas. Calcular la desviación típica del error de estimación del valor aditivo de los toros X e Y del ejemplo 3.5. Desviación típica de la producción de leche, 1.200 kg/año.

Resolución: Las precisiones son:

$$\rho_{A,\bar{D}}(10) = \sqrt{\frac{10}{10 - 1 + \frac{4}{0.25}}} = 0.63 \qquad \rho_{A,\bar{D}}(100) = \sqrt{\frac{100}{100 - 1 + \frac{4}{0.25}}} = 0.93$$

y la relación de respuestas, que coincide con el cociente de las precisiones, es

$$R_{100} / R_{10} = 0.93 / 0.63 = 1.48$$

esperamos por tanto un 48% más de respuesta usando 100 hijas. Las desviaciones típicas de los errores de estimación de los valores aditivos de los toros X e Y serán

$$\sigma_e(10) = \sqrt{1200^2 \cdot 0.25(1 - 0.63^2)} = 466$$

$$\sigma_e(100) = \sqrt{1200^2 \cdot 0.25(1 - 0.93^2)} = 220$$

y los intervalos de. 95% de probabilidad aproximadamente el doble de esas cantidades. En la práctica los toros son evaluados además con datos de medias hermanas, de su madre y en general de todos sus parientes, lo que mejora sustancialmente la precisión.

3.3. Índices de selección

Los índices de selección hacen su aparición en un artículo de Karl Pearson (1903) sobre selección natural, aunque tardarían mucho tiempo en ser aplicados a la mejora genética animal, y expresados de forma simple en los trabajos de Lush sobre selección combinada (Lush, 1947). En 1936 Smith (1936) propone los índices de selección para varios caracteres, aunque al parecer es a Fisher a quien debe considerársele como el autor del método, puesto que Smith dice en la sección de agradecimientos que la parte I de su artículo (Teoría) era poco

más que la transcripción de las instrucciones que había recibido de Fisher. Posteriormente Hazel (1947) aplicó estos índices a la mejora animal, y hoy en día se siguen utilizando tanto en experimentos de laboratorio como en núcleos de selección comerciales, particularmente en aves, cerdos y conejos por razones que veremos más adelante.

3.3.1. INDICES DE SELECCION PARA UN CARACTER

La idea básica de los índices de selección es estimar el valor aditivo de cada individuo con toda la información disponible en la población. Antes de que se generalizara el uso de ordenadores, los índices se restringían a los parientes más próximos, y su eficacia no era muy diferente de los índices que consideran toda la información, puesto que la información relevante se encuentra en los parientes más próximos, pero hoy en día no hay razones para no usar modelos más completos, que por otra parte resultan más fáciles de manejar desde un punto de vista informático.

Aquí el criterio de selección es un conjunto de parientes del individuo, aunque se puede tomar el conjunto de todos los datos de la población, presentes y pasados. Como se trata de varios datos, la forma de estimación es por regresión lineal múltiple. Cuando el carácter se distribuye de forma Normal, como el valor aditivo también se distribuye de forma Normal, al estar determinado por muchos genes de pequeño efecto cada uno, la relación entre el valor aditivo y el criterio es lineal, puesto que la relación entre variables que se distribuyen conjuntamente de forma Normal es lineal. Sin embargo hay casos en los que esto no es así y el carácter puede no distribuirse normalmente. En esos casos el índice es la mejor de las aproximaciones lineales, aunque no se puede garantizar que otra aproximación no lineal pudiera ser más eficaz.

En nuestro caso el criterio es $C = [y_1, y_2, \dots, y_n]$, y la estimación del valor aditivo se realiza por regresión múltiple.

$$A = \mathbf{b}' \mathbf{y} + \mathbf{e} = b_1 y_1 + b_2 y_2 + \dots + b_n y_n + \mathbf{e} = \hat{A} + \mathbf{e}$$

donde \hat{A} es la estima del valor aditivo

$$\hat{A} = \mathbf{l} = \mathbf{b}' \mathbf{y}$$

se le llama “índice de selección”. Encontraremos los valores de \mathbf{b} que minimizan el riesgo cuadrático medio (ver Apéndice I). El riesgo es (ecuación 3.22)

$$R(A, l) = E(A - l)^2 = [E(A) - E(l)]^2 + \text{var}(A) + \text{var}(l) - 2 \text{cov}(A, l)$$

Si los valores de \mathbf{y} están *centrados*,

$$E(l) = E(\mathbf{b}'\mathbf{y}) = \mathbf{b}' E(\mathbf{y}) = 0$$

$$\text{var}(l) = \text{var}(\mathbf{b}'\mathbf{y}) = \mathbf{b}' \text{var}(\mathbf{y}) \mathbf{b} = \mathbf{b}' \mathbf{V} \mathbf{b}$$

$$\text{cov}(A, l) = \text{cov}(A, \mathbf{b}'\mathbf{y}) = \mathbf{b}' \text{cov}(A, \mathbf{y}) = \mathbf{b}' \mathbf{c} = \mathbf{c}'\mathbf{b}$$

Si derivamos respecto a \mathbf{b} para obtener el valor que minimiza el riesgo, teniendo en cuenta que $E(A)$ y $\text{var}(A)$ son características de la población y constantes respecto a \mathbf{b} ,

$$\partial R / \partial \mathbf{b} = 2 \mathbf{b}' \mathbf{V} - 2 \mathbf{c}' = 0$$

por tanto $\mathbf{b}' = \mathbf{c}' \mathbf{V}^{-1}$, y la estima del valor aditivo,

$$\hat{A} = l = \mathbf{c}' \mathbf{V}^{-1} \mathbf{y}$$

Obsérvese que no se ha invocado la Normalidad de los datos en ningún momento, por lo que el índice minimiza el riesgo cuadrático medio entre los estimadores lineales independientemente de que los datos sean Normales o no.

EJEMPLO 3.7

Enunciado: Supongamos que los datos disponibles para evaluar a una coneja por tamaño de camada son los datos de uno de sus partos, los de dos medias hermanas y los de su madre. La heredabilidad del carácter es 0.10 y la varianza fenotípica 9 gazapos al cuadrado. Esto implica que la varianza aditiva es de 0.9 gazapos². Calcular el índice de selección.

Resolución:

$$\mathbf{y}' = [y_l, y_{HS}, y_{HS}, y_M]$$

$$\begin{aligned} \mathbf{c}' = \text{cov}(A_i, \mathbf{y}') &= [\text{cov}(A_i, y_l), \text{cov}(A_i, y_{HS}), \text{cov}(A_i, y_{HS}), \text{cov}(A_i, y_M)] = \\ &= [\sigma_A^2, \frac{1}{4} \sigma_A^2, \frac{1}{4} \sigma_A^2, \frac{1}{2} \sigma_A^2] = [0.9, 0.225, 0.225, 0.45] \end{aligned}$$

$$\begin{aligned}
 \mathbf{V} = \text{var}(\mathbf{y}) &= \begin{bmatrix} \text{var}(y_i) & \text{cov}(y_i, y_{HS}) & \text{cov}(y_i, y_{HS}) & \text{cov}(y_i, y_M) \\ & \text{var}(y_{HS}) & \text{cov}(y_{HS}, y_{HS}) & \text{cov}(y_{HS}, y_M) \\ & & \text{var}(y_{HS}) & \text{cov}(y_{HS}, y_M) \\ & & & \text{var}(y_M) \end{bmatrix} = \\
 &= \begin{bmatrix} \sigma_P^2 & (1/4)\sigma_A^2 & (1/4)\sigma_A^2 & (1/2)\sigma_A^2 \\ & \sigma_P^2 & (1/4)\sigma_A^2 & 0 \\ & & \sigma_P^2 & 0 \\ & & & \sigma_P^2 \end{bmatrix} = \begin{bmatrix} 9 & 0.225 & 0.225 & 0.45 \\ & 9 & 0.225 & 0 \\ & & 9 & 0 \\ & & & 9 \end{bmatrix}
 \end{aligned}$$

(como la matriz es simétrica se representa sólo el triángulo superior). Invertiendo y multiplicando se llega a los valores del índice.

$$\begin{aligned}
 \hat{A} = I = [0.9, 0.225, 0.225, 0.45] & \begin{bmatrix} 0.1115 & -0.0027 & -0.0027 & -0.0056 \\ & 0.1112 & -0.0027 & 0.0001 \\ & & 0.1112 & 0.0001 \\ & & & 0.1114 \end{bmatrix} \begin{bmatrix} y_i \\ y_{HS} \\ y_{HS} \\ y_M \end{bmatrix} \\
 &= 0.0966 y_i + 0.0220 y_{HS} + 0.0220 y_{HS} + 0.0452 y_M
 \end{aligned}$$

Precisión y predicción de la respuesta

$$\sigma_i^2 = \text{var}(\mathbf{c}' \mathbf{V}^{-1} \mathbf{y}) = \mathbf{c}' \mathbf{V}^{-1} \text{var}(\mathbf{y}) \mathbf{V}^{-1} \mathbf{c} = \mathbf{c}' \mathbf{V}^{-1} \mathbf{V} \mathbf{V}^{-1} \mathbf{c} = \mathbf{c}' \mathbf{V}^{-1} \mathbf{c}$$

$$\text{cov}(A, I) = \text{cov}(A, \mathbf{c}' \mathbf{V}^{-1} \mathbf{y}) = \mathbf{c}' \mathbf{V}^{-1} \text{cov}(A, \mathbf{y}) = \mathbf{c}' \mathbf{V}^{-1} \mathbf{c} = \sigma_i^2$$

$$\rho_{A,I} = \frac{\text{cov}(A, I)}{\sigma_A \sigma_I} = \frac{\sigma_i}{\sigma_A} \quad (3.15)$$

$$R = \rho_{A,I} \cdot i \cdot \sigma_I = i \cdot \sigma_i \quad (3.16)$$

$$\sigma_e^2 = \sigma_A^2 (1 - \rho_{A,I}^2) = \sigma_A^2 - \sigma_i^2$$

Nos encontramos aquí con el mismo fenómeno que al hablar de la prueba de la descendencia. La varianza del índice aumenta con la cantidad de información, puesto que se van añadiendo

sumandos positivos conforme aumenta la cantidad de información. Sin embargo la varianza del error de estimación disminuye, como cabe esperar.

EJEMPLO 3.8

Enunciado: Calcular la respuesta esperada para una intensidad de selección correspondiente a una presión del 25% (ver Ejemplo 3.1), y la precisión del índice del ejemplo 3.7.

Resolución:

$$\sigma_i^2 = [0.9, 0.225, 0.225, 0.45] \begin{bmatrix} 0.1115 & -0.0027 & -0.0027 & -0.0056 \\ & 0.1112 & -0.0027 & 0.0001 \\ & & 0.1112 & 0.0001 \\ & & & 0.1114 \end{bmatrix} \begin{bmatrix} 0.9 \\ 0.225 \\ 0.225 \\ 0.45 \end{bmatrix} =$$

$$= 0.12$$

$$\rho_{A,I} = \sqrt{\frac{0.12}{0.9}} = 0.36$$

$$\sigma_e^2 = 0.9 - 0.12 = 0.78 ; \quad \sigma_e = 0.88$$

$$R = 1.28 \cdot \sqrt{0.12} = 0.44 \text{ gazapos por generación.}$$

Con selección individual se habría obtenido 0.38 gazapos por generación, la precisión sería de 0.32 y la desviación típica del error de 0.90.

Propiedades de los índices de selección

Los índices de selección gozan de ciertas propiedades óptimas *dentro de los estimadores lineales*, como hemos precisado al principio de este apartado.

1) *Los índices de selección minimizan el Riesgo cuadrático medio de la estimación.*

2) *Los índices maximizan la precisión.* El valor de \mathbf{b} que maximiza la precisión es precisamente el del índice; esto es, $\mathbf{b} = \mathbf{V}^{-1}\mathbf{c}$. Para hallar el valor de \mathbf{b} que maximiza la precisión, tomamos logaritmos y derivamos e igualamos a cero.

$$\rho_{A,I} = \frac{\text{cov}(A, \mathbf{y})}{\sigma_A \sqrt{\text{var}(\mathbf{b}' \mathbf{y})}} = \frac{\mathbf{b}' \text{cov}(A, \mathbf{y})}{\sigma_A \sqrt{\mathbf{b}' \mathbf{V} \mathbf{b}}} = \frac{\mathbf{b}' \mathbf{c}}{\sigma_A \sqrt{\mathbf{b}' \mathbf{V} \mathbf{b}}}$$

$$\log(\rho_{A,I}) = \log(\mathbf{b}' \mathbf{c}) - \frac{1}{2} \log(\mathbf{b}' \mathbf{V} \mathbf{b}) - \log(\sigma_A)$$

$$\frac{\partial \log(\rho_{A,I})}{\partial \mathbf{b}'} = \frac{\mathbf{c}}{\mathbf{b}' \mathbf{c}} - \frac{2 \mathbf{V} \mathbf{b}}{2 \mathbf{b}' \mathbf{V} \mathbf{b}} = 0 \quad \longrightarrow \quad \mathbf{b} = \mathbf{V}^{-1} \mathbf{c}$$

3) *Los índices maximizan la Respuesta a la selección.* Los índices maximizan la precisión, que es $\rho_{A,I} = \sigma_I / \sigma_A$. Como la Respuesta es

$$R = i \cdot \sigma_I = i \cdot \rho_{A,I} \cdot \sigma_A \text{ (ecuación 3.15),}$$

y como σ_A es una característica de la población, el índice también maximiza la Respuesta.

