ITEA (1998), Vol. **94A N.º 1**, 5-41

LA CONTROVERSIA BAYESIANA EN MEJORA ANIMAL

A. Blasco
Departamento de Ciencia Animal
Universidad Politécnica de Valencia
Apartado 22012
46071 Valencia

RESUMEN

Se examinan los presupuestos de las escuelas frecuentista y bayesiana para la inferencia científica. Se discute la aproximación bayesiana cuando hay información a priori, cuando se admite que la hay bajo ciertas hipótesis, cuando la información a priori es vaga y cuando no hay información previa al experimento. Los métodos más usuales en mejora genética animal (Indices de selección, BLUP, Máxima verosimilitud, REML) son presentados bajo el punto de vista de las dos escuelas y se examinan sus propiedades en uno y otro caso. Se expone también la predicción bayesiana del valor genético y de los parámetros genéticos. Finalmente se discuten las aproximaciones frecuentista y bayesiana desde un punto de vista ontológico y práctico, y se exponen casos en los que la inferencia bayesiana da soluciones que aún no ofrece la escuela frecuentista.

Palabras claves: Análisis frecuentista, Análisis Bayesiano, Mejora genética.

SUMMARY
THE BAYESIAN CONTROVERSY IN ANIMAL BREEDING

The frequentist and bayesian approaches to scientific inference are discussed. The Bayesian approach is discussed in the cases in which prior information is available, prior information is available under certain hypotheses, prior information is vague and there is no prior information. The most usual methods in animal breeding (Selection index, BLUP, ML, REML) are presented under the hypotheses of both schools of inference, and their properties are examined in both cases. Bayesian prediction of genetic values and genetic parameters are presented. Finally, the frequentist and bayesian approaches are compared from an ontological and practical point of view. Some problems for which the bayesian methods offer solutions that are not found by using frequentist methods are exposed.

Key words: Frequentist analysis, Bayesian analysis, Animal breeding.

"Un tema constante en el desarrollo de la estadística ha sido la búsqueda de una justificación a lo que los estadísticos hacen. Leyendo los libros de texto uno puede hacerse la idea equivocada de que 'Student' propuso su test porque era el test insesgado uniformemente más potente para la media de una distribución Normal, pero sería más preciso decir que el concepto de test insesgado uniformemente más potente ha gana-

do mucho de su predicamento porque produce el test t, y como todo el mundo sabe el test t es algo que está muy bien".

DAWID (comentario en una discusión del artículo de O. Bandorff-Nielsen "Plausibility inference", J. Royal Stat. Soc. Serie B 38: 123-125, 1976).

Introducción

La reciente irrupción de métodos bayesianos en la mejora genética animal ha sumido en una cierta perplejidad a los mejoradores. Los partidarios del uso de estas técnicas aseguran que dan respuesta a problemas mal resueltos en mejora animal, por ejemplo, citando a GIANOLA *et al.* (1994)[1]:

"Supongamos que nos encontramos en el caso donde se aplican análisis de máxima verosimilitud (ML) y BLUP sobre un modelo animal. Con respecto a la primera cuestión [¿Qué puede decirse acerca de la varianza genética en la población base?] obtendríamos una estimación puntual de la varianza genética y una simple medida de incertidumbre que, técnicamente hablando, sólo tendría sentido en muestras grandes y si los datos se distribuyen de forma normal. Respecto a la segunda cuestión [¿Cómo estimar la incertidumbre de los valores de mejora cuando no se conocen las medias y varianzas?] tendríamos que utilizar las estimas máximo verosímiles de las varianzas como si fueran los parámetros verdaderos, ignorando su error. La cuestión relativa a la respuesta a la selección [¿Cómo estimar la eficacia de la selección, añadiendo a esa estima una medida de su incertidumbre?] incluso es más complicada. Se pueden derivar estimas de la respuesta a partir de un modelo animal, pero sus propiedades son desconocidas... Una alternativa es adoptar la posición bayesiana. Para cada una de las cuestiones expuestas la respuesta Bayesiana reside en calcular la densidad marginal posterior del parámetro de interés. Esta distribución cuantifica exactamente la incertidumbre acerca del parámetro desconocido".

Pero también conocidos estadísticos de la mejora animal muestran su escepticismo, cuando no su abierta oposición al uso de estos métodos:

"He indicado por qué algunos estimadores bayesianos tienen propiedades poco afortunadas, pues suelen conducir a estimas de cero para las componentes de varianza entre grupos, en experimentos bien diseñados. Sin embargo aún se advoca por este tipo de estimador" (THOMPSON, 1987).

"La controversia entre las escuelas de inferencia frecuentista y bayesiana tiene unas implicaciones enormes. Desde el punto de vista del análisis de datos la consecuencia es que todos los procedimientos expuestos en libros como el de Snedecor y Cochran 'Métodos estadísticos' no es sólo que son malos, sino que además están mal fundamentados. Por otra parte, cada informe sobre un trabajo de investigación debería contener afirmaciones del investigador del tipo '*mi probabilidad* de que un parámetro θ caiga entre 2,3 y 5,7 es 0,90'" (KEMPTHORNE, 1984).

Aunque hasta el momento me he referido a dos escuelas de inferencia (frecuentista y bayesiana), puesto que son las únicas

1. Las citas textuales no figuran en el idioma original. La traducción no es literal.

que tienen relevancia hoy en día (al menos en el campo de la mejora genética animal), la inferencia que proviene del examen de la verosimilitud está situada de forma un tanto intermedia entre ambas. Por un lado la verosimilitud juega un papel central en la inferencia bayesiana como portadora de la información que proviene de los datos, y por otro el método de máxima verosimilitud goza de propiedades atractivas desde el punto de vista frecuentista[2]. Normalmente el mejorador está interesado en soluciones eficaces para estimar el valor de mejora de sus animales, y no necesariamente interesado en la filosofía subyacente, por lo que resulta tranquilizador saber que cuando el número de datos es grande los resultados son los mismos en los tres casos, o que los métodos de mejora animal corrientemente utilizados pueden derivarse de una u otra forma según los presupuestos de una u otra escuela. ROBINSON (1991) llega a sugerir que en el caso de la estima de valores de mejora la distancia entre escuelas es mínima, y que la apreciación de que no es así se debe esencialmente a la insistencia de los mejoradores adscritos a una u otra escuela en subrayar las diferencias entre ellas más que las similitudes. Creo que estos planteamientos andan lejos de resolver el problema. La estadística tiene como objetivo proponer cuánto conocemos sobre algo y con qué grado de certeza o de duda, por lo que el caso de muestras grandes y representativas es generalmente el caso en el que menos se nota el auxilio de esta rama de las matemáticas[3]. Además, en la mayor parte de experimentos se intenta maximizar la eficacia respecto a los medios empleados, por lo que no se pretende utilizar grandes muestras sino muestras de un tamaño suficiente como para poder extraer conclusiones relevantes. Aunque es reconfortante saber que el BLUP y el REML tienen una interpretación adecuada en ambas escuelas, las diferencias entre ellas son demasiado notorias, desde el punto de vista práctico, como para ignorarlas: por ejemplo:

1. La forma de presentar los resultados es completamente diferente. La escuela bayesiana trabaja con probabilidades, y ante un problema como el de estimar la heredabilidad su resultado es una distribución de densidad de probabilidad de este parámetro, dada una muestra determinada. La escuela frecuentista aporta ante el mismo problema un valor estimado y un intervalo de confianza. Por supuesto, la escuela frecuentista puede en ocasiones ofrecer las distribuciones de la muestra en el caso de que se repitiera infinitas veces el experimento, o una aproximación por ejemplo, utilizando técnicas de bootstrap, que últimamente han alcanzado cierta relevancia en algunos problemas de genética molecular, pero la inter-

2. No he incluido la inferencia fiducial de Fisher o la estructural de Fraser porque no han tenido incidencia en los métodos utilizados en mejora animal y por otra parte están casi abandonadas. El lector interesado puede consultar el primero de los dos artículos sobre inferencia de la Enciclopedia of Statistical Sciences (KOTZ y JOHNSON, 1982), Vol 3, pp. 89-104. Una visión frecuentista de la inferencia se encuentra en STUART y ORD (1991), mientras que para la visión bayesiana se puede consultar BERNARDO y SMITH (1994). La visión de verosimilitud se puede encontrar en EDWARDS (1992). El excelente libro de BARNETT (1982) no recoge los últimos avances en la materia, por lo que la comparación de sistemas de inferencia queda un poco anticuada desde el punto de vista práctico.

3. Esto no siempre es cierto, hay ocasiones en que con una muestra grande el resultado de un test frecuentista es opuesto al de uno bayesiano (paradoja de Lindley, ver p. ej. PRESS 1989, pp. 36-37), pero aún siendo un asunto importante, el tema excede los límites de este artículo.

pretación es completamente diferente, pues la distribución que se obtiene es la del estimador, no la del parámetro sobre el que se intentan hacer inferencias. Además, algunos conceptos de uso frecuente en mejora animal, como el de estimadores insesgados o la diferencia entre efectos fijos y aleatorios, no tienen cabida en la escuela bayesiana[4].