4) *Los índices maximizan la probabilidad de ordenar correctamente a los animales con arreglo a su valor aditivo* (Henderson , 1963). Esta última propiedad es importante en especies en las que se venden genes de un animal evaluado con precisión, como es el caso del vacuno de leche, puesto que el ganadero puede escoger al reproductor en base a sus méritos genéticos. Esta propiedad requiere que el carácter se distribuya de forma Normal

3.3.2. SELECCIÓN PARA VARIOS CARACTERES

El beneficio que producen las plantas o los animales no depende de un solo carácter habitualmente sino de una combinación de ellos. Sopesando económicamente los caracteres podemos estimar el beneficio que produce un individuo:

$$B = a_1 y_1 + \dots + a_n y_n = \mathbf{a}' \mathbf{y}$$

donde a_i representa el beneficio obtenido por incremento de una unidad del carácter i (hay formas más complejas de calcular los pesos económicos, pero quedan fuera del alcance de este libro. Este índice indica el valor económico del individuo, pero no estamos interesados en

él al elegir reproductores, sino en el valor económico que *transmiten a la descendencia*; es decir, nuestro interés reside en encontrar el *valor aditivo económico* del individuo⁽¹⁶⁾

$$A_E = a_1 A_1 + \dots + a_n A_n = \mathbf{a}' \mathbf{u}$$

puesto que nos interesa lo que producirán sus hijos, no él mismo. Aquí \mathbf{u} es el vector de valores aditivos de cada carácter. Como antes, el *criterio* de selección que utilizaremos para estimar este valor será un conjunto de valores de varios caracteres medidos en cada individuo. Obsérvese que es posible que el conjunto de caracteres que determinan el *valor aditivo económico* no sea el mismo conjunto de caracteres que se puede medir. Por ejemplo, el índice de conversión puede ser un carácter económicamente importante, pero al ser difícil de medir podríamos seleccionarlo de forma indirecta a través de la velocidad de crecimiento, sin que la velocidad de crecimiento esté necesariamente considerada como un carácter económicamente relevante, por lo que pudiera no estar incluida en el *valor aditivo económico*. Por tanto, el criterio de selección puede contener datos de caracteres que no están incluidos en el *valor aditivo económico*, y a su vez el *valor aditivo económico* puede incluir valores aditivos de caracteres que no se miden y por tanto no están en el criterio de selección.

Llamemos, como antes, al criterio de selección $\mathbf{C} = [y_1, y_2 \dots y_m]$. De momento no consideraremos datos de los parientes, sino solamente del propio individuo. Obsérvese que los subíndices no indican que el carácter y_i se corresponde con el valor aditivo A_i ; si ordenamos el criterio adecuadamente podemos hacer que en los primeros casos sea así, pero no en los últimos, en donde puede estarse midiendo un carácter no incluido en el *valor aditivo económico*, o puede haber un carácter incluido en este *valor aditivo económico* que no se esté midiendo. Este punto se aclara en el ejemplo 3.9. Estimaremos el *valor aditivo económico* por regresión, como antes. Para cada carácter u_i

$$\hat{u}_i = \hat{A}_i = \mathbf{c}_i' \mathbf{V}^{-1} \mathbf{y}$$

Considerando todos los caracteres

$$\mathbf{u} = \mathbf{C}' \mathbf{V}^{-1} \mathbf{y}$$

donde \mathbf{C} es una matriz en la que en cada fila hay un vector $\mathbf{c}_i = \text{cov}(A_i, \mathbf{y})$ abarcando todos los caracteres $i = 1, \dots, n$

¹⁶ En la literatura el valor aditivo económico aparece con el nombre de "genotipo agregado", bastante más oscuro que el nombre que aquí proponemos.

$$A_E = I = \mathbf{a}' \mathbf{u} = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{y}$$

EJEMPLO 3.9

Enunciado: En porcino el principal coste de producción es la alimentación, pudiendo llegar a ser en torno a 2/3 del coste total de producción por lechón. Las canales se pagan de acuerdo a su contenido en carne, por lo que se estima este contenido midiendo en animal vivo la grasa dorsal, que supone un 70% de la grasa total, mediante aparatos de ultrasonidos. Deseamos seleccionar, pues, a favor de índice de conversión (IC) y contra contenido en grasa (GD), pero el consumo de pienso individual es caro de medir y se utilizará la velocidad de crecimiento (VC) para hacer una estimación indirecta, pues esta última medida sólo supone realizar una o dos pesadas del animal. Sabemos que un ahorro en 0.1 puntos de índice conversión implica un ahorro de 3 € por animal, y que una reducción en 1mm de espesor de capa de grasa dorsal implica un beneficio de 1€ por canal vendida. La siguiente matriz muestra las heredabilidades (diagonal) y las correlaciones genéticas (encima de la diagonal) y fenotípicas (bajo la diagonal) de los caracteres. Se ofrecen también las medias y desviaciones fenotípicas.

	<i>IC</i>	<i>VC</i>	<i>GD</i>		<i>m</i>	σ_P
<i>IC</i>	$\begin{bmatrix} 0.2 & -0.7 & 0.4 \\ -0.6 & 0.3 & 0.1 \\ 0.4 & 0.2 & 0.5 \end{bmatrix}$			<i>IC</i>	2.5	0.25
<i>VC</i>				<i>VC</i>	700	72 g/día
<i>GD</i>				<i>GD</i>	15	3 mm

Calcular el índice de selección que estima el valor aditivo económico.

Resolución: De estas tablas se deduce inmediatamente

$$\text{var}(y_{IC}) = 0.25^2 = 0.063 \text{ (kg pienso / kg peso vivo)}^2;$$

$$\text{var}(y_{VC}) = 72^2 = 5184 \text{ (g/día)}^2;$$

$$\text{var}(y_{GD}) = 9 \text{ mm}^2$$

$$\sigma_{A_{GD}}^2 = h_{GD}^2 \cdot \sigma_{P_{GD}}^2 = 0.5 \cdot 9 = 4.5 \text{ mm}^2;$$

$$\sigma_{A_{VC}}^2 = 0.3 \cdot 5184 = 1555 \text{ (g/día)}^2;$$

$$\sigma_{A_{IC}}^2 = 0.2 \cdot 0.063 = 0.013 \text{ (kg pienso / kg peso vivo)}^2$$

$$\text{cov}(A_{IC}, A_{VC}) = r_A(IC, VC) \cdot \sigma_{A_{IC}} \sigma_{A_{VC}} = -0.7 \sqrt{0.013 \cdot 1555} = -3.1 \text{ (kg p/kg pv) \cdot (g/día)}$$

$$\text{cov}(A_{IC}, A_{GD}) = 0.4 \sqrt{0.013 \cdot 4.5} = 0.10 \text{ (kg p/kg pv) \cdot mm}$$

$$\text{cov}(A_{GD}, A_{VC}) = 0.1 \sqrt{4.5 \cdot 1555} = 8.4 \text{ mm \cdot g/día}$$

$$\text{cov}(P_{GD}, P_{VC}) = r_P(GD, VC) \cdot \sigma_{P_{GD}} \cdot \sigma_{P_{VC}} = 0.2 \cdot 3 \cdot 72 = 43 \text{ mm \cdot g/día}$$

$$\text{cov}(P_{IC}, P_{VC}) = -11 \text{ (kg p/kg pv) \cdot (g/día)}$$

$$\text{cov}(P_{IC}, P_{GD}) = 0.30 \text{ (kg p/kg pv)} \cdot \text{mm}$$

$$A_E = a_{IC} \cdot A_{IC} + a_{GD} \cdot A_{GD}$$

$$a_{IC} = \frac{3 \text{ €}}{-0.1 \text{ (kg p/kg pv)}} = -30 \frac{\text{€}}{\text{kg p/kg pv}}$$

$$a_{GD} = -1 \frac{\text{€}}{\text{mm}}$$

$$\mathbf{a}' = [a_{IC}, a_{GD}] = [-30 \quad -1]$$

$$\mathbf{u}' = [A_{IC}, A_{GD}]$$

$$\mathbf{y}' = [y_{VC}, y_{GD}] \text{ donde los valores de } y \text{ están } \underline{\text{centrados}}$$

$$\mathbf{C}' = \text{cov}(\mathbf{u}, \mathbf{y}') =$$

$$= \text{cov} \left(\begin{bmatrix} A_{IC} \\ A_{GD} \end{bmatrix}, [y_{VC} \quad y_{GD}] \right) = \begin{bmatrix} \text{cov}(A_{IC}, A_{VC}) & \text{cov}(A_{IC}, A_{GD}) \\ \text{cov}(A_{GD}, A_{VC}) & \text{var}(A_{GD}) \end{bmatrix} = \begin{bmatrix} -3.1 & 0.1 \\ 8.4 & 4.5 \end{bmatrix}$$

$$\mathbf{V} = \text{var}(\mathbf{y}) = \begin{bmatrix} \text{var}(y_{VC}) & \text{cov}(y_{VC}, y_{GD}) \\ \text{cov}(y_{VC}, y_{GD}) & \text{var}(y_{GD}) \end{bmatrix} = \begin{bmatrix} 5184 & 43 \\ 43 & 9 \end{bmatrix}; \quad \mathbf{V}^{-1} = \begin{bmatrix} 2.0 & -9.6 \\ -9.6 & 1157 \end{bmatrix} 10^{-4}$$

$$\hat{A}_E = \mathbf{I} = [-30 \quad -1] \begin{bmatrix} -3.1 & 0.1 \\ 8.4 & 4.5 \end{bmatrix} \begin{bmatrix} 2.0 & -9.6 \\ -9.6 & 1157 \end{bmatrix} \begin{bmatrix} y_{VC} \\ y_{GD} \end{bmatrix} 10^{-4} = 0.0242 y_{VC} - 0.949 y_{GD}$$

Por ejemplo, un individuo que haya crecido 770 g/día y tenga un espesor de grasa dorsal de 12mm tiene un valor económico aditivo de (recuérdese que los valores de \mathbf{y} están *centrados*):

$$\hat{A}_E = 0.0242 (770 - 700) - 0.949 (12 - 15) = 4.54 \text{ €}$$

EJEMPLO 3.10

Enunciado: Podríamos haber utilizado datos de parientes en la estimación, y de hecho en los programas modernos de mejora se usan los datos de toda la población, presentes y pasados, como indicamos al hablar de índices para un carácter. Calcular un índice como el del ejemplo 3.9 usando los datos del individuo y de un medio hermano.

Resolución:

$$A_E = a_{IC} \cdot A_{IC} + a_{GD} \cdot A_{GD}$$

$$\mathbf{a}' = [a_{IC}, a_{GD}] = [-30 \quad -1]$$

$$\mathbf{u}' = [A_{IC}, A_{GD}]$$

$$\mathbf{y}' = [y_{VC}, y_{GD}, y_{VC}^{HS}, y_{GD}^{HS}] \quad \text{donde los valores de } y \text{ estan } \underline{\text{centrados}}$$

$$\begin{aligned} \mathbf{C}' = \text{cov}(\mathbf{u}, \mathbf{y}') &= \text{cov}\left(\begin{bmatrix} A_{IC} \\ A_{GD} \end{bmatrix}, \begin{bmatrix} y_{VC} & y_{GD} & y_{VC}^{HS} & y_{GD}^{HS} \end{bmatrix}\right) = \\ &= \begin{bmatrix} \text{cov}(A_{IC}, A_{VC}) & \text{cov}(A_{IC}, A_{GD}) & (1/4)\text{cov}(A_{IC}, A_{VC}) & (1/4)\text{cov}(A_{IC}, A_{GD}) \\ \text{cov}(A_{GD}, A_{VC}) & \text{var}(A_{GD}) & (1/4)\text{cov}(A_{GD}, A_{VC}) & (1/4)\text{var}(A_{GD}) \end{bmatrix} = \\ &= \begin{bmatrix} -3.1 & 0.1 & -0.78 & 0.025 \\ 8.4 & 4.5 & 2.1 & 1.1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \mathbf{V} = \text{var}(\mathbf{y}) &= \begin{bmatrix} \text{var}(y_{VC}) & \text{cov}(y_{VC}, y_{GD}) & (1/4)\text{var}(A_{VC}) & (1/4)\text{cov}(A_{GD}, A_{VC}) \\ & \text{var}(y_{GD}) & (1/4)\text{cov}(A_{GD}, A_{VC}) & (1/4)\text{var}(A_{GD}) \\ & & \text{var}(y_{VC}) & \text{cov}(y_{GD}, y_{VC}) \\ & & & \text{var}(y_{GD}) \end{bmatrix} = \\ &= \begin{bmatrix} 5184 & 43 & 389 & 2.1 \\ & 9 & 2.1 & 1.1 \\ & & 5184 & 43 \\ & & & 9 \end{bmatrix} \end{aligned}$$

$$\hat{A}_E = l = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{y} = 0.0238 y_{VC} - 0.934 y_{GD} + 0.0036 y_{VC}^{HS} - 0.114 y_{GD}^{HS}$$

Observese que los datos del medio hermano tienen mucho menor peso que los del individuo, y que este peso no solo depende del parentesco sino tambien de los parametros geneticos.

Precision y prediccion de la respuesta

$$\sigma_l^2 = \text{var}(\mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{y}) = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \text{var}(\mathbf{y}) \mathbf{C} \mathbf{V}^{-1} \mathbf{a} = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{V} \mathbf{V}^{-1} \mathbf{C} \mathbf{a} = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{C} \mathbf{a}$$

$$A_E = \mathbf{a}' \mathbf{u} = \mathbf{u}' \mathbf{a} \quad ; \quad l = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{y}$$

$$\text{cov}(l, A_E) = \text{cov}(\mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{y}, \mathbf{u}' \mathbf{a}) = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \text{cov}(\mathbf{y}, \mathbf{u}') \mathbf{a} = \mathbf{a}' \mathbf{C}' \mathbf{V}^{-1} \mathbf{C} \mathbf{a} = \sigma_l^2$$

$$\sigma_{A_E}^2 = a_1^2 \sigma_{A_1}^2 + a_2^2 \sigma_{A_2}^2 + \dots + a_n^2 \sigma_{A_n}^2 + 2a_1 a_2 \text{cov}(A_1, A_2) + \dots + 2a_{n-1} a_n \text{cov}(A_{n-1}, A_n)$$

$$\rho_{A_E, I} = \frac{\text{cov}(I, A_E)}{\sigma_{A_E} \sigma_I} = \frac{\sigma_I}{\sigma_{A_E}}$$

$$\sigma_e^2 = \sigma_{A_E}^2 - \sigma_I^2$$

$$R = \rho_{A_E, I} \cdot i \cdot \sigma_{A_E} = i \cdot \sigma_I$$

La respuesta para cada carácter es

$$R_i = \frac{\text{cov}(I, u_i)}{\sigma_{A_i} \sigma_I} \cdot i \cdot \sigma_{A_i} = \text{cov}(I, u_i) \cdot \frac{i}{\sigma_I} = \text{cov}(\mathbf{a}'\mathbf{C}'\mathbf{V}^{-1}\mathbf{y}, u_i) \cdot \frac{i}{\sigma_I} = \mathbf{a}'\mathbf{C}'\mathbf{V}^{-1} \text{cov}(\mathbf{y}, u_i) \cdot \frac{i}{\sigma_I}$$

el vector de respuestas es

$$\mathbf{R}' = \mathbf{a}'\mathbf{C}'\mathbf{V}^{-1}\mathbf{C} \cdot \frac{i}{\sigma_I}$$

EJEMPLO 3.11

Enunciado: Calcular la Respuesta esperada y la precisión del índice del ejemplo 3.9, para una intensidad de selección de 1.28.

Resolución: Aplicando las fórmulas,

$$\sigma_I^2 = 9.3 \text{ €}^2 ; \quad \sigma_I = 3.05 \text{ €}$$

$$\begin{aligned} \sigma_{A_E}^2 &= \mathbf{a}_{IC}^2 \sigma_{A_{IC}}^2 + \mathbf{a}_{GD}^2 \sigma_{A_{GD}}^2 + 2 \mathbf{a}_{IC} \mathbf{a}_{GD} \text{cov}(A_{IC}, A_{GD}) = \\ &= (-30)^2 0.013 + (-3)^2 4.5 + 2 (-30) (-3) 0.10 = 70 \text{ €}^2 ; \end{aligned}$$

$$\sigma_{A_E} = 8.4 \text{ €}$$

$$\sigma_e^2 = 70 - 9.3 = 60.7 \text{ €}^2 ; \quad \sigma_e = 7.8 \text{ €}$$

$$\rho_{A_E, I} = 3.05 / 8.4 = 0.36$$

$$R = 1.28 \cdot 3.05 = 3.9 \text{ € por generación}$$

$$\mathbf{R}' = [-0.17 \quad -4.1] \cdot \frac{1.28}{3.05} = [-0.07 \quad -1.7] ; \text{ esto es,}$$

$R_{IC} = -0.07$ (kg p/kg pv) por generación

$R_{GD} = -1.7$ mm por generación

Por selección individual habrían mejorado -0.06 (kg p/kg pv) y -1.9 mm por generación. Aquí la grasa se ve perjudicada porque la presión de selección es menor sobre ese carácter cuando se consideran los dos caracteres; esto es, hay individuos que no habrían sido seleccionados si el criterio fuera exclusivamente grasa dorsal, pero que lo son debido a que tienen buen índice de conversión estimado. Por el contrario, el índice de conversión, al que también le sucede lo mismo, compensa este efecto a través de la información que suministra el carácter grasa dorsal, que tiene una heredabilidad elevada y una correlación no muy baja con el índice de conversión.

Un error en los pesos económicos infraestimando un 25% el de el índice de conversión y sobreestimando un 25% el de grasa dorsal; esto es:

$$\mathbf{a}' = [-22.5 \quad -1.25]$$

daría lugar a una varianza del índice de 8.8 €^2 y a una respuesta de 3.8 € por generación.

Un error más grave

$$\mathbf{a}' = [-15 \quad -2]$$

da una respuesta de 4.7 € por generación, lo que supone una sobreestimación de la respuesta de un 20%. Un error semejante en los parámetros genéticos, en el que las covarianzas genéticas son respectivamente infraestimadas y sobreestimadas,

$$\mathbf{C}' = \begin{bmatrix} -3.1 & 0.2 \\ 4.2 & 4.5 \end{bmatrix}$$

da lugar a una varianza del índice de 16.1 , y a una respuesta de 5.1 € por generación, lo que supone una sobreestimación de la respuesta en un 30%.

- 1) Es posible que haya individuos sin datos (por ejemplo, en el caso del tamaño de camada o la producción de leche los machos no tienen datos), lo que no impide la estimación del valor genético de estos individuos con los datos disponibles. Simplemente no figuran en el vector \mathbf{y} , pero su parentesco sí que es considerado en \mathbf{c}' y \mathbf{V} , como se ve en el ejemplo 3.8.
- 2) La inversión de \mathbf{V} es difícil si el número de individuos es grande, por lo que se ha recurrido a resolver los índices usando otro tipo de ecuaciones que se explican en el siguiente apartado, al hablar del BLUP.
- 3) Para construir un índice hace falta conocer la heredabilidad del carácter, y en el caso de índices para varios caracteres hace falta conocer las correlaciones genéticas y fenotípicas entre los caracteres. En la práctica se sustituye el valor de estos parámetros genéticos por estimaciones realizadas con los propios datos o por estimaciones encontradas en la literatura, pero el error de estimación que se comete al estimar estos parámetros no es tenido en cuenta cuando se calcula la varianza del error o el intervalo de confianza, por lo que en realidad la varianza del error y el intervalo de confianza que se calculan son menores que los reales. Esta dificultad sólo puede resolverse aplicando la teoría Bayesiana (Apéndice II), de lo contrario hay que tratar a los parámetros genéticos como si fueran los verdaderos.
- 4) Los pesos económicos se miden en (€/unidad del carácter), por ejemplo €/g ó €/mm, por lo que el valor aditivo económico se mide en €. El valor aditivo económico estimado es el criterio que se usa para seleccionar, por tanto se seleccionan aquellos animales *cuyos hijos* darán un beneficio (en €) máximo.
- 5) El índice tiene en cuenta las relaciones genéticas entre caracteres para optimizar el beneficio. Es decir, no necesariamente mejora cada uno de los caracteres (por ejemplo, si la correlación entre los caracteres es negativa podría empeorar alguno de ellos). Obsérvese que se seleccionan los animales que transmiten a la descendencia un valor económico máximo; un mismo valor aditivo económico puede lograrse con un valor aditivo pobre para GD pero alto para IC o mediante valores aditivos intermedios para ambos caracteres, en cualquier caso se seleccionan aquellos individuos de los que se espera que su descendencia de lugar a un rendimiento económico máximo.
- 6) No podemos implementar muchos caracteres en un índice (no más de tres o cuatro). La razón es compleja de describir y está relacionada con los errores de estimación. Por azar

pueden aparecer estimas de parámetros genéticos incoherentes. Por ejemplo, si estimamos la heredabilidad mediante correlaciones entre medios hermanos y por azar los individuos no emparentados se parecieran más entre sí que los medios hermanos, tendríamos estimaciones de la heredabilidad negativas. Esto será más probable que ocurra cuantos más caracteres consideremos. De igual modo es posible que aparezcan por azar relaciones entre los caracteres incoherentes, y asimismo será más frecuente cuantos más caracteres incluyamos en el índice (Hill y Thompson, 1978). Aunque se han propuesto algunas soluciones parciales o aproximadas al problema (“bending”, Hayes y Hill, 1986, por ejemplo), los beneficios marginales de incluir muchos caracteres en el índice no suelen ser grandes, por lo que hay que extremar las precauciones.

7) Los índices son *relativamente* robustos ante los errores en los pesos económicos. Errores del 200 y del 300% no suelen afectar mucho al resultado final (Smith 1983), como se puede comprobar en nuestro ejemplo. Sin embargo a medio o largo plazo puede que las predicciones no sean lineales; es decir, no se obtenga el mismo beneficio al seleccionar una unidad del carácter. Este problema no es sencillo de resolver, y se suele recomendar recalcularse los pesos económicos con el paso del tiempo. Por otra parte hay empresas que calculan los pesos económicos de manera más compleja; por ejemplo, teniendo en cuenta su cuota de mercado y la cuota que aspiran obtener en competencia con las demás compañías. La existencia de limitaciones a la producción (cuotas lecheras, por ejemplo), añade nuevas complicaciones. Por último, los intereses de los integrantes de la cadena que va de la producción al consumo no siempre coinciden; por ejemplo, al ganadero le interesa lo que le paga el matadero por sus canales, pero al carnicero le interesa el porcentaje de cortes caros de la canal y al consumidor la calidad de la carne. Estos intereses están con frecuencia contrapuestos y es difícil integrarlos en programas de mejora. Blasco (1996) ofrece una extensa revisión de estos y otros problemas relacionados con la estimación de los pesos económicos.

8) No ocurre lo mismo con los errores en los parámetros genéticos (Meyer y Hill, 1983), particularmente en el caso de las correlaciones genéticas, y los índices son sensibles a errores en la estimación de estas correlaciones. El problema es por un lado que hace falta una cantidad considerable de datos para estimar una correlación genética con precisión (más de 1.000, por ejemplo) y por otro que los parámetros genéticos cambian con la selección y deberían ser recalculados en cada generación, sin datos suficientes para una precisión razonable en muchas ocasiones (estas son también razones para no incluir muchos caracteres en el índice). Cuando se dispone de todos los datos usados en la selección y de las relaciones entre parientes completas, no es necesario el recálculo de los parámetros genéticos porque

serven los de la generación base, o pueden usarse todos ellos para estimar los parámetros de la generación base mejorando la precisión.

9) Los índices necesitan que los datos estén centrados. Naturalmente esto no ocurre en los programas reales: los animales crecen más en invierno, crecen menos si provienen de camadas numerosas, y según en la granja en la que estén se les prodigan mejores o peores cuidados que afectan también al crecimiento o a otros caracteres productivos. Hay varios procedimientos más o menos artesanales para centrar los datos: comparaciones entre contemporáneos, o entre individuos de la misma estación, interpolaciones para que los datos se ajusten a un estándar común, etc. En el apartado siguiente, al hablar del BLUP, se aborda formalmente la corrección de estos datos.