2. Los resultados aportados por estas escuelas no siempre coinciden, particularmente en muestras pequeñas y cuando la escuela bayesiana hace uso de información *a priori* (ver, por ejemplo, WANG *et al.*, 1994).

3. Algunos problemas mal resueltos en la escuela frecuentista o en la que utiliza la verosimilitud tienen solución en la escuela bayesiana, como hemos comentado al principio al exponer la cita de GIANOLA *et al.* (1994).

Dadas las ventajas claras de la escuela bayesiana en los puntos 1 y 3, surgen dos cuestiones obvias: ¿Por qué aparecen ahora los métodos bayesianos en mejora animal si la metodología bayesiana es incluso más antigua que los análisis frecuentistas? Y si es una metodología que aventaja sustancialmente a las que hasta ahora se utilizan, ¿por qué no las ha desplazado?[5]. La respuesta a la primera pregunta es más simple que la repuesta a la segunda. Operativamente hablando, la metodología bayesiana requiere habitualmente, y particularmente en el caso de la mejora animal, la resolución de integrales complicadas frecuentemente sólo accesibles por métodos aproximados (ver, por ejemplo, CANTET *et al.*, 1992). La aparición de métodos nuevos para obtener las distribuciones de probabilidad bayesianas (singularmente el muestreo de GIBBS; ver SORENSEN *et al.*, 1994 y GIANOLA *et al.*, 1994) ha facilitado enormemente el abordaje de estos problemas, haciéndolos operativamente factibles. La respuesta a la segunda pregunta requiere un examen detallado de las características de las dos escuelas en su forma de abordar la inferencia y los problemas de decisión[6]. El objetivo de este artículo es exponer los puntos de vista de estas dos escuelas de inferencia en lo que atañe a los métodos utilizados hoy en mejora genética animal, con los pros y los contras de las distintas aproximaciones.

---

4. Esta última afirmación puede resultar chocante al mejorador habituado a trabajar con los métodos BLUP (Best Linear Unbiased Predicition, HENDERSON 1973) o REML (máxima verosimilitud residual, también llamada máxima verosimilitud modificada o restringida, PATTERSON y THOMPSON, 1971). El propio Henderson ha comentado muchas veces que métodos sesgados pueden ser más eficaces que el BLUP (ver HENDERSON, 1984, por ejemplo). Fisher y Yates consideraban que la distinción entre efectos fijos y aleatorios aportaba más confusión que claridad al análisis de la varianza (ver, p. ej. el prólogo de Yates a la reedición de las obras de FISHER, 1990), y Fisher consideraba la propiedad de insesgamiento un tanto irrelevante (FISHER, 1956, pp. 146-147), esencialmente debido a que no era invariante funcional; (por ejemplo, si bien la esperanza de la media muestral es la media poblacional, la esperanza del cuadrado de la media muestral no es el cuadrado de la media poblacional).

5. O, como se preguntaba Brad Efron en un breve artículo del que lo más interesante es la discusión que lo acompaña, ¿por qué no somos todos bayesianos? (EFRON, 1986).

6. Hay una respuesta más obvia, los mejoradores son educados en Universidades donde sólo se les explica el método frecuentista, presentándose generalmente una caricatura del método bayesiano. De todas formas el fenómeno no es ajeno al "revival" que vive la escuela Bayesiana desde mediados de los 50.

La escuela frecuentista

La escuela frecuentista desarrolló sus métodos en los años 30 y 40. Los nombres más notorios de esta escuela son el polaco Jerzy Neyman, el británico Egon Pearson –hijo de Karl Pearson, uno de los fundadores de la estadística moderna–, quienes trataron los problemas de inferencia, y Abraham de Wald que trató los problemas de teoría de la decisión.

Inferencia y precisión

Aunque no es en realidad necesario exponer sus fundamentos, dado que sus métodos se explican de forma estándar hoy en día en la mayor parte de cursos de estadística, comentaré los dos procedimientos de inferencia usados en Mejora Animal: la estimación[7] y el test de hipótesis (aunque algunos artículos recientes están introduciendo la teoría de la decisión en problemas de Mejora, la producción científica en este campo es aún muy reducida; ver por ejemplo, WOOLIAMS y MEUWISSEN, 1993). Se considera que el valor del parámetro que se quiere estimar (la heredabilidad de un carácter en una población, por ejemplo) es una cantidad fija determinada, y la inferencia se basa en observar cómo se distribuirían hipotéticas estimas del parámetro de interés si la experiencia se repitiera infinitas veces.

Intervalos de confianza: Si el experimento se repitiera infinitas veces aparecerían infinitos intervalos de confianza que incluirían al valor del parámetro en un 95% de los casos. Obsérvese que cuando se da un intervalo de confianza, por ejemplo $h^2 = 0.20 \pm 0.05$, no se está indicando que la probabilidad de que la heredabilidad se encuentre entre 0.15 y 0.25 es del 95%, sino que si se repitiera el experimento infinitas veces aparecerían infinitos intervalos de confianza (de los que el intervalo [0.15, 0.25] es un ejemplo) que contendrían al valor verdadero en un 95% de los casos.

Test de hipótesis: Asumiendo unos riesgos determinados de considerar una hipótesis verdadera cuando no lo es, o de rechazarla cuando es verdadera, se crean unas regiones de aceptación y de rechazo sobre la distribución de los resultados de hipotéticas infinitas repeticiones del experimento, bajo la hipótesis que se quiere contrastar y bajo su alternativa, y se observa si la muestra cae en la región de aceptación o de rechazo . Por ejemplo, cuando la hipótesis es que dos razas tienen la misma media para un determinado carácter, si la diferencia que se observa es lo suficientemente pequeña como para no caer en la región de rechazo, bien se acepta que efectivamente las dos razas tienen la misma media o bien simplemente no se afirma que las razas sean diferentes (dependiendo de la potencia[8] del test). Obsérvese que aumentando el tamaño de la muestra se acabaría por rechazar siempre esta hipótesis, por lo que en experimento bien diseñados la cuestión de si dos razas tienen la misma media para un carácter o no, es una cuestión de tipo práctico: antes de empezar el experimento

7. No distinguiré entre estimación puntual y por intervalos porque esta distinción no es necesaria para los propósitos del artículo. Fisher, con la aspereza que le caracterizaba, acusó a Neyman y Pearson de inventar el concepto de estimación puntual sólo para, confiando en la ignorancia de los estudiantes, pretender haber traído algo original a la estadística con el concepto de intervalos de confianza (FISHER, 1956, p. 143).

8. El test que considera dos riesgos es tal vez el más frecuente, aunque experimentadores hay que diseñan con potencias del 50%, lo que equivale a ignorar el de segunda especie.

se debe decidir qué valor es lo suficientemente relevante como para considerar esas medias diferentes. El problema surge en experimentos no bien diseñados, en datos de campo o en los caracteres para los cuales no se diseñó el experimento, en esos casos la respuesta del test (aceptar o rechazar si dos valores son diferentes) es insatisfactoria .

Los índices de selección y el BLUP como estimadores frecuentistas

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). El BLUP aparece, sin este nombre, en 1949 como un método máximo verosímil, lo que resultó no ser cierto (HENDERSON, 1949, 1950) y condujo a que no se le prestara una importancia especial hasta años más tarde. La historia de su concepción y desarrollo está narrada por HENDERSON (1973). Examinaremos aquí las propiedades frecuentistas de estos métodos.

Consideremos el modelo lineal

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{u} + \mathbf{e} \qquad (1)$$

donde y es el vector de datos, **b** el de efectos fijos (por ejemplo, rebaño-año-estación, covariables, etc), **u** el de efectos aleatorios (valores de mejora de los individuos), **e** el de residuos y X, Z son matrices de incidencia. El estimador

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

donde **A** y **V** son las matrices de varianzas-covarianzas de los efectos aleatorios y de los datos respectivamente, y $\hat{\mathbf{b}}$ es la estima por mínimos cuadrados de **b,** da la *mejor predicción*[9] *lineal insesgada (Best Linear Unbiased Prediction)* de **u.** Caso de que no existan efectos fijos, suprimiendo $\mathbf{X}\hat{\mathbf{b}}$ de la fórmula se obtiene la expresión de los índices de selección, por lo que estos pueden considerarse un caso particular.

Invertir **V** no es sencillo cuando se disponen de miles de datos, como es el caso habitual de la mejora genética animal. HENDERSON (1963) demostró que la solución de las ecuaciones

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

coincidía con las estimas de $\hat{\mathbf{b}}$ y $\hat{\mathbf{u}}$ anteriores, y posteriormente (HENDERSON, 1976) mostró cómo calcular $\mathbf{A}^{-1}$ sin necesidad de invertir **A,** lo que convirtió en factible al método, que hoy ha alcanzado una difusión universal entre los mejoradores.