3.4. Modelos lineales. El BLUP

3.4.1. BLUP

Al deducir los índices hemos puesto como condición que los valores estuvieran *centrados*. Si los datos no están centrados no podemos proceder como hicimos en los índices de selección, porque entonces $E(\mathbf{yy}')$ ya no es la matriz de varianzas covarianzas fenotípica ni $E(\mathbf{Ay}')$ es un vector de covarianzas. En 1949 Henderson propuso resolver el problema utilizando una solución máximo verosímil para estimar simultáneamente las medias y los valores genéticos, lo que resultó no ser el caso. El método quedó olvidado hasta que Searle, un estadístico alumno y colaborador de Henderson le encontró propiedades interesantes. Henderson demostró en 1963 que se podía expresar la solución en forma operativamente cómoda, puesto que, como indicamos en el caso de los índices de selección, la inversión de \mathbf{V} es difícil si no imposible cuando hay muchos individuos a evaluar. A esta forma de resolver el problema se le conoce como “ecuaciones del modelo mixto” (MME en inglés), y es la forma habitual de resolver el BLUP hoy en día⁽¹⁷⁾. De todas formas estas ecuaciones requieren la inversión de la matriz de parentesco \mathbf{G} , algo también complejo y que limitó el desarrollo del BLUP hasta que Henderson (1976) descubrió que sorprendentemente se

¹⁷ El nombre BLUP aparece por primera vez en un artículo de Goldberg (1962), quien dedujo el BLUP de forma independiente a Henderson y en un contexto alejado de la genética. Algunos autores pretenden dar la primacía del BLUP a Goldberg, pero las ecuaciones del modelo mixto son muy anteriores, y es obvio que Henderson desconocía el trabajo de Goldberg cuando mostró en 1963 la equivalencia de sus ecuaciones y el BLUP.

podía calcular la inversa de \mathbf{G} directamente y de forma sencilla. Desde entonces es el método más utilizado en mejora genética animal, y empieza a imponerse en mejora de plantas. Las iniciales BLUP corresponden a *Best Linear Unbiased Prediction*, el mejor de los predictores lineales insesgados, y derivan de sus propiedades, que veremos a continuación. Ronningen (1971) y Dempfle (1977) llamaron la atención sobre el hecho de que el BLUP podría considerarse como un estimador bayesiano, y Dempfle (1977) hizo notar que podría también considerarse como una mezcla de estimadores en los que uno extrae la información de los datos y otro la información “a priori” que sobre la población se dispone (la media y la matriz \mathbf{G}). Finalmente, en una extensa revisión, Blasco (2001) discute el BLUP como estimador frecuentista y bayesiano.

El BLUP como índice corregido

Una solución obvia al problema de que los datos no estén centrados es estimar las medias y centrar los datos. Cuando los datos tienen diferentes medias, los datos pueden representarse mediante el modelo

$$\mathbf{y} = \mathbf{m} + \mathbf{e}$$

donde \mathbf{m} es el vector de medias de los datos; esto es, $E(\mathbf{y}) = \mathbf{m}$. Como varios grupos de datos tienen la misma media (por ejemplo, los que nacieron en la misma estación), el modelo se representa como

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

donde \mathbf{b} contiene las medias comunes a varios individuos (los efectos de estación, por ejemplo) y \mathbf{X} es una matriz de incidencia; esto es, de unos y ceros indicando la presencia o ausencia de un efecto para un individuo concreto. En el caso de que hayan covariables, una columna de \mathbf{X} contiene los valores de la covariable para cada individuo. Los errores se considera que tienen media cero y están incorrelacionados.

EJEMPLO 3.12

En la tabla siguiente se indican los datos de tamaño de camada obtenidos por dos conejas en dos estaciones distintas, y el peso de ambas

	CONEJA 1	CONEJA 2

	INVIERNO	PRIMAVERA	INVIERNO	PRIMAVERA
PARTO 1	12		9	
PARTO 2		7		11
PARTO 3				8

Llamando $E1$ y $E2$ a los efectos de estación, $P1$, $P2$, $P3$ a los efectos de parto,

$$\left. \begin{array}{l} 12 = E1 + P1 + e1 \\ 7 = E2 + P2 + e2 \\ 9 = E1 + P1 + e3 \\ 11 = E2 + P2 + e4 \\ 8 = E2 + P3 + e5 \end{array} \right\} \begin{array}{l} \left[\begin{array}{c} 12 \\ 7 \\ 9 \\ 11 \\ 8 \end{array} \right] = \left[\begin{array}{ccccc} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array} \right] \cdot \left[\begin{array}{c} E1 \\ E2 \\ P1 \\ P2 \\ P3 \end{array} \right] + \left[\begin{array}{c} e1 \\ e2 \\ e3 \\ e4 \\ e5 \end{array} \right] \\ \mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{e} \end{array}$$

Con esta notación, $E(\mathbf{y}) = \mathbf{Xb}$. Una primera idea para centrar los datos puede ser estimar las medias \mathbf{b} por mínimos cuadrados, restárselas a \mathbf{y} creando un nuevo vector de datos corregido $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} = \hat{\mathbf{e}}$, y aplicar la fórmula anterior al vector \mathbf{y}^* , puesto que los estimadores mínimo cuadráticos son insesgados y cumplen que $E(\hat{\mathbf{b}}) = \mathbf{b}$, con lo que

$$E(\mathbf{y}^*) = E(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = E(\mathbf{y}) - \mathbf{X}E(\hat{\mathbf{b}}) = \mathbf{Xb} - \mathbf{Xb} = \mathbf{0}$$

y el vector \mathbf{y}^* estaría centrado. El estimador mínimo cuadrático es

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (3.14)$$

Frecuentemente $\mathbf{X}'\mathbf{X}$ no tiene inversa, y hay que añadir alguna condición arbitraria al sistema de ecuaciones para que la nueva matriz $\mathbf{X}'\mathbf{X}$ tenga inversa, por ejemplo se obliga a que los efectos de parto sumen cero, o se da un valor de cero a uno de ellos y se refieren los demás efectos a ese efecto nulo. Así, en el ejemplo 3.12, para indicar que los efectos de parto suman cero se añade un cero a la columna de datos y a la de errores, y a la matriz \mathbf{X} se le añade la fila $[0 \ 0 \ 1 \ 1 \ 1]$.

El modelo es, sin embargo, incorrecto, puesto que los individuos están emparentados, los datos \mathbf{y} correlacionados, y por tanto los errores correlacionados. Para tener en cuenta el que los datos están correlacionados, el modelo que se utiliza es

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \quad (3.15)$$

donde \mathbf{u} contiene los valores aditivos de los individuos (tengan datos o no) y \mathbf{Z} es la matriz de diseño correspondiente. En el caso de que hubiera individuos sin datos (por ejemplo, un macho), se añade una columna de ceros en la matriz \mathbf{Z} , y el valor aditivo del macho en el vector \mathbf{u} . A este modelo se le conoce como "modelo mixto", puesto que contiene tanto efectos fijos como aleatorios. Cuando \mathbf{u} incluye los valores aditivos de cada uno de los individuos (y no solamente el de los machos, por ejemplo), al modelo mixto se le llama también "modelo animal" o "modelo planta".

EJEMPLO 3.13

Con el ejemplo 3.12, se desea plantear las ecuaciones del modelo mixto para las dos conejas y para un macho, considerando sólo el efecto de estación. Llamando u_1 , u_2 a los valores aditivos de las conejas, y u_3 al valor aditivo del macho, el modelo es ahora

$$\left. \begin{array}{l} 12 = E1 + u1 + e1 \\ 7 = E2 + u1 + e2 \\ 9 = E1 + u2 + e3 \\ 11 = E2 + u2 + e4 \\ 8 = E2 + u2 + e5 \end{array} \right\} \begin{array}{l} \left[\begin{array}{c} 12 \\ 7 \\ 9 \\ 11 \\ 8 \end{array} \right] = \left[\begin{array}{ccccc} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array} \right] \cdot \left[\begin{array}{c} E1 \\ E2 \\ P1 \\ P2 \\ P3 \end{array} \right] + \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{array} \right] \cdot \left[\begin{array}{c} u1 \\ u2 \\ u3 \end{array} \right] + \left[\begin{array}{c} e1 \\ e2 \\ e3 \\ e4 \\ e5 \end{array} \right] \\ \mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{Z} \mathbf{u} + \mathbf{e}
 \end{array}$$

En este modelo, por conveniencia de cálculo, los efectos aleatorios se refieren a la media; esto es, la media de los efectos aleatorios es cero.

$$E(\mathbf{y}) = \mathbf{X}\mathbf{b} \quad ; \quad E(\mathbf{u}) = \mathbf{0} \quad ; \quad \text{var}(\mathbf{u}) = \mathbf{G} \quad ; \quad \text{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}' \text{var}(\mathbf{u}) \mathbf{Z} + \text{var}(\mathbf{e}) = \mathbf{Z}' \mathbf{G} \mathbf{Z} + \mathbf{I}\sigma_e^2$$

Con este modelo, la estima por mínimos cuadrados generalizados de \mathbf{b} es

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

en donde si $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$ no tiene inversa se actúa como en el caso del estimador por mínimos cuadrados (3.14). El vector $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$ está centrado, puesto que se comprueba de forma análoga que $E(\hat{\mathbf{b}}) = \mathbf{b}$. El BLUP es, entonces

$$\hat{u} = \mathbf{c}'\mathbf{V}^{-1}\mathbf{y}^* = \mathbf{c}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

esto es, un índice al que se le han corregido los datos centrándolos apropiadamente. Henderson (1963) demostró que este estimador era el mejor entre los lineales insesgados, como veremos más adelante.

Propiedades del BLUP

El BLUP es un índice de selección con los datos corregidos, por lo que las propiedades de los índices de selección son aplicables al BLUP. En realidad las propiedades de los índices son válidas en tanto en cuanto los datos que se utilicen estén bien centrados; esto es, que hayan sido corregidos apropiadamente.

El BLUP y la selección

Una de las condiciones para aplicar del modelo mixto (ecuación 3.15) es que la esperanza de los efectos aleatorios $E(\mathbf{u})$ sea nula. Obviamente, si hay selección, esto no es cierto, puesto que los individuos de las últimas generaciones tendrán por término medio valores superiores a los de las primeras. Esto implica que no se podría aplicar el BLUP a todas las generaciones de selección disponibles. Sin embargo, Henderson (1975) demostró que con ciertas condiciones, que en esencia consisten en:

- 1) Analizar los datos con la matriz de parentesco completa.
- 2) Utilizar en el análisis todos los datos usados para la selección.
- 3) Usar en el análisis el mismo modelo que se utilizó al seleccionar (es decir, si se usan efectos de estación en el momento de la selección, estos efectos se tienen en cuenta en el análisis de datos)⁽¹⁸⁾.
- 4) Utilizar los parámetros genéticos correctos. Los parámetros genéticos cambian de generación en generación como resultado de la selección, pero aquí se requiere utilizar en el BLUP los parámetros genéticos de la generación base, antes de seleccionar.

¹⁸ Esta condición no es en realidad tan restrictiva, pero no podemos entrar en detalles sin alcanzar una complejidad intolerable para los objetivos de este capítulo.

entonces puede ignorarse el efecto de la selección y aplicar el BLUP como si los datos no estuvieran seleccionados, puesto que las estimas que se obtendrán serán a su vez estimas BLUP correctas. Esto permite obtener los valores aditivos de los individuos generación tras generación, libres de los factores ambientales, y por tanto permite monitorizar el proceso de selección sin necesidad de una población control. La Respuesta a la selección puede estimarse simplemente como la media de los valores aditivos de los individuos de cada generación. Sin embargo, Thompson (1986) y Sorensen y Johansen (1992) han llamado la atención sobre la extremada sensibilidad de esta Respuesta a los parámetros genéticos que se utilizan para calcular el BLUP, por lo que si la condición 4 no se cumple, la Respuesta estimada no será la correcta.

Dos de los problemas habituales en la estimación de la Respuesta son tenidos en cuenta, sin embargo, por el BLUP, como demuestran Sorensen y Kenedy (1985). En primer lugar las estimas BLUP tienen en cuenta la reducción de la varianza debida a selección (efecto Bulmer), y por otra parte los errores típicos de las estimas BLUP tienen en cuenta el efecto de la deriva genética, cuando se dispone de matrices de parentesco completas. Con el modelo infinitesimal no hay fijación ni pérdida de genes, sólo se alteran sus frecuencias, y la única causa de esta alteración debida a la deriva genética es el aumento del parentesco con la selección, lo que queda recogido en la evaluación con la matriz de parentesco completa.

3.4.2. LAS ECUACIONES DEL MODELO MIXTO

Invertir V no es sencillo, por lo que Henderson (1963) derivó unas ecuaciones para obtener \hat{b} y \hat{u} que simplificaban el problema, puesto que en ellas no era necesario invertir la matriz de varianzas covarianzas fenotípica. A esas ecuaciones se les conoce como Ecuaciones del Modelo Mixto. En realidad Henderson derivó en 1949 las ecuaciones del modelo mixto creyendo que daban lugar a estimas máximo verosímiles de los efectos aditivos, lo que como hemos comentado antes resultó no ser el caso, pero posteriormente Henderson (1963) demostró que daban la misma solución para \hat{b} y para \hat{u} que las estimas BLUP. No es posible encontrar estimas máximo verosímiles de los efectos aleatorios. Para hallar una estima máximo verosímil de un parámetro se toma la función de densidad de probabilidad de los datos y , que es función de los valores de ese parámetro, y se fijan los valores del parámetro hasta encontrar el valor que maximiza la probabilidad de los datos. El problema es que al ser los valores aditivos unos efectos aleatorios, no pueden “fijarse”,

puesto que entonces dejan de ser aleatorios, se estimarían exactamente igual que los efectos fijos y no se tendría en cuenta las correlaciones entre ellos.

Ecuaciones del modelo mixto

La forma más sencilla de deducir las ecuaciones del modelo mixto es aplicar el argumento original de Henderson (1949), interpretándolo correctamente. Henderson(1949) quería hallar una estima máximo verosímil de los valores aditivos \mathbf{u} . La verosimilitud se representa mediante la función $f(\mathbf{y}|\mathbf{u})$, que indica la densidad de probabilidades de la muestra \mathbf{y} *dado un valor de \mathbf{u} concreto*. Según sea el valor de \mathbf{u} , la muestra tiene una probabilidad mayor o menor. Las estimas máximo verosímiles son aquéllas que, si fueran el verdadero valor, harían máxima la probabilidad de encontrar los datos \mathbf{y} . Parece lógico multiplicar los valores de $f(\mathbf{y}|\mathbf{u})$ por la probabilidad de que, efectivamente, la \mathbf{u} que figura en $f(\mathbf{y}|\mathbf{u})$ sea la verdadera. Esto es, se trataría de encontrar el valor de \mathbf{u} que hace máximo

$$\varphi = f(\mathbf{y}|\mathbf{u}) \cdot f(\mathbf{u})$$

Para hacerlo supondremos que tanto los datos como los valores aditivos se distribuyen de forma Normal, y que los efectos \mathbf{b} son “fijos”; esto es, la función de densidad de \mathbf{y} se escribe correctamente $f(\mathbf{y}|\mathbf{b},\mathbf{u})$, cuya media es $E(\mathbf{y}) = \mathbf{Xb} + \mathbf{Zu}$, y cuya varianza, considerando fijados a \mathbf{b} y a \mathbf{u} es $\text{var}(\mathbf{y}) = I\sigma_e^2$ (ya que si \mathbf{b} y \mathbf{u} están fijados, sólo varía el error \mathbf{e}). Por su parte, $f(\mathbf{u})$ tiene de media cero y de matriz de varianzas-covarianzas \mathbf{G} , matriz que incluye las relaciones de parentesco de toda la población. La expresión anterior es, entonces,

$$\begin{aligned} \varphi &\propto \exp [(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})' (1/\sigma_e^2) (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})] \cdot \exp(\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}) = \\ &= \exp [(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})' (1/\sigma_e^2) (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}] \end{aligned}$$

donde \propto significa “proporcional a”. Tomando logaritmos, derivando e igualando a cero para obtener el máximo,

$$(\partial \log \varphi / \partial \mathbf{b}) = 2 \mathbf{X}' (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) (1/\sigma_e^2) = 0$$

$$(\partial \log \varphi / \partial \mathbf{u}) = 2 \mathbf{Z}' (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) (1/\sigma_e^2) + 2 \mathbf{G}^{-1} \mathbf{u} = 0$$

La matriz \mathbf{G} suele expresarse como $\mathbf{G} = \mathbf{A}\sigma_A^2$ donde \mathbf{A} recoge el doble de los coeficientes de parentesco entre los individuos. Así, las ecuaciones quedan

$$\mathbf{X}'\mathbf{X} \hat{\mathbf{b}} + \mathbf{X}'\mathbf{Z} \hat{\mathbf{u}} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{Z}'\mathbf{Z} \hat{\mathbf{b}} + \mathbf{Z}'\mathbf{Z} \hat{\mathbf{u}} + \mathbf{A}^{-1}\hat{\mathbf{u}} (\sigma_e^2/\sigma_A^2) = \mathbf{Z}'\mathbf{y}$$

y en forma matricial, y llamando $\alpha = \sigma_e^2/\sigma_A^2$

$$\sigma_P^2 = \sigma_A^2 + \sigma_e^2 = h^2\sigma_P^2 + \sigma_e^2 \quad \longrightarrow \quad \alpha = \frac{\sigma_e^2}{\sigma_A^2} = \frac{\sigma_P^2(1-h^2)}{\sigma_P^2 \cdot h^2} = \frac{1-h^2}{h^2}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad (3.16)$$

A nadie se le escapa que invertir \mathbf{A} es difícil cuando hay un gran número de datos disponible. Henderson (1976) y Quaas (1976) encontraron métodos sencillos para calcular \mathbf{A}^{-1} directamente, y hoy en día hay programas estándar para ello.

EJEMPLO 3.14

Enunciado: Crear las ecuaciones del modelo mixto para el ejemplo 3.13 teniendo en cuenta sólo el efecto de estación. Las dos conejas son medio hermanas y el macho es hermano de la primera. La heredabilidad del tamaño de camada es 0.1 y la varianza fenotípica 9.

Resolución: Las ecuaciones del modelo lineal son

$$\left. \begin{array}{l} 12 = E1 + u1 + e1 \\ 7 = E2 + u1 + e2 \\ 9 = E1 + u2 + e3 \\ 11 = E2 + u2 + e4 \\ 8 = E2 + u2 + e5 \end{array} \right\} \begin{bmatrix} 12 \\ 7 \\ 9 \\ 11 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} E1 \\ E2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} u1 \\ u2 \\ u3 \end{bmatrix} + \begin{bmatrix} e1 \\ e2 \\ e3 \\ e4 \\ e5 \end{bmatrix}$$

$$\alpha = (1 - 0.1) / 0.1 = 9$$

La matriz de parentesco es, dado que las conejas son medio hermanas, y el macho es hermano de la primera,

$$\mathbf{A} = \begin{bmatrix} 1 & 0.25 & 0.5 \\ & 1 & 0.25 \\ & & 1 \end{bmatrix}; \quad \mathbf{A}^{-1} = \begin{bmatrix} 1.36 & -0.18 & -0.64 \\ & 1.09 & -0.18 \\ & & 1.36 \end{bmatrix}; \quad \mathbf{A}^{-1} \cdot \alpha = \begin{bmatrix} 12 & -1.6 & -5.7 \\ & 9.8 & -1.6 \\ & & 12 \end{bmatrix}$$

Las ecuaciones del modelo mixto son

$$\left[\begin{array}{cc|ccc} 2 & 0 & 1 & 1 & 0 \\ 0 & 3 & 1 & 2 & 0 \end{array} \right] \begin{bmatrix} E1 \\ E2 \\ u1 \\ u2 \\ u3 \end{bmatrix} = \begin{bmatrix} 21 \\ 26 \\ 19 \\ 28 \\ 0 \end{bmatrix};$$

$$\left[\begin{array}{ccc|ccc} 1 & 1 & & & & \\ 1 & 2 & & & & \\ 0 & 0 & & & & \end{array} \right] + \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & & & \\ 0 & 3 & 0 & & & \\ 0 & 0 & 0 & & & \end{array} \right] + \left[\begin{array}{ccc|ccc} 12 & -1.6 & -5.7 & & & \\ -1.6 & 9.8 & -1.6 & & & \\ -5.7 & -1.6 & 12 & & & \end{array} \right] \begin{bmatrix} u1 \\ u2 \\ u3 \end{bmatrix} = \begin{bmatrix} 19 \\ 28 \\ 0 \end{bmatrix};$$

$$\left[\begin{array}{ccccc|ccc} 2 & 0 & 1 & 1 & 0 & E1 & & & \\ 0 & 3 & 1 & 2 & 0 & E2 & & & \\ 1 & 1 & 14 & -1.6 & -5.7 & u1 & = & 19 & \\ 1 & 2 & -1.6 & 13 & -1.6 & u2 & = & 28 & \\ 0 & 0 & -5.7 & -1.6 & 12 & u3 & = & 0 & \end{array} \right]; \rightarrow \begin{bmatrix} E1 \\ E2 \\ u1 \\ u2 \\ u3 \end{bmatrix} = \begin{bmatrix} 10.5 \\ 8.7 \\ -0.012 \\ 0.011 \\ -0.004 \end{bmatrix}$$

Los valores de u_1 y u_2 se dan respecto a la media (recuérdese que $E(\mathbf{u}) = \mathbf{0}$), y los valores de los efectos de estación incluyen la media (¹⁹).

El error de estimación

Como vimos antes, y se detalla en el apéndice I, la varianza de los errores de estimación de los efectos fijos coincide con la varianza de estos efectos, mientras que la varianza de los errores de estimación de los efectos aleatorios es $\text{var}(\mathbf{u} - \hat{\mathbf{u}})$. La matriz de varianzas-covarianzas de los errores de estimación no la deduciremos (ver detalles, por ejemplo en Rico, 1999). Llamando

$$\mathbf{Q} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \equiv \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$$

a la matriz de coeficientes, la varianza de los errores de estimación es

$$\text{var} \begin{bmatrix} \hat{\mathbf{b}} \\ \mathbf{u} - \hat{\mathbf{u}} \end{bmatrix} = \mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}$$

donde \mathbf{C}^{ij} es la parte de la inversa de la matriz de coeficientes \mathbf{Q}^{-1} correspondiente a la partición \mathbf{C}_{ij} (nótese que \mathbf{C}^{ij} no es \mathbf{C}_{ij}^{-1} sino la parte de \mathbf{Q}^{-1} que corresponde a \mathbf{C}_{ij}).

¿Fijos o aleatorios?

¹⁹ Téngase en cuenta que los resultados no son exactos, puesto que se ha redondeado sucesivas veces para facilitar la reproducción del ejemplo por parte del lector

Si las estimas de efectos aleatorios aprovechan mejor la información, ¿porqué no estimar los efectos de rebaño como aleatorios? En primer lugar porque hace falta disponer de una matriz equivalente a la de parentesco para los efectos de rebaño, y además es necesario conocer las covarianzas entre efectos de rebaño y efectos aditivos, ya que al ser ahora todos los efectos aleatorios, la matriz de varianzas-covarianzas de los efectos aleatorios incluye todas las relaciones posibles.

Podría simplificarse el problema suponiendo que la varianza de rebaño es la misma para todos los rebaños, pero esto es notoriamente falso, ya que rebaños buenos muestran menos variación, presumiblemente por tener mejor controlado el ambiente. Podríamos intentar estimar en cada rebaño la varianza del efecto debido a rebaño, pero eso implicaría separar la varianza aditiva de la ambiental con los datos de cada rebaño, habitualmente insuficientes para obtener una precisión adecuada. Finalmente, los mejores rebaños suelen importar el mejor semen, por lo que hay covarianzas entre efectos de rebaño y efectos aditivos, difíciles de precisar. Este último hecho movió a Henderson (1973) a considerar al efecto de rebaño como fijo. Así desaparecen en la estimación las varianzas y covarianzas asociadas al rebaño.

OBSERVACIONES SOBRE EL BLUP

1) Siendo el BLUP un índice de selección corregido, son de aplicación las observaciones que se han realizado acerca de los índices. Por ejemplo, como en caso de índices de selección, hay que conocer los verdaderos valores de \mathbf{c} y \mathbf{V} , de lo contrario la estimación no es BLUP. En conclusión, no es posible encontrar estimas BLUP sino aproximaciones mejores o peores según la precisión con la que se estimen las varianzas y covarianzas.

2) Los errores de estimación son difíciles de calcular, puesto que hay que invertir matrices muy grandes. Esto ha conducido a una aplicación indiscriminada del BLUP en la que no se ofrecen los errores de estimación con la excusa de su dificultad de cálculo. Estos errores pueden ser elevados si los valores aditivos se estiman con pocos datos o las conexiones entre efectos fijos y aleatorios es débil.