El BLUP es, sin embargo, un estimador insesgado en un sentido diferente al que se le da en la estadística frecuentista. Tomemos uno de los componentes de $\hat{\mathbf{b}}$ cualquiera, al que llamaremos también $\hat{b}$ (sin subíndice, por comodidad en la notación). Si tomásemos infinitas muestras obtendríamos infinitos $\hat{b}$ que cumplirían

$$E(\hat{b}) = b$$

es decir; en cada repetición hipotética del experimento los valores de $\hat{b}$ que se obtienen se agrupan en torno al valor verdadero b. Sin embargo esta condición es imposible

9. Henderson introdujo la costumbre de llamar *predicciones* a las estimas de los valores aleatorios. Parece innecesario.

de cumplir para los valores aleatorios u, puesto que, por ser aleatorios, cambian en cada repetición hipotética del experimento. Cuando los valores son aleatorios se dice que su estimador es insesgado si

$$E(\hat{u}) = E(u)$$

Esta condición es mucho menos sugestiva que la anterior. Ahora el estimador no se distribuye alrededor del valor verdadero, puesto que en cada repetición del experimento no sólo tengo un û diferente sino que también u ha cambiado. Esto afecta también a la interpretación del riesgo de estimación, que es lo que en definitiva se trata de minimizar. En el caso anterior

$$\text{Riesgo } (\hat{b}) = [\text{sesgo } (\hat{b})\,]^2 + \text{varianza } (\hat{b})$$

ahora

$$\text{Riesgo } (\hat{u}) = [\text{sesgo}(\hat{u})]^2 + \text{varianza } (\hat{u}) + \text{varianza } (u)$$

La cantidad varianza (u)[10] no se puede modificar, puesto que depende de la naturaleza (habitualmente será la varianza aditiva, dominante o de efectos maternos). El problema está, por otra parte, indeterminado, puesto que hay infinitos estimadores que, según su sesgo y varianza, tienen el mismo valor de riesgo. Para determinar el problema se elige el de varianza mínima de entre los que tienen sesgo nulo (en el sentido que hemos dado a la palabra sesgo cuando hablamos de factores aleatorios) .

Para poder utilizar el método BLUP hay que conocer las matrices A y V, de lo contrario el método ya no es BLUP. Como su valor verdadero se desconoce, se han hecho varias propuestas para que el método conserve buenas propiedades, más adelante volveré sobre este asunto[11].

Los métodos basados en esperanzas de formas cuadráticas

El estudio de los componentes de la variación (ANOVA) fue formalizado por FISHER en 1922 sobre bases muy intuitivas, como le gustaba subrayar al propio Fisher. Su formalización se hizo para modelos equilibrados con factores aleatorios únicamente. Tras varios intentos de generalización a modelos desequilibrados por varios autores, HENDERSON publica en 1953 tres métodos de análisis de varianza de dificultad de computación creciente: el primero trata del análisis de modelos desequilibrados con factores aleatorios y los otros dos ya tratan de modelos en los que hay efectos fijos y aleatorios simultáneamente –modelos mixtos–. El método 2 no admite interacciones entre efectos fijos y aleatorios, pero el método 3 resuelve cualquier modelo. Estos métodos son implementados por HARVEY (1960) para estimar heredabilidades y correlaciones genéticas en un progra-

10. Aquí sesgo significa E(û) - E(u), que es una expresión diferente a la que habitualmente utiliza la estadística frecuentista. En el caso de efectos aleatorios el sesgo debería ser E(û|u) - u, valor distinto de cero, con lo que el método BLUP sería un método sesgado. Llamar al BLUP insesgado porque se llama sesgo a otra cosa, parece un uso un tanto liberal del lenguaje.

11. Es tan habitual considerar al BLUP como "el mejor estimador", que se olvida a veces que sólo es el mejor de entre los de una clase muy concreta, y sólo cuando se conocen las componentes de varianza de los efectos aleatorios sin error, por lo que el uso de la palabra "Best" resulta un tanto equívoco.

ma que fue muy popular entre los genetistas[12].

Posteriormente se demostró que el método de Fisher era óptimo para diseños equilibrados, pero de los métodos de Henderson, aparte de ser insesgados no se conoce que tengan alguna otra propiedad favorable. Otro intento de encontrar estimas con propiedades adecuadas es el desarrollado por RAO en 1971 con sus estimas insesgadas de varianza mínima (MIVQUE). Este método no es iterativo, pero es función de unos valores previos. El método produce estimas insesgadas de error mínimo sólo cuando los valores previos que se le suministran son los verdaderos componentes de varianza, lo que –al menos para el profano– resulta un tanto peculiar[13]. El método se basa, además, en la normalidad de los caracteres que se analizan. Cuando no es así, el método se llama MINQUE porque tiene la propiedad de que minimiza la norma (módulo al cuadrado) del vector que define la distancia entre el valor estimado y el verdadero. Esta última forma de proceder, un tanto oscura, no ha gozado de popularidad no ya entre los mejoradores sino ni siquiera entre los estadísticos profesionales.

Para resolver dificultades de cómputo, HENDERSON (1984) y SCHAEFFER (1986) presentaron sendos métodos basados también en esperanzas de formas cuadráticas. Todos estos métodos gozan de una propiedad típicamente frecuentista: el insesgamiento, pero no son necesariamente los estimadores más eficientes (de mínimo riesgo) y, sobre todo, tienen un inconveniente que los hace particularmente inadecuados para su uso en mejora genética animal: son métodos sesgados cuando los datos están seleccionados (ver, p. ej., GIANOLA *et al.*, 1989). Este inconveniente es importante, puesto que la mayor parte de bases de datos que se usa en mejora genética animal consiste en datos seleccionados. Escoger sólo a las generaciones base para hacer el análisis no resuelve el problema, puesto que en ese caso es probable que el número de datos sea insuficiente para obtener una precisión aceptable.

El método de máxima verosimilitud

El concepto de verosimilitud y el método de máxima verosimilitud fueron formalizados por FISHER entre 1912 y 1922. En

12. Una advertencia respecto a esta breve nota histórica y a las que siguen, precedentes hay de casi todo. Alguien aducirá que el MIVQUE fue propuesto por LAMOTTE en 1970 en notas internas de su Universidad, pero se puede aducir que en 1968 un estudiante de Searle terminó su tesis doctoral sobre MIVQUE para modelos sencillos. De hecho, rastreando el método, en 1906, en un tratado de geodesia, aparece la primera descripción de lo que hoy podemos llamar MIVQUE. En 1861 hay indicios del uso del ANOVA para un factor, y su autor dice que "no es procedimiento que sea nuevo" (ver SCHÉEFFE, 1959). Máxima verosimilitud restringida –con este nombre–se encuentra para un caso simple en THOMPSON –otro Thompson– 1962, y el título del artículo de PATTERSON y THOMPSON (1971) es casi el mismo que el del capítulo 24 del libro de ANDERSON y BANCROFT (1952), quienes ya sugirieron estos procedimientos. Por su parte el método de máxima verosimilitud, aunque propuesto por Fisher y examinadas sus propiedades en 1922, se encuentra antes en el propio FISHER (1912) y en varios autores desde BERNOUILLI (1782, en KENDALL, 1961).

13. Además pronto se descubrió que sustituyendo en la fórmula como valores iniciales los que se deberían obtener en la primera iteración, se obtenía una fórmula idéntica a una de las soluciones de algoritmos que dan lugar a estimas REML, por lo que –con la notable excepción del propio Rao, quien en 1988 produjo un libro sobre estimación de componentes de varianza prácticamente dedicado al MIVQUE (RAO y KLEFFE, 1988)–, el método ha caído en desuso en Mejora Animal.

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).

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(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(y|u). Se supone que f(y|u) es función de u para cada muestra 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.

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–[15].

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

15. Se genera, además, un problema de notación, puesto que hablar de varias funciones de densidad $f(y|m_1)$, $f(y|m_2)$, $f(y|m_3)$, ...para un y dado, es idéntico matemáticamente a hablar de una cierta función L(m|y), con la diferencia de que mientras f es una función de densidad, que representa probabilidades, L ni es de densidad ni representa probabilidades. Problema al que contribuyen textos clásicos como el de de Kendall, utilizando para la verosimilitud la notación $L(x|\theta)$ (STUART and ORD, 1991), o la de miembros de la escuela frecuentista distinguiendo entre "dado un parámetro" (que siempre está fijado) y "dada la muestra" (que es la variable aleatoria); para el

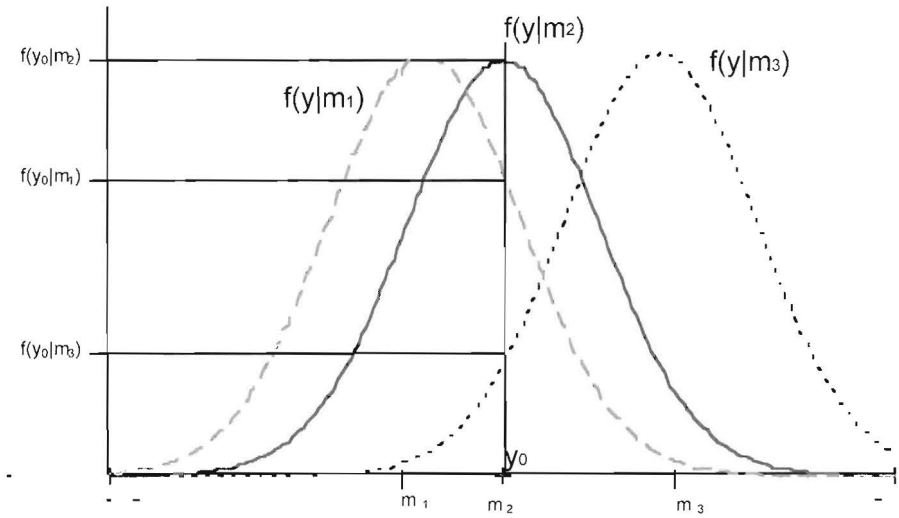


Figura 1. Valores de verosimilitud para distintas medias

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 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[16].