3) Obsérvese que el BLUP es el mejor de los estimadores pero sólo dentro de la clase de los insesgados. Estimadores sesgados pueden ser mejores que el BLUP, aunque no tenemos medios de obtenerlos de forma general. Cuando los efectos fijos se estiman con pocos datos, la escasa precisión de su estimación afecta también a los efectos aleatorios (ver ejemplo

3.15). Por ejemplo, en vacuno de leche es frecuente que ciertos datos provengan de pequeñas granjas. En ese caso si se consideran como efectos aleatorios las granjas con pocos datos tendrán un efecto menor que las granjas con más datos, lo que contribuye a aumentar la precisión. El problema es que al desconocerse las varianzas de granja y sus covarianzas con los efectos aditivos, las estimas están sesgadas. En ocasiones, sin embargo, es preferible el sesgo al exceso de imprecisión. Estos dilemas suelen resolverse mediante simulación.

4) Se ha exagerado a menudo la eficiencia del BLUP respecto a los índices de selección. En realidad el BLUP sólo es ligeramente más eficiente que un índice que tenga los datos bien corregidos (sea corrigiéndolos mediante mínimos cuadrados como en el ejemplo 3.12, o por métodos más aproximados como la comparación entre contemporáneos). Blasco et al. (1985) en conejos, Sorensen y Johansen (1995) en cerdos y Demfple (1980) en vacuno de leche, apenas encuentran ligeras ventajas del BLUP frente a los índices corregidos por métodos tradicionales. Las ventajas del BLUP provienen del enorme desarrollo de la computación en los últimos años, que permite resolver sistemas gigantescos de ecuaciones con rapidez y a bajo coste. La forma de almacenar los datos en el ordenador y la facilidad con la que se plantean y resuelven estos sistemas hace que el BLUP sea un instrumento cómodo de utilizar. Los resultados del BLUP permiten además monitorizar mejor la respuesta a la selección y la evolución de los efectos ambientales, siempre y cuando los parámetros genéticos que se introduzcan sean fiables.

5) No es infrecuente que en la bibliografía se encuentren respuestas a la selección estimadas con BLUP en las que se observa una tendencia fenotípica casi nula, una tendencia genética favorable y una tendencia ambiental desfavorable. No se insistirá nunca suficientemente en el hecho de que las tendencias dependen de los parámetros genéticos utilizados para calcularlas. Un genetista que no obtuviera resultados, podría camuflar su fracaso utilizando parámetros genéticos tales (que incluso podrían provenir de la bibliografía) que se observaran tendencias genéticas positivas y tendencias ambientales negativas. Es prudente, por tanto, utilizar parámetros genéticos que no muestren un deterioro del ambiente, a no ser que haya alguna razón para creer que el ambiente efectivamente se ha deteriorado.

6) No es necesario que los datos se distribuyan de forma Normal para aplicar el BLUP. Es cierto que algunas propiedades requieren Normalidad (al igual que en el caso de los índices), pero si los datos no se distribuyen de forma Normal, el BLUP es la mejor aproximación lineal a la estima del valor aditivo. En particular, la minimización del riesgo medio cuadrático y la maximización de la precisión no requieren la hipótesis de normalidad.

3.4.3. OTROS MODELOS

Modelo Padre

El modelo animal requiere la evaluación de todos los animales, tanto los que están presentes en las explotaciones como sus antecesores. Esto ocasiona tener que resolver gigantescos sistemas, que en el caso del vacuno de leche suponen varios millones de ecuaciones. Cuando, además, los modelos son complejos, como en el caso de los análisis de supervivencia o longevidad, para disminuir el número de ecuaciones la estimación se concentra en los machos, que son al fin y al cabo la clave del comercio de genes en el caso del vacuno de leche. Un modelo de padre se escribe exactamente igual y se resuelve exactamente igual, pero el vector u contiene solamente los efectos de macho, por lo que el dato de la hembra se supone explicado por los efectos fijos y por el efecto del padre, conteniendo el error tanto el efecto de la madre como los posibles efectos aleatorios ambientales no considerados. Para evitar sesgos debido a que no se considera el efecto de la madre, suele añadirse al modelo un efecto del padre de la madre, con lo que se espera corregir el posible sesgo de una vaca en la que no se ha considerado el valor genético de su madre para explicar su producción. Hay que considerar también que la varianza de los efectos aleatorios ya no es la varianza aditiva sino la varianza de padres, que **como vimos en el capítulo anterior** es $\frac{1}{4}$ de la varianza aditiva.

EJEMPLO 3.15

Para ver cómo se produce la estimación, pondremos un ejemplo sencillo. Disponemos de dos rebaños R_1 y R_2 , y de dos machos emparentados s_1 y s_2 que tienen dos hijas cada uno, cada una en un rebaño distinto. Los datos de lactación de sus hijas son:

	s_1	s_2
R_1	HIJA 1 y_1, y_2	HIJA 1 y_4
R_2	HIJA 2 y_3	HIJA 2 y_5

Las ecuaciones del modelo mixto dan lugar al sistema

$$3R_1 + 2s_1 + s_2 = y_1 + y_2 + y_4$$

$$2R_2 + s_1 + s_2 = y_3 + y_5$$

$$2R_1 + R_2 + (1+a_{11})s_1 + a_{12}s_2 = y_1 + y_2 + y_3$$

$$R_1 + R_2 + a_{12}s_1 + (1+a_{22})s_2 = y_4 + y_5$$

donde a_{ij} es el elemento ij de la matriz $A^{-1} \cdot \alpha$. Si resolvemos, despejando sale

$$R_1 = (y_1 + y_2 + y_4 - s_1 - s_2) / 3$$

$$R_2 = (y_3 + y_5 - s_1 - s_2) / 2$$

es decir, cada efecto de rebaño (efecto fijo) se estima a partir de los datos del propio rebaño, descontando - por decirlo así - los efectos genéticos. Así, si un rebaño tiene tendencia a traer semen de padres muy buenos, esto no afectará al efecto de rebaño.

$$s_1 = (y_1 + y_2 + y_3 - 2R_1 - R_2)(1+a_{11}) - [a_{12} / (1+a_{11})] s_2$$

$$s_2 = (y_4 + y_5 - R_1 - R_2)(1+a_{22}) - [a_{12} / (1+a_{22})] s_1$$

es decir, los individuos se estiman a partir de los datos de sus hijas descontando el efecto de rebaño, pero también intervienen en la estimación los parientes, y en la misma forma que intervenían en los índices de selección. Obsérvese que en este modelo sólo intervienen como parientes los machos, luego hay relaciones útiles de -parentesco que se desaprovechan, cosa que no ocurre en otros modelos. Si los machos no están emparentados, $a_{12} = 0$ y entonces no intervienen en la evaluación de otro animal.

Modelos con más de un efecto aleatorio

Modelo de Repetibilidad: Cuando un animal posee datos repetidos (por ejemplo, varios tamaños de camada o varias lactaciones) es frecuente incluir un factor aleatorio en el modelo que dé cuenta del ambiente permanente que afecta a las distintas medidas del animal, puesto que rara vez la repetibilidad coincide con la heredabilidad (sólo cuando coinciden, no hay ambiente permanente). Los efectos permanentes se consideran incorrelacionados y con la misma varianza, por lo que su matriz de varianzas covarianzas es la identidad multiplicada por la varianza de efectos permanentes.

Modelo de efectos maternos: En otras ocasiones se añade un efecto aleatorio que recoge el ambiente materno; por ejemplo, en el carácter peso al destete, el provenir de la misma madre, que puede ser mejor o peor lechera, es un efecto ambiental aleatorio que afecta a todos los individuos de la misma camada. Añadir estos efectos no es difícil, y no causa dificultades mayores salvo en modelos muy complejos en los que se desea también utilizar la parte genética de los efectos maternos separada de la parte ambiental (por ejemplo; si una madre es buena lechera parte puede deberse a causas genéticas que se traducen en un peso al destete de los hijos). En esos casos hay que estimar no sólo los parámetros genéticos del efecto materno, sino sus correlaciones con el efecto directo. Podría ocurrir, por ejemplo, que parte de los genes que hacen que una madre sea buena lechera sean los mismos que producen que ella haya tenido un buen peso al destete, con lo que habría una correlación genética entre efectos maternos y directos (propios del individuo).

Estimación cuando los valores aditivos no tienen media cero

En ocasiones los valores aditivos no tienen de media cero, por ejemplo, cuando se importan animales de un valor genético superior es frecuente hacer grupos genéticos correspondientes a los orígenes de importación, de forma que el valor aditivo de un individuo es

valor aditivo = valor medio del grupo + valor aditivo dentro del grupo

El valor medio del grupo se estima como un efecto fijo, por lo que el valor aditivo del individuo k dentro del grupo i pasa a ser

$$\text{valor aditivo}_{ik} = g_i + u_{ik}$$

Los efectos de grupo g_i pasan a estar incluidos en el vector \mathbf{b} del modelo mixto y el objetivo pasa a ser, de forma general, estimar una combinación de efectos fijos y aleatorios, normalmente estimar un parámetro del tipo

$$w = \mathbf{k}'\mathbf{b} + u$$

La solución se deriva de forma análoga y es, como probablemente cabría esperar,

$$w = \mathbf{k}'\hat{\mathbf{b}} + \mathbf{c}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

BLUP Multicarácter

Al igual que se utilizan índices par varios caracteres, es frecuente utilizar el BLUP multicarácter (al fin y al cabo el BLUP no es sino un índice con los efectos ambientales corregidos). Algunas dificultades se presentan cuando los modelos utilizados en los distintos caracteres no son los mismos (por ejemplo, si un carácter es el tamaño de camada y otro el índice de conversión, el primero tiene un efecto aleatorio permanente y el segundo no). Cae fuera de los objetivos de este libro el tratar estos modelos en detalle, para los que nos referimos al libro de Rico (1999) que contiene una completa casuística del BLUP explicada con numerosos ejemplos.

3.4.4. OTRAS INTERPRETACIONES DEL BLUP

El BLUP como el mejor de los estimadores lineales insesgados

Supongamos que decidimos estimar los valores aditivos con una función lineal de todos los datos disponibles, presentes y recogidos en el pasado,

$$\hat{u} = a_1 y_1 + a_2 y_2 + \dots + a_n y_n = \mathbf{a}'\mathbf{y}$$

Si tomamos una función de pérdidas cuadrática, el Riesgo es (Apéndice I, ecuación 3.21)

$$R(u, \hat{u}) = E(\hat{u} - u)^2 = [E(u) - E(\hat{u})]^2 + \text{var}(u) + \text{var}(\hat{u}) - 2 \text{cov}(u, \hat{u}) \quad (3.18)$$

Por comodidad algebraica suele tomarse, como vimos en el modelo mixto de la ecuación 3.15, $E(u) = 0$. Los otros componentes del riesgo son

$$E(\hat{u}) = E(\mathbf{a}'\mathbf{y}) = \mathbf{a}' E(\mathbf{y}) = \mathbf{a}'\mathbf{X}\mathbf{b}$$

$$\text{var}(\hat{u}) = \text{var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}' \text{var}(\mathbf{y}) \mathbf{a} = \mathbf{a}' \mathbf{V} \mathbf{a}$$

$$\text{cov}(u, \hat{u}) = \text{cov}(u, \mathbf{a}'\mathbf{y}) = \mathbf{a}' \text{cov}(u, \mathbf{y}) = \mathbf{a}' \mathbf{c} = \mathbf{c}'\mathbf{a}$$

con lo que el riesgo es

$$R(u, \hat{u}) = [\mathbf{a}' E(\mathbf{y})]^2 + \text{var}(u) + \mathbf{a}' \mathbf{V} \mathbf{a} - 2 \mathbf{c}'\mathbf{a}$$

Si derivamos respecto a \mathbf{a} para obtener el valor que minimiza el riesgo,

$$\partial R / \partial \mathbf{a} = 2 \mathbf{a}' E(\mathbf{y}) E(\mathbf{y}') + 2 \mathbf{a}' \mathbf{V} - 2 \mathbf{c}' = 0 \quad (3.19)$$

nos encontramos con que necesitamos conocer la media de los datos $E(\mathbf{y})$ para poder ofrecer un estimador de \mathbf{a} . Si los datos estuvieran centrados, entonces $E(\mathbf{y})=0$, y el valor de \mathbf{a} es el de los índices de selección. Una forma de evitar el problema es usar un subconjunto de estimadores, aquéllos que cumplen la condición $E(\hat{u}) = E(u)$, con lo que la fórmula del riesgo (3.18) pasa a ser

$$R(u, \hat{u}) = \text{var}(u) + \mathbf{a}' \mathbf{V} \mathbf{a} - 2 \mathbf{a}' \mathbf{c}$$