Cuando Fisher propuso el método no quedaba claro que este estimador fuera particularmente bueno. Fisher lo propuso porque le parecía que la verosimilitud suminis-

primer caso usan la notación $(x|\theta)$ y para el segundo $(\theta;x)$, con lo que se puede encontrar la verosimilitud con cuatro notaciones, $L(\theta|x)$, $L(x|\theta)$, $f(x|\theta)$ y $L(\theta;x)$. La primera de ellas es la que menos confusión ocasiona.

16. Suponiendo, claro está, que el sistema del que se muestrea genere las muestras con probabilidad máxima. Esta crítica la hizo Euler cuando Bernouilli presentó su trabajo, y es razonable, puesto que suponer esto puede conducir a sesgos en muestras pequeñas. Por ejemplo, en el caso de la varianza, si tomamos una muestra de dos individuos el estimador máximo verosímil tenderá a estimar varianzas pequeñas, puesto que en las distribuciones con varianzas más pequeñas los valores de $f(y|\sigma^2)$ son más elevados (de hecho esto es exactamente lo que ocurre, el estimador máximo verosímil de la varianza el la varianza muestral, s^2, que es sesgado y tiende a infraestimarla).

traba 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[17]. 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[18]. 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.

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[19],

17. De hecho VON MISES (1957, pp. 157-158) acusa a Fisher de exponer con gran cuidado las diferencias entre probabilidad y verosimilitud, para acto seguido usar la verosimilitud como se usa la palabra probabilidad en el sentido cotidiano. No le falta razón.

18. Un ejemplo conocido es la paradoja de James-Stein: supongamos que disponemos de K poblaciones $N(m_i, \sigma^2)$ independientes y se intenta estimar las m_i para que el riesgo $E[\Sigma(m_i - m_i)^2]$ sea mínimo. El estimador máximo verosímil está compuesto por las medias muestrales de cada población, $m=[\bar{x}_1, \ldots, \bar{x}_K]$. JAMES y STEIN (1961) encontraron un estimador de menor riesgo que el máximo verosímil para K>2.

19. "Cantidad de información" es un término con un contenido emocional que intentaré expresar de forma intuitiva. Llamando l al logaritmo de la verosimilitud, dl(u|y)/du es la pendiente de la curva del logaritmo de la 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

$$E[\delta \ell(u|\mathbf{y})/\delta u]^2$$

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)[20]. 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.

Si la verosimilitud indica un grado de creencia racional en una estimación, lo lógico sería utilizar el cociente entre dos verosimilitudes para decidir sobre la importancia de un efecto en un modelo. Lamentablemente ya hemos visto que las verosimilitudes no son cantidades que se puedan comparar como se comparan las probabilidades; una verosimilitud cuatro veces mayor que otra no proporciona un "grado de creencia racional" cuatro veces mayor. Sin embargo el cociente de verosimilitudes tiene propiedades frecuentistas interesantes y es habitual el uso del test de razones de verosimilitud. Si en un modelo se ignora un factor aleatorio, los cocientes de verosimilitudes se distribuirían, si se repitiera infinitas veces la experiencia, como una χ^2_1, lo que permite utilizarlo de forma frecuentista examinando las regiones de aceptación y rechazo para su uso[21].

El REML como estimador frecuentista

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.

---

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

$$L(q|x_1,\ldots x_n) = L(q|x_1) \cdot L(q|x_2) \cdot \ldots \cdot L(q|x_n),$$

por tanto

$$\log L(q|x_1,\ldots x_n) = \log L(q|x_1) + \log L(q|x_2) + \ldots + \log L(q|x_n)$$

20. Recuerdo que esta cantidad es equivalente matemáticamente a $-E[dL^2(u|\mathbf{y})/du^2]$, que es como suele encontrarse en la literatura (ver STUART y ORD, 1991, pp. 659).

21. EDWARDS (1992) recomienda, de todas formas, la comparación de verosimilitudes, pero basado en argumentos ajenos a la escuela frecuentista.

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.

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

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

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

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

El valor que hace máximo $L(\sigma^2|\mathbf{y})$ es

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

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,

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

con lo que la estima ML de la varianza es

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

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 **K,** el método consiste en hacer

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

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

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

Cumple la condición.

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

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

Si se hace la multiplicación $\mathbf{K'K}$ se deduce inmediatamente que

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

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

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

y aunque se puede demostrar formalmente para el caso general[22], es sencillo ver en nuestro ejemplo que

$$\mathbf{y'K(K'K)^{-1}K'y} = [\mathbf{y} - \mathbf{1}(1/n)\Sigma y_i]'$$
$$[\mathbf{y} - \mathbf{1}(1/n)\Sigma y_i] = (\mathbf{y} - \mathbf{1}\hat{\mu})'(\mathbf{y} - \mathbf{1}\hat{\mu})$$

y la estima REML de la varianza es

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

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 m por el estimado, sino que al deducir la fórmula del estimador de la varianza aparece una expresión, $(1/n)\Sigma y_i$, 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

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

mientras que la estima REML es

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

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

Con la matriz **K** propuesta, el análisis de la varianza se hace sobre un vector

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

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

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

con el mismo resultado. Esto no quiere decir que cualquier **K** valga. Una matriz con la mitad de sus filas compuestas por ceros también cumple $\mathbf{K'1} = 0$, 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 **K** el máximo número de contrastes lineales independientes. De hecho no importa la **K** concreta, puesto que en la fórmula final no aparece. Obsérvese que **K** es de dimensiones (n-1) x 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 **K'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.

---

22. Se puede demostrar que, en general,

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

En nuestro caso, $\mathbf{X} = \mathbf{1}$, $\mathbf{V} = \mathbf{I}\sigma^2$, con lo que

$\mathbf{X}(\mathbf{X'V^{-1}X})^{-1}\,\mathbf{X'V^{-1}y} = \mathbf{1}\,[(1/\sigma^2)\mathbf{1'1}]^{-1}\,(1/\sigma^2)\mathbf{1'y} = 1\,[(1/\sigma^2)\,n\,]^{-1}\,(1/\sigma^2)\Sigma y_i = \mathbf{1}\,(1/n)\Sigma y_i$

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. Sin embargo el REML se ha impuesto como el "método a escoger", tal vez, parafraseando a Dawid en la cita que da comienzo a este artículo, porque "todo el mundo sabe que es algo que está muy bien".

La escuela 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, formalizando el mismo principio de inferencia[23]. 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 GIANOLA *et al.*, 1996 para una revisión sobre los métodos bayesianos en mejora genética animal).

Inferencia y precisión

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 $f(h^2|y)$ siendo y el vector de valores observados (figura 2). 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

23. Bayes no llegó a publicar ningún trabajo matemático en vida y era miembro de la Royal Society por unos estudios sobre metafísica. El manuscrito original no se conserva y debemos confiar en la adscripción a Bayes que su amigo R. Price hizo al presentar la publicación en la Royal Society. Aparentemente el principio había sido formulado con anterioridad, y STIGLER (1983) se lo atribuye a Saunderson, un profesor de óptica ciego autor de un gran número de publicaciones sobre varias ramas de las matemáticas.

probabilidad del 95%, o qué probabilidad tiene el que h^2 esté entre tal y tal valor. En la Figura 2 el área entre los valores 0.1 y 0.3 es la probabilidad de que h^2 se encuentre entre esos dos valores. 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 $f(h^2|y)$ que pueden ser usados como estimación puntual, y cuyo uso depende de la preferencia del investigador. Por ejemplo,

– *La moda:* Es el valor de h^2 que maximiza $f(h^2|\mathbf{y})$; esto es, el valor más probable de h^2 dada la muestra y. Puede demostrarse que es el valor que maximiza el riesgo de estimación cuando la función de pérdidas es del tipo todo-nada; esto es, la función de pérdidas toma valor 0 cuando $h^2 = \hat{h}^2$ y valor 1 en caso contrario, lo que equivale en realidad a no asignar una función de pérdidas definida.

– *La mediana:* Es tan probable que el valor verdadero sea superior como inferior a esta estima. Minimiza el riesgo de estimación cuando la función de pérdidas es $|h^2 - \hat{h}^2|$.

– *La media:* Es el estimador que minimiza el riesgo mínimo cuadrático, que es por otra parte el de uso más corriente en mejora animal: $E(h^2 - \hat{h}^2)^2$.

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,

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

Obsérvese que $f(h^2|\mathbf{y})$ es una función de h^2, pero no de **y,** que está fijada; por tanto $f(\mathbf{y}|h^2)$ es aquí función de h^2, pero no de y, que es exactamente la definición que antes vimos de verosimilitud. Por la misma razón $f(\mathbf{y})$ 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[24], llamada *a priori*[25] 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.

Obsérvese que:

1) Es discutible qué valor de estimación puntual de h^2 se debe ofrecer. En casos

24. Obviamente no hay probabilidades "absolutas" al margen de cualquier experimento. Los valores de la heredabilidad serán más o menos probable dadas unas circunstancias determinadas. Está mal escrito, por tanto, $f(h^2)$, y debiera escribirse siempre $f(h^2|H)$, donde H es el conjunto de hipótesis bajo las cuales los valores de h2 son más o menos probables. El teorema de Bayes se reescribiría, pues, conteniendo en cada probabilidad un "dado H", pero no se hace por simplificar la escritura.

25. JEFFREYS (1931) distingue entre información previa ("prior") e información *a priori*. Para esta exposición la distinción no me parece relevante.

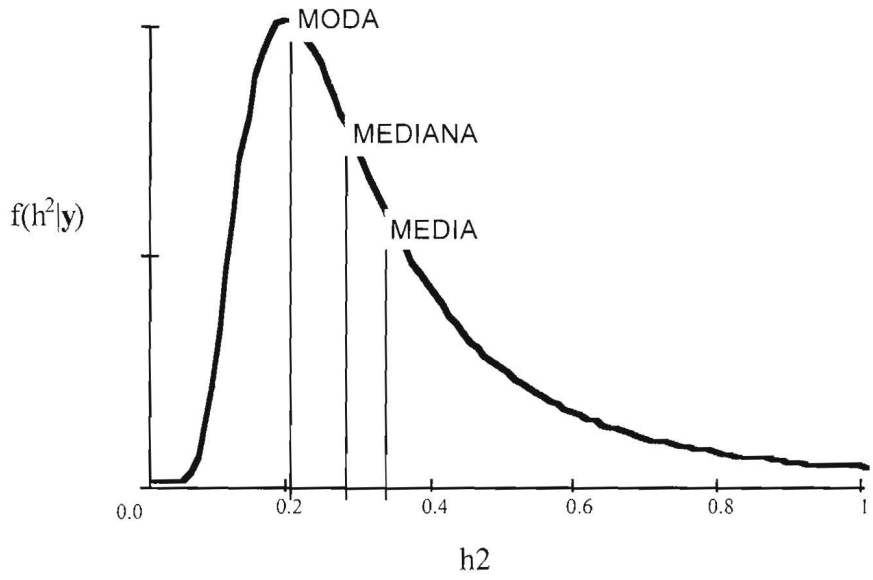


Figura 2. Densidad posterior de la heredabilidad

extremos, en los que la función de densidad posterior sea bimodal, podría ocurrir que el valor medio, que es el que minimiza el riesgo cuadrático, fuera mucho menos probable que los valores de las modas (figura 3). En este caso una estimación puntual no describe bien la incertidumbre en torno al verdadero valor de h^2.

2) Aunque la función de densidad posterior sea asimétrica, se puede calcular siempre un intervalo de confianza[26]. Si se desea que el valor más probable esté en el centro del intervalo, en ese caso el intervalo es asimétrico[27].

3) Al no haber repeticiones hipotéticas del experimento, tampoco hay nada que

26. Los bayesianos suelen hablar de intervalos o regiones 'de credibilidad', más que 'de confianza', puesto que son regiones en las que el verdadero valor del parámetro se encuentra con una probabilidad determinada.

27. Una dificultad adicional, si se quiere usar el sistema de intervalos de confianza para la inferencia científica, es que hay que conocer exactamente la densidad posterior, por lo que hay que conocer el valor de f(y). Este valor es difícil de computar porque

$$f(\mathbf{y}) = \int f(\mathbf{y},h^2)\, f(h^2) d\,(h^2)$$

como y es un vector, esta integral es multidimensional. Aunque los elementos de y fueran independientes, esta integral pasaría a ser una integral múltiple de tantas dimensiones como datos, y en ambos casos el problema es irresoluble incluso por métodos numéricos. Hoy en día las técnicas de muestreo de Gibbs que se mencionan en el apéndice I han resuelto este problema.

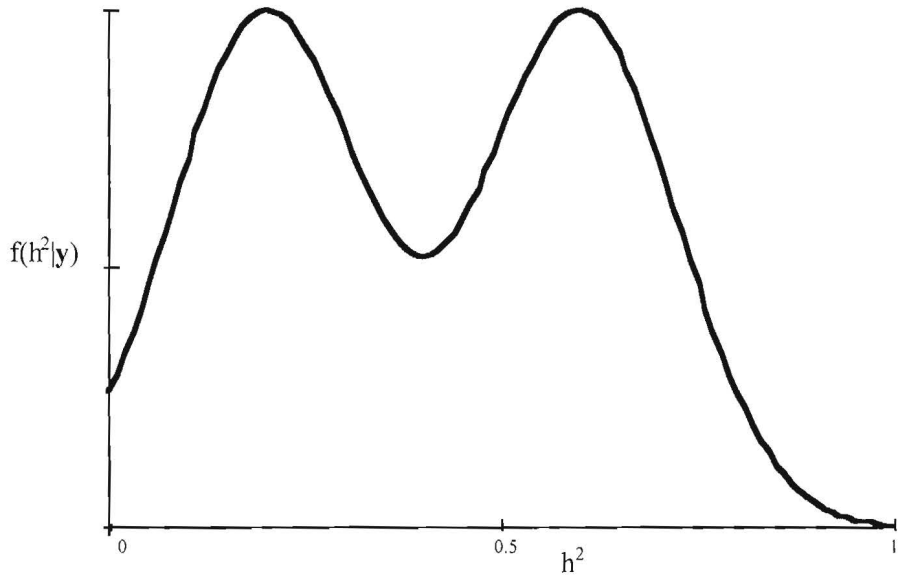


Figura 3. Densidad posterior bimodal de una heredabilidad

pueda ser considerado como "sesgo". Tampoco hay factores fijos, dado que a todos los factores desconocidos se les asigna una distribución aleatoria a priori basada en las opiniones del investigador.

Los tests de hipótesis tienen una interpretación más clara en la escuela bayesiana. Por ejemplo, se desea probar si las medias de dos razas son iguales. Para ello se calcula el cociente entre la probabilidad de que sean iguales y la probabilidad de que difieran:

$$P(m_1 - m_2 = 0|\mathbf{y}) / P(m_1 - m_2 \neq 0|\mathbf{y})$$

si este cociente es igual a 7, por ejemplo, esto quiere decir que la probabilidad de que ambas medias sean iguales es 7 veces mayor que la probabilidad de que sean diferentes.

El uso de la información previa

Cuando claramente hay información a priori

En ocasiones existe información a priori sobre los parámetros a estimar. En esos casos el uso de esta información previa no representa ningún problema, como veremos a continuación con un ejemplo de Fisher (para quien, dicho sea de paso, el bayesianismo era una auténtica "bestia negra")[28].

28. El lector puede creer que exagero. LANE (1980) usa exactamente esta expresión al revisar las opiniones de Fisher sobre el bayesianismo.

Supongamos un carácter controlado por un gen con dos alelos y con dominancia, como el caso del color negro que domina sobre el marrón de la piel de ciertos ratones. Se desea averiguar si un ratón negro proveniente de un apareamiento entre heterocigotos es homocigoto o heterocigoto. Para ello se cruza ese ratón negro con uno marrón, y se obtienen siete descendientes, todos ellos negros.

La probabilidad a priori de que un ratón negro proveniente de un apareamiento entre heterocigotos sea homocigoto es 1/3, puesto que un apareamiento Aa x Aa da tres tipos de animales negros: AA, Aa, aA.

$$f(AA) = 1/3$$

A partir de ahora representaremos por Aa a los animales heterocigotos, sean estos Aa o aA. La probabilidad a priori de que un ratón negro proveniente de un apareamiento entre heterocigotos sea heterocigoto es, por la misma razón, 2/3.

$$f(Aa) = 2/3$$

Una vez realizado el experimento, la verosimilitud de que el ratón negro sea homocigoto es 1; es decir, dado que se observan siete descendientes negros, en el caso de que el padre negro fuera homocigoto (AA), la probabilidad de ese suceso sería 1, puesto que los descendientes del cruce con el ratón marrón (aa) cabe esperar que sean negros (Aa).

$$L(AA|\ y = 7 \text{ descendientes, todos negros}) = \text{Prob}(y = 7 \text{ descendientes, todos negros} \mid AA) = 1$$

Una vez realizado el experimento, la verosimilitud de que el ratón negro sea heterocigoto es $(1/2)^7 = 1/128$; es decir, dado que se observan siete descendientes negros, en el caso de que el padre negro fuera heterocigoto (Aa), la probabilidad de ese suceso sería 1/2, puesto que los descendientes del cruce con el ratón marrón (aa) cabe esperar que sean negros (Aa) y marrones (aa) en la misma proporción.