Debemos, pues, minimizar el riesgo, *sujeto a la condición* de que $E(u) = \mathbf{a}' \mathbf{X} \mathbf{b} = E(\hat{u})$; esto es, de entre todos los estimadores posibles sólo examinaremos los insesgados. En nuestro modelo, $E(u)=0$, por tanto la condición de insesgamiento implica que $\mathbf{a}' \mathbf{X} \mathbf{b}=0$, y como \mathbf{b} es un vector de constantes esto a su vez implica que $\mathbf{a}' \mathbf{X} = \mathbf{0}$, con lo que el riesgo se puede escribir

$$R(u, \hat{u}) = \text{var}(u) + \mathbf{a}' \mathbf{V} \mathbf{a} - 2 \mathbf{a}' \mathbf{c} + 2 \mathbf{a}' \mathbf{X} \mathbf{q}$$

donde al vector de parámetros \mathbf{q} se le conoce como vector de multiplicadores de Lagrange. Obsérvese que como $\mathbf{a}' \mathbf{X} = \mathbf{0}$, el término que se ha añadido, $2\mathbf{a}' \mathbf{X} \mathbf{q}$, es nulo, por lo que el Riesgo no se ha modificado, es un mero artificio para imponer la condición de insesgamiento. Para minimizar el riesgo derivaremos respecto a los parámetros e igualaremos a cero.

$$\partial R(u, \hat{u}) / \partial \mathbf{a}' = 2 \mathbf{V} \mathbf{a} - 2 \mathbf{c} + 2 \mathbf{X} \mathbf{q} = \mathbf{0}$$

$$\partial R(u, \hat{u}) / \partial \mathbf{q} = 2 \mathbf{a}' \mathbf{X} = 0$$

En este sistema de ecuaciones se minimiza el riesgo a la par que se mantiene la condición de insesgamiento (segunda ecuación). Despejando \mathbf{a} de la primera ecuación y sustituyendo en la segunda,

$$\mathbf{a} = \mathbf{V}^{-1} (\mathbf{X} \mathbf{q} - \mathbf{c})$$

$$\mathbf{a}'\mathbf{X} = \mathbf{X}'\mathbf{a} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{q} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{c} = 0$$

$$\mathbf{q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{c}$$

$$\mathbf{a} = \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{c} - \mathbf{V}^{-1}\mathbf{c}$$

$$\hat{u} = \mathbf{a}'\mathbf{y} = \mathbf{c}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{c}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{c}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

El BLUP es, pues, el estimador *dentro del subconjunto de estimadores lineales insesgados* que minimiza el Riesgo (concretamente el riesgo cuadrático medio).

El BLUP como estimador Bayesiano

La forma esencial de trabajar de la escuela bayesiana consiste en, dados los datos observados en el experimento, describir toda la incertidumbre que puede existir en torno a lo que se quiere estimar; esto es, representar la probabilidad de cada valor posible. En un contexto bayesiano no hay “efectos fijos”, todos los efectos son aleatorios, puesto que se representa la probabilidad de que los efectos tomen tal o cual valor, lo que implica que deben tratarse como variables aleatorias (Apéndice II). Trataremos, pues de encontrar un estimador para el vector $\mathbf{t}' = [\mathbf{b}' \ \mathbf{u}']$ a partir de los datos \mathbf{y} . Para ello determinaremos primero $f(\mathbf{t} | \mathbf{y})$, que es la función de densidad de probabilidad de \mathbf{t} *dados los datos* y luego hallaremos la moda (el valor más probable) de esta función. Aplicando el teorema de Bayes,

$$f(\mathbf{t} | \mathbf{y}) = f(\mathbf{y} | \mathbf{t}) f(\mathbf{t}) / f(\mathbf{y})$$

Como $f(\mathbf{y})$ no depende de \mathbf{t} , podemos decir que es constante y representar $f(\mathbf{t} | \mathbf{y})$ como proporcional a los otros dos términos; esto es,

$$f(\mathbf{t} | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{t}) f(\mathbf{t})$$

Si llamamos $\mathbf{W} = [\mathbf{X} \ \mathbf{Z}]$, como ya vimos al derivar las ecuaciones del modelo mixto, $f(\mathbf{y} | \mathbf{t}) \sim N([\mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u}], I\sigma_e^2) = N(\mathbf{W}\mathbf{t}, I\sigma_e^2)$; tenemos, pues, por un lado que

$$f(\mathbf{y} | \mathbf{t}) \propto \exp[(\mathbf{y} - \mathbf{W}\mathbf{t})' (\mathbf{y} - \mathbf{W}\mathbf{t})]$$

Por otra parte $f(\mathbf{t})$, que es la distribución *a priori* de \mathbf{t} , podemos suponer que es también Normal. Esto es razonable para los valores aditivos \mathbf{u} , pero para los efectos ambientales \mathbf{b} es discutible, y luego lo discutiremos. La media y la matriz de varianzas covarianzas *a priori* de \mathbf{t} son

$$\mathbf{m}^* = E(\mathbf{t}') = E[\mathbf{b}' \mathbf{u}'] = [\mathbf{m}_b' \mathbf{0}]$$

$$\mathbf{V}^* = \text{Var}(\mathbf{t}') = \text{Var}[\mathbf{b}' \mathbf{u}'] = \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{bmatrix}$$

donde \mathbf{m}_b y \mathbf{S} son la media y la matriz de varianzas covarianzas *a priori* de los efectos ambientales.

$$f(\mathbf{t}) \propto \exp[(\mathbf{t} - \mathbf{m}^*)' \mathbf{V}^{*-1} (\mathbf{t} - \mathbf{m}^*)]$$

$$f(\mathbf{t} | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{t}) f(\mathbf{t}) \propto \exp [(\mathbf{y} - \mathbf{Wt})' (\mathbf{y} - \mathbf{Wt}) + (\mathbf{t} - \mathbf{m}^*)' \mathbf{V}^{*-1} (\mathbf{t} - \mathbf{m}^*)]$$

Hallaremos la moda, que es el valor que maximiza $f(\mathbf{t} | \mathbf{y})$.

$$\partial f(\mathbf{t} | \mathbf{y}) / \partial \mathbf{t} = 2 \mathbf{W}'(\mathbf{y} - \mathbf{Wt}) + 2 \mathbf{V}^{*-1}(\mathbf{t} - \mathbf{m}^*) = 0$$

$$[\mathbf{W}'\mathbf{W} + \mathbf{V}^{*-1}] \mathbf{t} = \mathbf{W}'\mathbf{y} + \mathbf{V}^{*-1} \mathbf{m}^* \quad (3.20)$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} + \mathbf{S}^{-1} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} + \mathbf{S}^{-1}\mathbf{m}_b \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Estas ecuaciones son muy parecidas a las del modelo mixto, pero al producto matricial de los efectos antes considerados como "fijos", $\mathbf{X}'\mathbf{X}$, se le añade la inversa de la matriz de varianzas covarianzas de estos efectos, justamente como ocurre en los efectos aditivos con la matriz \mathbf{G} . En el caso de una ignorancia total sobre los valores *a priori* de los efectos fijos, estos podrían variar en el intervalo $]-\infty, +\infty[$, con lo que su varianza tendería a infinito y \mathbf{S}^{-1} tendería a cero. Si \mathbf{S}^{-1} se anula tenemos entonces las ecuaciones del modelo mixto.

El BLUP no es, pues, sino un estimador bayesiano que considera que la distribución *a priori* de los efectos aditivos es normal de media cero y varianza \mathbf{G} , y la de los efectos ambientales es plana (todos los posibles valores tienen la misma probabilidad *a priori*) y varía a lo largo de toda la recta real (pueden tomar cualquier valor). Este último punto es ilógico y demuestra que el BLUP no es el mejor estimador posible desde un punto de vista Bayesiano. Presumiblemente se puede construir un estado de creencias *a priori* sobre los efectos fijos y tratarlos como a los aleatorios, pero esto no es siempre fácil de hacer, como ya discutimos en el apartado 3.4.2. *¿Fijos o aleatorios?*

El BLUP y las estimas por mínimos cuadrados

Al principio del apartado 3.4.2 comentamos que el BLUP no puede estimarse de forma máximo-verosímil porque al condicionar sobre los efectos aleatorios estos quedaban “finados” y su estima era la misma que en el caso de mínimos cuadrados. Para ver las diferencias entre el BLUP y las estimas de mínimos cuadrados, vamos a combinar dos estimadores. El primero es el estimador de mínimos cuadrados de \mathbf{t} ; esto es, considerando \mathbf{b} y \mathbf{u} fijos, con lo que $\mathbf{y}|\mathbf{t} \sim N([\mathbf{Xb}+\mathbf{Zu}], I\sigma_e^2)$

$$\mathbf{t}_1 = (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}' \mathbf{y}$$

La información a priori, como en el caso anterior, está contenida en la media y la varianza a priori de los efectos genéticos y ambientales. El segundo estimador es simplemente la media a priori de los efectos:

$$\mathbf{t}_2 = \mathbf{m}^*$$

Las varianzas de los estimadores son

$$\text{var}(\mathbf{t}_1) = (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}' \text{var}(\mathbf{y}) \mathbf{W} (\mathbf{W}'\mathbf{W})^{-1} = (\mathbf{W}'\mathbf{W})^{-1} (1/\sigma_e^2)$$

$$\text{var}(\mathbf{t}_2) = \mathbf{V}^*$$

Cuando se dispone de dos estimadores incorrelacionados \mathbf{t}_1 y \mathbf{t}_2 , y se desea combinarlos para optimizar la estimación, el nuevo estimador \mathbf{t} sopesa ambos estimadores con unos pesos que son proporcionales a las inversas de sus varianzas (Apéndice III). Combinando ambos estimadores

$$\mathbf{t} = [\mathbf{W}'\mathbf{W} + \mathbf{V}^{*1}]^{-1} [\mathbf{W}'\mathbf{W} (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}' \mathbf{y} + \mathbf{V}^{*1} \mathbf{m}^*] = [\mathbf{W}'\mathbf{W} + \mathbf{V}^{*1}]^{-1} [\mathbf{W}'\mathbf{y} + \mathbf{V}^{*1} \mathbf{m}^*]$$

este estimador es idéntico al estimador Bayesiano de la ecuación 3.20, y por tanto a las ecuaciones del modelo mixto. El BLUP no es, pues, sino un estimador de mínimos cuadrados tradicional, combinado con otro estimador que recoge la información a priori.

3.5. Cálculo de los valores de mejora. Software disponible

Se requiere software para la selección por BLUP y, según su complejidad, para los índices de selección. En la actualidad, en especies de producción intensiva, la selección más frecuente se realiza mediante un BLUP para varios caracteres, en los que los modelos

aplicados a cada carácter no siempre coinciden. Un programa muy completo y de uso muy frecuente, que puede obtener BLUP multivariantes con modelos distintos para cada carácter, es el programa PEST de Groeneveld, Kovak y Wang (1990), que se puede adquirir por un módico precio para usos académicos, y con un coste algo mayor para usos comerciales.

Otro programa, en este caso público, más limitado, es el de Misztal ([PONER REFERENCIA WEB](#)). Un excelente programa público para usos didácticos es el GENUP([PONER REFERENCIA WEB](#)), que contiene un amplio abanico de programas relativos a la mejora genética, incluyendo índices de selección multicarácter, en los que el alumno va rellenando paso a paso y de forma interactiva los elementos requeridos para la solución de los problemas que se plantean.

Los programas que calculan componentes de varianza proporcionan también los valores de mejora, pero estos programas quedan fuera de los alcances de este capítulo.

Anexo 1. Riesgo de un estimador

La forma habitual de considerar la incertidumbre en torno a la estimación es primero definir qué entendemos como error de estimación. Llamaremos error de estimación a la diferencia $e = (\theta - t)$ entre el valor del parámetro y su estimación (aunque podríamos haber definido el error de otra forma, por ejemplo, como una proporción del verdadero valor, $e = \theta \cdot t$). Un error en la estimación tiene consecuencias económicas, que deben ser medidas por una función de pérdidas, función que depende tanto del verdadero valor como del estimador. En ocasiones un error en la estimación ocasiona una catástrofe, por lo que esa función de pérdidas puede ser discontinua y tener elevados valores asociados a ciertos valores del parámetro o a ciertas diferencias entre el parámetro y su estimación, pero en general se toman valores de funciones de pérdidas sencillos, como $|e| = |\theta - t|$ ó $e^2 = (\theta - t)^2$, siendo esta última función la más utilizada (nótese que las pérdidas mayores tienen más importancia al estar elevadas al cuadrado). Nuestro objetivo es encontrar un estimador que minimice las pérdidas, pero esto no es posible hacerlo para todos los valores del parámetro (por ejemplo, el estimador $t=7$ es el mejor valor cuando $\theta=7$, pero no en otros casos), por lo que hay que tomar una alternativa, como minimizar la máxima pérdida o minimizar la media de las pérdidas. Lo más frecuente es minimizar la media de las pérdidas, función a la que se llama *Riesgo* de un estimador. Tomando como función de pérdidas el cuadrado del error de estimación, y llamando $E_e = E(e)$, $E_\theta = E(\theta)$ y $E_t = E(t)$, el riesgo es

$$R(\theta, t) = E(e)^2 = E(e - E_e + E_e)^2 = E(e - E_e)^2 + (E_e)^2 + 2 E(e - E_e) E_e = \text{var}(e) + (E_e)^2$$

ya que $E(e - E_e) E_e = E(e)E_e - (E_e)^2 = 0$. Por tanto,

$$R(\theta, t) = \text{var}(\theta - t) + (E_\theta - E_t)^2$$

Cuando el efecto θ a estimar es un *efecto fijo*, $E_\theta = \theta$, y $\text{var}(\theta) = 0$, con lo que la media del error es $\theta - E_t$ y la varianza del error $\text{var}(\theta - t) = \text{var}(t)$, con lo que la fórmula del riesgo pasa a ser

$$R(\theta, t) = (\theta - E_t)^2 + \text{var}(t)$$

Cuando el efecto θ a estimar es un *efecto aleatorio*, el Riesgo es

$$R(\theta, t) = (E_\theta - E_t)^2 + \text{var}(\theta) + \text{var}(t) - 2 \text{cov}(\theta, t) \quad (3.22)$$

Si t es un estimador de regresión lineal; esto es, del tipo $t = b \cdot C$, resulta que

$$\text{var}(t) = \text{var}(b C) = b^2 \text{var}(C) = \frac{\text{cov}^2(C, \theta)}{\sigma_C^4} \cdot \sigma_C^2 = \frac{\text{cov}^2(C, \theta)}{\sigma_C^2}$$

$$\text{cov}(t, \theta) = \text{cov}(b C, \theta) = b \text{cov}(C, \theta) = \frac{\text{cov}(C, \theta)}{\sigma_C^2} \cdot \text{cov}(C, \theta) = \frac{\text{cov}^2(C, \theta)}{\sigma_C^2} = \text{var}(t)$$

con lo que la fórmula del riesgo pasa a ser

$$R(\theta, t) = (E_\theta - E_t)^2 + \text{var}(\theta) - \text{var}(t) \quad (3.23)$$

La cantidad $\text{var}(\theta)$ es una característica de la población y no depende del tamaño de muestra ni del estimador que se use. Al aumentar el tamaño muestral, se reduce la varianza del error $\text{var}(\theta - t)$; en el caso de los efectos fijos $\text{var}(t)$ disminuye, y en el caso de los efectos aleatorios $\text{var}(t)$ aumenta aproximándose a $\text{var}(\theta)$. En ambos casos la varianza del error disminuye, pero el comportamiento diferente de la varianza del estimador $\text{var}(t)$ según el efecto sea fijo o aleatorio es un típica fuente de confusiones.

A la media del error de estimación $E_e = E_\theta - E_t$ se le llama *sesgo* del estimador, y en el caso de los efectos fijos es considerado como una propiedad atractiva el que un estimador sea *insesgado*; esto es, que tenga sesgo nulo, $E(t) = \theta$, y por tanto se distribuya alrededor del valor verdadero en cada repetición conceptual de la experiencia. Esta propiedad es mucho menos atractiva en el caso de los efectos aleatorios, puesto que en cada repetición del experimento no sólo cambia t sino también θ , con lo que el estadístico no se distribuye alrededor del valor verdadero. Algunos estadísticos frecuentistas como Fisher, consideran la propiedad de insesgamiento como más bien irrelevante, puesto que transformaciones de un estadístico insesgado pierden la propiedad de insesgamiento; por ejemplo, la raíz cuadrada de un estimador insesgado de la varianza no es un estimador insesgado de la desviación típica, por lo que es inútil utilizar estimadores insesgados de la varianza si el interés está en obtener desviaciones típicas.

En general, el estadístico que minimiza el riesgo depende de θ , no hay un único estadístico que lo minimice. Por eso se le busca dentro de algún subconjunto que resuelva esta indeterminación, por ejemplo el estadístico insesgado de varianza mínima. En ese caso el riesgo del estimador coincide con la varianza del error de estimación.

Anexo 2. La inferencia bayesiana

La escuela bayesiana fue fundada por Laplace por medio de varios trabajos publicados de 1774 a 1812, y durante el siglo XIX ocupó un papel preponderante en la inferencia científica (Stigler, 1986). Antes que Laplace, y sin que al parecer éste tuviera conocimiento, se había presentado en la Royal Society de Londres un trabajo póstumo atribuido a un oscuro clérigo, el reverendo Thomas Bayes (quien no publicó trabajos matemáticos en vida), formalizando el mismo principio de inferencia. Al parecer este principio había sido formulado anteriormente, y Stigler (1983) lo atribuye a Sauderson, un profesor de óptica ciego, autor de numerosos trabajos en diversos campos de la matemática. Los trabajos sobre verosimilitud de Fisher en los años 20 y los de la escuela frecuentista en los 30 y 40 hicieron casi desaparecer a la escuela bayesiana, hasta que comenzó un "revival" a mediados de los 50 que dura *in crescendo* hasta nuestros días. En mejora genética animal el bayesianismo fue introducido por Daniel Gianola, primero en trabajos sobre caracteres umbral en colaboración con J.L. Foulley, y posteriormente en artículos en los que se desarrollan aplicaciones a prácticamente todos los campos de la mejora genética animal (ver Blasco, 2001, para una revisión sobre los métodos bayesianos en mejora genética).

La forma esencial de trabajar de la escuela bayesiana consiste en, dados los datos observados en el experimento, describir toda la incertidumbre que puede existir en torno a un parámetro, usando como medida natural de la incertidumbre la probabilidad de que el parámetro tome determinados valores. Por ejemplo, en el caso de la heredabilidad se obtendría la función de densidad de probabilidad $f(h^2|\mathbf{y})$ siendo \mathbf{y} el vector de valores observados. Una vez obtenida esa distribución se pueden hacer inferencias de múltiples maneras: por ejemplo, se puede desear averiguar entre qué valores se encuentra h^2 con una probabilidad del 95%, o qué probabilidad tiene el que h^2 esté entre tal y tal valor. En los casos en los que es necesaria una estimación puntual de h^2 , por ejemplo para un índice de selección, hay varios parámetros de la función de densidad $f(h^2|\mathbf{y})$ que pueden ser usados como estimación puntual, y cuyo uso depende de la preferencia del investigador. Por ejemplo, la *moda*, que es el valor más probable de h^2 dada la muestra \mathbf{y} ; la *mediana*, cuyo valor hace tan probable que el valor verdadero sea superior como inferior a esta estima y minimiza el riesgo de estimación cuando la función de pérdidas es $|h^2 - \hat{h}^2|$; o la *media*, que es el estimador que minimiza el riesgo mínimo cuadrático $E(h^2 - \hat{h}^2)^2$.

Para poder hacer todas estas inferencias es menester disponer de la función de densidad de probabilidad $f(h^2|\mathbf{y})$. De acuerdo con las leyes de la probabilidad, la probabilidad $P(A,B)$ de que se presenten dos sucesos simultáneamente es

$$P(A,B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

con lo que

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

En nuestro caso,

$$f(h^2|\mathbf{y}) = f(\mathbf{y}|h^2) f(h^2) / f(\mathbf{y}) = cte \cdot f(\mathbf{y}|h^2) f(h^2)$$

donde f significa “función de densidad”, pero no es necesariamente la misma para $\mathbf{y}|h^2$ que para h^2 . Obsérvese que $f(h^2|\mathbf{y})$ es una función de h^2 , pero no de \mathbf{y} , que está fijada; por tanto $f(\mathbf{y}|h^2)$ es aquí función de h^2 , pero no de \mathbf{y} , que es exactamente la definición de verosimilitud. Por la misma razón $f(\mathbf{y})$ es una constante, ya que no depende de h^2 e \mathbf{y} está fijado. Finalmente, $f(h^2)$ es la densidad de probabilidad de h^2 al margen de nuestro experimento.

Las críticas al bayesianismo tienen que ver con esta última probabilidad llamada *a priori* porque no depende de los datos, es previa al experimento. En ocasiones esta información está claramente determinada; por ejemplo, la probabilidad *a priori* de obtener un individuo recesivo en el cruce de dos heterocigotos es $1/4$, al margen del experimento, pero en el caso de la heredabilidad no está claro qué se quiere decir con esta probabilidad previa. En muchas ocasiones es difícil cuantificar la información *a priori* de una forma tan objetiva como la de los ejemplos que acabamos de citar. En esos casos los estadísticos no bayesianos consideran que no es posible aplicar el teorema de Bayes y el problema no tiene solución por la vía de las probabilidades. Dentro del campo bayesiano se ha intentado dar respuesta a esta dificultad de varias formas, bien definiendo la probabilidad como un estado de creencias del investigador, quien define $f(h^2)$ según su opinión y los experimentos realizados previamente o consultados en la literatura, o bien eliminando en la práctica la influencia de la probabilidad *a priori* a base de aumentar el tamaño muestral. Si se dispusiera de suficientes datos, la probabilidad *a priori* no influiría en la distribución de la densidad posterior de probabilidades, por tanto se deben hacer experimentos con un número de datos suficiente como para que la función *a priori* carezca de relevancia. En ese caso la función de densidad de probabilidades *a priori* se busca de forma relativamente arbitraria (se procura que coincida en lo posible con una opinión defendible; p. ej., que no sea muy probable que la heredabilidad tenga un valor de 0.95), y habitualmente se procura que facilite los cálculos de la función posterior y que no conduzca a paradojas o a resultados inadmisibles. Es frecuente en ese caso probar varias funciones *a priori* diferentes y alguna función de referencia (p. ej., un *a priori* plano en el que todos los valores presentan la misma probabilidad) para comprobar que el resultado final (la función posterior) apenas se altera. Cuando no hay información *a priori*, o cuando se desea actuar como si no la hubiera, el bayesianismo se enfrenta a la dificultad de que es imposible realizar inferencias, puesto que la probabilidad *a priori* es necesaria para poder aplicar el teorema de Bayes, y cualquier forma que tenga esa probabilidad es de alguna manera informativa. Se ha sugerido suponer que cuando no hay información sobre los distintos sucesos posibles, hay que asignarles a todos la misma probabilidad *a priori*. En el caso de variables continuas esto implica representarlas como una recta paralela al eje de las X en un intervalo concreto, por ejemplo al intervalo $[0,1]$ para el caso de la heredabilidad, por lo que se les conoce también como *a prioris planos* o *no informativos*, siendo este último nombre inapropiado, puesto que sí que son informativos (no es lo mismo decir que se ignora la probabilidad de los distintos sucesos que decir que todos tienen la misma probabilidad). Estos *a prioris planos* son frecuentes en la literatura como funciones de referencia. Otras soluciones más complejas aunque escasamente aplicadas en el campo de la mejora genética son discutidas por Blasco (2001).

Apéndice III

Cuando se dispone de dos estimadores incorrelacionados t_1 y t_2 , y se desea combinarlos para optimizar la estimación, el nuevo estimador t sopesa ambos estimadores con unos pesos w_1 y w_2

$$t = w_1 t_1 + w_2 t_2$$

$$\text{var}(t) = w_1^2 \text{var}(t_1) + w_2^2 \text{var}(t_2)$$

Vamos a calcular w_1 y w_2 de forma que la varianza de t sea mínima y que ambos sumen la unidad; esto es, consiste el problema en saber qué peso se le da a un estimador respecto al otro.

$$w_1 + w_2 = 1$$

$$\text{var}(t) = w_1^2 \text{var}(t_1) + (1 - w_1)^2 \text{var}(t_2)$$

$$d\text{var}(t)/dw_1 = 2 \text{var}(t_1) w_1 - 2 \text{var}(t_2) + 2 w_1 \text{var}(t_2) = 0$$

$$w_1 = \frac{\text{var}(t_2)}{\text{var}(t_1) + \text{var}(t_2)} = \frac{\frac{1}{\text{var}(t_1)}}{\frac{1}{\text{var}(t_1)} + \frac{1}{\text{var}(t_2)}}$$

y análogamente con w_2

$$w_2 = \frac{\text{var}(t_1)}{\text{var}(t_1) + \text{var}(t_2)} = \frac{\frac{1}{\text{var}(t_2)}}{\frac{1}{\text{var}(t_1)} + \frac{1}{\text{var}(t_2)}}$$

Cuestiones

Libros recomendados***Referencias***