$$L(Aa|\ y = 7 \text{ descendientes negros}) = \text{Prob}(y = 7 \text{ descendientes todos negros} \mid Aa) = \frac{1}{128}$$

Por tanto, la probabilidad de que el padre sea AA debiera ser

$$L(AA|y) \cdot f(AA) = 1 \cdot 1/3 = 1/3$$

y de que sea heterocigoto

$$L(Aa|y) \cdot f(Aa) = (1/128)\,(2/3) = 1/192$$

pero como estas cantidades no suman 1 y ambos son sucesos excluyentes que no admiten otra posibilidad, hay que dividir por la suma de ambas:

$$f(y) = \text{Prob}(7 \text{ descendientes todos negros}) = \text{Prob}(\text{padre AA}) + \text{Prob}(\text{padre Aa}) = (1/3) + (1/192)$$

Probabilidad de padre homocigoto, dado que se obtuvieron siete descendientes negros al cruzar con un ratón marrón

$$f(AA|y) = L(AA|y) \cdot f(AA) / f(y) = (1/3) / (1/3 + 1/192)$$

Probabilidad de padre heterocigoto, dado que se obtuvieron siete descendientes negros al cruzar con un ratón marrón

$$f(Aa|y) = L(Aa|y) \cdot f(Aa) = (1/192) / (1/3 + 1/192)$$

que es la respuesta buscada.

Cuando la información a priori se admite como conocida

Es frecuente en mejora animal admitir que se conocen los parámetros genéticos, por ejemplo para usarlos en estimaciones BLUP o en índices de selección, aunque

sólo se tengan estimas de los mismos. En el caso de los índices se requiere además el conocimiento de las medias (es decir, de los efectos fijos). Este uso de la información previa está admitido en mejora genética animal, y no genera ninguna polémica. Consideremos el modelo

$$\mathbf{y} = \mathbf{Z}\,\mathbf{u} + \mathbf{e}$$

donde **y** es el vector de datos, **u** el valor aditivo de un individuo, **e** la componente ambiental y **Z** la matriz de incidencia. Hallaremos la función de densidad posterior f(**u**|**y**).

Se considera que **u** se distribuye de forma Normal con media cero y matriz de varianzas-covarianzas genéticas conocida **A**

$$f(\mathbf{u}) = cte \cdot \exp\{(-1/2\,\sigma_u^2)\,\mathbf{u}'\mathbf{A}^{-1}\,\mathbf{u}\}$$

$$f(\mathbf{y}|\mathbf{u})_{\text{como función de u}} = cte \cdot \exp\{(-1/2\,\sigma_e^2)\,(\mathbf{y} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Zu})\}$$

$$f(\mathbf{u}|\mathbf{y}) = f(\mathbf{y}|\mathbf{u})_{\text{como función de u}} \cdot f(\mathbf{u}) =$$

$$= cte \cdot \exp\{(-1/2\,\sigma_e^2)\,[(\mathbf{y} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Zu}) + \mathbf{u}'\mathbf{A}^{-1}\,\mathbf{u}\,\alpha]\}$$

donde $\alpha = \sigma_e^2 / \sigma_u^2$

Esta función de densidad posterior f(u|y) tiene el máximo en

$$\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\,\alpha)^{-1}\,\mathbf{Z}'\,\mathbf{y}$$

lo que corresponde a la mejor predicción lineal de u, y es un resultado conocido de los índices de selección. Como la distribución es Normal, el máximo corresponde a la moda, la media y la mediana simultáneamente. A partir de f(u|y) se pueden hacer predicciones como hemos visto anteriormente.

Cuando la información a priori es vaga

En muchas ocasiones es difícil cuantificar la información a priori de una forma tan objetiva como la de los ejemplos que acabo 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:

1) Definiendo la probabilidad como un estado de creencias del investigador

La probabilidad definida como la frecuencia con la que se puede presentar un suceso es un concepto demasiado restrictivo como para dar solución a nuestro problema, por lo que se considera que la probabilidad es un valor subjetivo que expresa la opinión del investigador acerca del grado de conocimiento existente sobre el parámetro que se desea estimar. Probabilidad subjetiva no quiere decir probabilidad arbitraria. Como hace notar KEYNES (1921), esta opinión subjetiva debe estar fundada de forma tal que el investigador pueda persuadir a los científicos de su entorno de que esa opinión es razonable.

Una forma de imaginar cómo puede establecerse esta probabilidad es pensar, de momento, en un problema distinto. Supongamos que intentamos evaluar el precio de un producto –p.ej. el kg de conejo–. Al margen de cualquier muestra que se tome, podemos afirmar que será menos de 10.000 pts el kg y más de 0 pts. Podemos afinar más diciendo que es poco probable que sea de 100 ptas./kg y también que es poco probable que esté a 1000 ptas./kg. Continuando, podemos poco a poco establecer una serie de sucesos que nos parecen igualmente probables, o unos más que otros, y llegar

a producir una distribución más o menos escalonada que será considerada como aceptable por muchos interlocutores. Esto, que es sencillo para problemas univariantes, se vuelve prácticamente imposible para problemas multivariantes. Por ejemplo, si hablamos de tres caracteres con varios efectos ambientales (lo que se llaman efectos fijos en el contexto frecuentista), hace falta determinar la probabilidad a priori que opinamos sobre el caso "heredabilidad del primer carácter 0,1, del segundo 0,2, del tercero 0,2, correlación genética del primero con el segundo -0,7, del segundo con el tercero 0,1, del primero con el tercero de 0,1, para un efecto de estación de -0,5 para el primer carácter, de 7,2 para el segundo, etc. ¡Y esto debe hacerse para cada caso posible! Determinar todas estas opiniones puede conducir al investigador a una crisis psicológica, por lo que en la mayor parte de problemas de mejora genética animal esta determinación de la opinión a priori debe ser abandonada.

2) 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 arbitraria, y habitualmente se procura que facilite los cálculos de la función posterior y que no conduzca a paradojas o a resultados inadmisibles. Este argumento fue introducido en la literatura bayesiana por JEFFREYS (1931) y ha tomado popularidad en los últimos años ante las dificultades que presenta el uso de la información previa cuando ésta es vaga. Su dificultad principal reside en la determinación del tamaño de la muestra en experimentos con pocos datos, en los que esa densidad previa puede influir el resultado final de forma relevante.

Cuando no hay información a priori

Hasta aquí, si bien discutible el procedimiento sobre cómo representar las creencias, no se presenta ninguna dificultad de tipo teórico. Habrá investigadores que prefieran no considerar a la probabilidad como un fenómeno subjetivo, pero si se da este paso adelante, la inferencia que se produce no da lugar a paradojas notables ni a contradicciones. El problema aparece al intentar representar la ignorancia *a priori*. Los primeros bayesianos (Bayes incluido) usaron una técnica a la que Keynes denominó *Principio de Indiferencia*[29], y que consiste en admitir el siguiente postulado: *Cuando no hay información sobre los distintos sucesos posibles, todos tienen la misma probabilidad a priori.* Admitir este postulado equivale a decir que en el viejo saco que contiene bolas blancas y negras usado para enseñar probabilidades, si no dispongo de ninguna información sobre en que proporción están, mi creencia a priori es que cual-

29. KEYNES (1921) ilustra su capítulo sobre el principio de indiferencia con una cita interesante que transcribo,

ABSOLUTO: Seguro, Señor, que no es muy razonable entregar mi amor a una dama de la que no conozco nada.

SIR ANTHONY: Estoy seguro, Señor, que es menos razonable poner objeciones a una dama de la que usted nada conoce.

quier combinación posible tiene la misma probabilidad. 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*. Este terrible postulado es origen no sólo de problemas ontológicos, sino matemáticos. Por ejemplo, si quiero representar ignorancia total sobre la heredabilidad, puedo decir que puede tomar cualquier valor en el intervalo [0,1] con la misma probabilidad. Entonces,

$$P(h^2 < 0,5) = 1/2$$

Pero el suceso $h^2 < 0,5$ es el mismo que $h^4 < 0,25$, por lo que su probabilidad es la misma:

$$P(h^4 < 0,25) = 1/2$$

Esto indica que es más probable que h^4 se encuentre cerca de 0 que cerca de 1, luego si admitimos este resultado queremos decir ni más ni menos que sobre h^2 no hay información a priori, pero sobre h^4 sí, lo que es absurdo. No hay una respuesta sencilla a este problema. Las soluciones propuestas son:

1) Admitir la validez de los a prioris planos

Es razonable representar la ausencia de información de forma plana sobre h^2 y no sobre h^4. Si admitiéramos que h^{16} tiene una densidad uniforme entre 0 y 1 ello implicaría que $P(h^{16} > 0,5) = 1/2$, por lo que consecuentemente $P(h^2>0,5^{1/8}) = P(h^2>0,92) = 1/2$, algo que no estaríamos dispuestos a admitir. Esta respuesta, propuesta por Box y Tiao (1973), resulta poco convincente. Como Edwards (1992) señala, parece lógico admitir que la ausencia de información sobre un parámetro implica también la ausencia de información sobre cualquier transformación. Decir que ignoramos qué probabilidades tienen los valores de un parámetro no parece lo mismo que decir que todos ellos tienen la misma probabilidad.

2) La ignorancia se representa respecto a la verosimilitud

Hay simplemente un problema de escala. Cambiando la escala es posible encontrar funciones de densidad a priori inmunes a reparametrizaciones. Para ello en primer lugar hay que definir bien qué quiere decir ausencia de información. De la expresión del teorema de Bayes vemos que la información de los datos está contenida en la verosimilitud $f(y|h^2)$, mientras que la densidad a priori $f(h^2)$ es independiente de los datos. Una distribución de densidad de probabilidades que exprese ignorancia se debe expresar en una escala en la cual la verosimilitud no cambie. Esta solución fue propuesta inicialmente por Jeffreys (1961) y está desarrollada con detalle en Box y Tiao (1973, cap. 1). En principio no siempre es posible encontrar una escala así, por lo que se recurre a aproximaciones que son buenas cuando el tamaño de la muestra es grande (Box y Tiao dicen de tamaño moderado, sin duda para evitar las críticas consecuentes). En el caso de que se intente expresar ignorancia de dos parámetros simultáneamente, por ejemplo, la media y la varianza, estas funciones no existen o conducen a paradojas. Las funciones a priori de Jeffreys no han tenido éxito en Mejora animal, a pesar de ser ampliamente usadas en otros campos de la estadística, precisamente porque los problemas de mejora animal suelen ser multivariantes de una u otra forma (hay efectos ambientales además de los genéticos, en muchas ocasiones los problemas son multicarácter, etc.).

Sucede también que muchas de las funciones de densidad a priori propuestas no son verdaderas probabilidades, puesto que su integral no es 1; por ejemplo, la función plana que da el mismo valor de probabilidad a todos los puntos entre $-\infty$ y $+\infty$. Estas funciones pueden en ocasiones dar lugar a densidades posteriores que no integren a 1 y que no sean, por tanto, verdaderas probabilidades, con lo que la inferencia se vuelve imposible. En muchas ocasiones es difícil de detectar que la función posterior encontrada no es una auténctica función de densidad, con lo que se corre el riesgo de realizar afirmaciones probabilísticas falsas. En un artículo reciente, HOBERT y CASELLA (1997) recomiendan utilizar densidades a priori propias para evitar estos problemas. En cualquier caso, si utilizamos densidades a priori impropias persiste siempre el problema ontológico de considerar una incertidumbre a priori mediante una función que no es una verdadera probabilidad.

3) Hallar una densidad posterior en la que la información a priori sea mínima

Para casos en los que haya objetivamente una ignorancia a priori, BERNARDO (1979) propuso el cálculo de funciones de densidad posterior de forma que las funciones a priori debían dar el máximo peso posible a los datos, lo que tiene una justificación intuitiva aparentemente razonable. La ventaja de este planteamiento es que se calcula la densidad posterior directamente, lo que asegura que se trata de una verdadera probabilidad. En cualquier caso, para quienes propugnan esta solución,

"El problema de caracterizar la ignorancia a priori o la información vaga a priori es mucho más complejo de lo que intuitivamente parece... 'vago' es un concepto demasiado vago para que pueda ser útil. No hay un a priori objetivo que represente la ignorancia... el *a priori* de referencia componente del análisis es simplemente una herramienta matemática" (BERNARDO y SMITH, 1994, pp. 298).

El principal problema de esta solución es operativo: estos aprioris son difíciles de determinar, y hay que calcularlos para cada modelo[30].

El BLUP y los índices de selección como estimadores bayesianos

RONNINGEN (1971) primero y DEMPFLE (1977) después, observaron que el BLUP puede considerarse como un estimador bayesiano particular. Dado el modelo mixto (1), la forma bayesiana de plantear la estimación del valor de mejora en animales consiste en estimar la función de densidad f(u|y) para poder hacer inferencias. En un contexto bayesiano no hay "efectos fijos", todos los efectos son aleatorios, por lo que el caso de índices de selección vuelve a ser un caso particular en el que no existe el vector **b.**

30. En el caso multivariante debe distinguirse entre el parámetro de interés y los de ruido. Los a prioris de referencia de Bernardo han sido criticados porque cambian según se dispone el orden de parámetros en el que se realiza el análisis (ver, p. ej. ROBERT, 1992, pp. 109-111). Naturalmente, si se admite que son meras herramientas matemáticas sin valor ontológico, este es un problema menor. Una dificultad seria reside en que aparentemente el cálculo de estas densidades posteriores es extremadamente complejo, particularmente cuando los datos están seleccionados y, como es habitual, el proceso de selección no se puede describir de forma sencilla (nunca se aplica la misma barrera de selección, los individuos tienen información variable, etc.). Aparecen en el proceso integrales múltiples difíciles de resolver.

Trataremos, pues de encontrar un estimador para el vector

$$\mathbf{t'} = [\,\mathbf{b'}\ \mathbf{u'}\,]$$

a partir de los datos y. Para ello determinaremos primero $f(\mathbf{t} \mid \mathbf{y})$. Sabemos que

$$f(\mathbf{t}|\mathbf{y}) = \text{cte.}\ f(\mathbf{y}|\mathbf{t})\ f(\mathbf{t})$$

Si llamamos $\mathbf{W} = [\mathbf{X}\ \mathbf{Z}]$ y suponemos que $f(\mathbf{y}|\mathbf{t}) \sim N([\mathbf{Xb}+\mathbf{Zu}], \mathbf{I}\sigma_e^2) = N(\mathbf{Wt}, \mathbf{I}\sigma_e^2)$ puesto que al estar **t** "dado" esto implica que es un valor fijado y constante,

$$f(\mathbf{y}|\mathbf{t}) = \text{cte.}\ \exp[\,(\mathbf{y} - \mathbf{Wt})'\ (\mathbf{y} - \mathbf{Wt}\,)]$$

La densidad a priori de **t** podemos suponer que es también Normal. Esto es razonable para los valores aditivos, pero para los efectos ambientales discutible, y luego lo discutiremos. La media y la varianza *a priori* de **t** son

$$E(\mathbf{t'}) = E\,[\mathbf{b'}\ \mathbf{u'}] = [\mathbf{m_b'}\ \mathbf{0}] = \mathbf{m}$$

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

donde **m** y **S** son la media y la varianza *a priori* de los efectos ambientales. Suponemos que los efectos ambientales expresados en **b** y los aditivos de **u** están incorrelacionados[31].

$$f(\mathbf{t}) = \text{cte.}\ \exp[(\mathbf{t} - \mathbf{m})'\ \mathbf{V^{-1}}\ (\mathbf{t} - \mathbf{m})]$$

$$f(\mathbf{t}|\mathbf{y}) = \text{cte.}\ f(\mathbf{y}|\mathbf{t})\ f(\mathbf{t}) = \text{cte.}\ \exp\,[(\mathbf{y} - \mathbf{Wt})'\ (\mathbf{y} - \mathbf{Wt}) + (\mathbf{t} - \mathbf{m})'\ \mathbf{V^{-1}}\ (\mathbf{t} - \mathbf{m})]$$

Hemos visto tres estimadores bayesianos: media, mediana y moda de $f(\mathbf{t}|\mathbf{y})$. Como la distribución es Normal, coinciden. Hallaremos, por ejemplo, la moda, que es el valor que maximiza $f(\mathbf{t}|\mathbf{y})$. Derivando respecto a **t** e igualando a cero se obtiene

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

que, en forma expandida, es

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

Estas ecuaciones son muy parecidas a las del modelo mixto, pero al producto matricial de los efectos antes considerados como "fijos", **X'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 **A**. 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 densidad *a priori* de los efectos aditivos es normal de media cero y varianza **A**, y la de los efectos ambientales es plana (todos los posibles valores tienen la misma probabilidad *a priori*[32] y varía a lo largo de toda la recta real (pueden tomar cualquier valor). Este último punto es discutible, puesto que cualquier efecto imaginable tiene en ganadería si no una distribución de densidad de pro-

31. Esto no es así en ciertas ocasiones. Se sabe, por ejemplo, que las mejores granjas de vacuno de leche gozan de un ambiente mejor y además usan semen de mejor calidad. Deshacerse de estas correlaciones fue una de las razones que impulsaron a HENDERSON (1973) a proponer el efecto granja como "fijo". Cuando el tamaño de las granjas es pequeño estos efectos están peor estimados que los aleatorios, lo que repercute en la estima de los efectos genéticos y genera un problema de difícil solución.

32. Para ser estricto debiera decir que todos los intervalos de valores de la misma longitud tienen la misma probabilidad, puesto que la probabilidad de un punto concreto es cero ya que hay infinitos puntos en cualquier intervalo de la recta real.

babilidades asociada, al menos unos límites que claramente no puede sobrepasar. Su uso queda justificado sólo como el de una densidad de referencia y por motivos operativos a la hora de calcular densidades posteriores.

Interpretación bayesiana de los métodos máximo verosímiles

Suponiendo que cualquier valor de h^2 es igualmente probable, $f(h^2)$=cte. (lo que sólo es en realidad admisible en el intervalo [0 1])

$$f(h^2|\mathbf{y}) = cte \cdot f(\mathbf{y}|h^2) \qquad [1]$$

con lo que se puede obtener $f(h^2|\mathbf{y})$ simplemente obteniendo $f(\mathbf{y}|h^2)$. En principio tampoco tenemos por qué saber cómo se distribuye $f(\mathbf{y}|h^2)$, pero esto es más fácil de resolver porque $f(\mathbf{y}|h^2)$ es una función de las muestras **y** que se toman, lo que facilita en muchas ocasiones el problema. Por ejemplo, en determinadas circunstancias, basándonos en experimentos pasados, podemos decir que y sigue una distribución Normal. Obsérvese, y es importante observarlo, que aunque la forma de la verosimilitud coincida con $f(h^2|\mathbf{y})$, la interpretación de la verosimilitud y de la densidad posterior es completamente diferente, puesto que $f(h^2|\mathbf{y})$ es una función de densidad de probabilidad y podemos hacer inferencias probabilísticas como hemos indicado antes.

En el caso particular de las estimas REML, fue HARVILLE (1974) quien primero se percató de que admitían una interpretación bayesiana. La moda de la densidad posterior $f(\mathbf{b},\sigma_u^2,\sigma_e^2|y)$ representa el máximo de esa distribución trivariante y tiene tres componentes, uno para **b** y otros dos para σ_u^2 y σ_e^2. La densidad posterior $f(\sigma_u^2|y) = \int\int f(\mathbf{b},\sigma_u^2,\sigma_e^2|y)\, d\mathbf{b}\, d\,\sigma_e^2$ tiene un máximo para un valor de que no tiene por qué coincidir con el componente de la moda de $f(\mathbf{b}, \sigma_u^2, \sigma_e^2|y)$. La estima máximo verosímil (ML) de σ_u^2 coincide con la componente correspondiente de la moda de $f(\mathbf{b}, \sigma_u^2, \sigma_e^2|y)$, mientras que la estima REML coincide con la moda de $f(\sigma_u^2, \sigma_e^2|y)$. Cuando hay varios efectos aleatorios, como por ejemplo en el modelo animal con repetibilidad:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e}$$

donde **p** son los efectos aleatorios de ambiente común entre dos mediciones del carácter, la estima REML de σ_u^2 coincide con la componente correspondiente de la moda de la distribución $f(\sigma_u^2, \sigma_p^2, \sigma_e^2|y)$ mientras que la estima ML coincide con la componente correspondiente de la moda de la distribución $f(\mathbf{b}, \sigma_u^2, \sigma_p^2, \sigma_e^2|y)$.

Obsérvese que el REML, cuya justificación desde el punto de vista frecuentista es oscura, aparece aquí justificado de forma bayesiana (con las limitaciones de considerar la densidad previa de los efectos "fijos" como plana).

Obsérvese también que desde un punto de vista bayesiano puede irse más lejos y, en el caso del modelo con varios efectos aleatorios, proponer la moda de la densidad

$$f(\sigma_u^2|y) = \int\int f(\sigma_u^2, \sigma_p^2, \sigma_e^2|y)\, d\,\sigma_p^2\, d\,\sigma_e^2$$

como estimador de la varianza σ_u^2. Esto es exactamente lo que proponen GIANOLA y FOULLEY (1990) con su estimador VEIL. La razón para proponer la moda, aparte de argumentos en torno a la función de pérdidas o a que sea el valor más probable, es de tipo práctico: se obtiene maximizando una función. Actualmente el desarrollo de técnicas con cadenas de Markov por simulación de Monte-Carlo (muestreo de Gibbs, etc.) hace innecesaria la obligatoriedad de recurrir a la moda, puesto que se estima la totalidad de la densidad posterior.

Los estimadores bayesianos del valor genético y los parámetros genéticos

Al margen de la interpretación bayesiana que tengan ciertos métodos de mejora, ¿cómo aborda un bayesiano la estima de los valores genéticos? En primer lugar, no existen "efectos fijos" en un contexto bayesiano, por lo que no sería de aplicación el modelo mixto expuesto en [1], sin embargo mantendremos la nomenclatura con el objeto de separar los efectos genéticos de los ambientales. La forma más amplia de abordar el problema consistiría en estimar la densidad posterior conjunta $f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y})$, y sus correspondientes densidades marginales, aunque pronto aparecen las mismas dificultades que en el caso de los estimadores frecuentistas[33].

En primer lugar, se desconoce la densidad a priori conjunta de **b** y **u**, por lo que se admite que son independientes: $f(\mathbf{b}, \mathbf{u}) = f(\mathbf{b})f(\mathbf{u})$. Como esto no es así en algunos casos importantes (ver nota 31), es frecuente tomar como valor a priori $f(\mathbf{b}) =$ constante, con lo que podemos obtener la solución BLUP con sus ventajas e inconvenientes[34]. En el caso de **u** se admite que se distribuye de forma Normal con media cero y varianza σ_u^2. Una forma alternativa de proceder sería dar información previa a $f(\mathbf{b})$, con lo que la solución final puede verse afectada si hay pocos datos. Ya hemos discutido antes el uso de la información a priori, y habitualmente se espera que sea irrelevante debido a la información que proporcionen los datos, lo que es de esperar que ocurra frecuentemente, dada la gran cantidad de datos disponibles en cualquier programa de mejora.

En segundo lugar, hay que estimar las componentes de varianza a partir de los datos. Esto también afecta a la forma de la densidad a priori, puesto que también se suele suponer la independencia de las distribuciones y esto no es necesariamente cierto. Se suele suponer que la densidad a priori de las componentes de varianza es una χ^2 invertida para calcular más fácilmente la función de densidad posterior. Esto, que parece enormemente arbitrario, no lo es tanto, puesto que estas funciones χ^2 pueden tener formas muy diversas en función de dos parámetros[35], y puede expresarse un amplio margen de opiniones previas en función de sus valores. En cualquier caso, desde la aparición de las técnicas de muestreo de Gibbs y de muestreo en "caja negra", ya no es estrictamente necesario usar funciones que acomoden particularmente bien a los cálculos, aunque se sigue haciendo para facilitar el cómputo de las densidades posteriores.

Una vez determinadas las densidades previas y la verosimilitud, se calcula la densidad posterior multiplicando ambas densidades. Si la función de densidad resultante es conocida, por ejemplo una Normal, se pueden aprovechar sus propiedades para producir inferencias (por ejemplo, que el

33. En realidad la forma más amplia incluiría las distribuciones de posibles matrices **Z** y **X**. Como no se sabe bien cómo abordar este problema, en realidad lo que se estima es la densidad posterior de $f(\mathbf{b},\mathbf{u}, \sigma_u^2, \sigma_e^2 | \mathbf{y}, \mathbf{X}, \mathbf{Z})$, aunque la forma habitual de expresarla sea la que indicamos en el texto. Se pueden incluir más efectos y más jerarquías, pero para el ejemplo es más que suficiente el modelo propuesto.

34. Nótese que aunque a priori sean independientes, en la densidad posterior típicamente no lo serán.

35. Parámetros a los que se les llama "grado de credibilidad" n y "parámetro de posición" S de forma un tanto equívoca, puesto que ambos parámetros afectan a la forma de la curva.

95% de probabilidad cae entre 1,96 y -1,96 veces su desviación típica). Cuando no es así se puede intentar realizar una estimación puntual, por ejemplo hallando la moda de la función –con lo que basta hallar el máximo–, o bien se pueden extraer valores de esa densidad posterior y trabajar con esos valores para representar aproximadamente esta función de densidad. 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 $f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e|\mathbf{y})$. Cada uno de ellos tiene un valor para σ^2_u y σ^2_e, por lo que realizando el cociente se obtienen 5.000 puntos muestreados al azar de la densidad posterior 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, así como cualquier intervalo de confianza.

El problema se centra ahora en cómo obtener esos valores tomados al azar de las funciones de densidad posterior. Hay un conjunto de funciones de los que se sabe cómo extraer valores al azar, y si la densidad posterior es una de ellas, no hay problemas para hacerlo. Cuando no es así, existen técnicas de muestreo "en caja negra" que permiten, con más o menos coste computacional, realizar ese muestreo[36]. En el caso multivariante, o que involucra estima de muchos parámetros, como es el caso habitual en problemas de mejora genética (piénsese que además de las estimas de componentes de varianza hay que estimar los valores genéticos), no es posible computacionalmente –al menos de momento– muestrear directamente de la función de densidad posterior, y hay que transformar el problema en univariante, o en problema de menos dimensiones, mediante técnicas de muestreo de Gibbs que se explican en el apéndice I. Las técnicas de muestreo de Gibbs permiten, además muestrear de la función de densidad posterior y no de funciones proporcionales, eludiendo el problema expuesto en la nota 27.

Discusión

Al comparar las ventajas e inconvenientes de las dos escuelas es menester distinguir los de tipo ontológico (relativas a la inferencia científica) de las de tipo práctico. Aunque las primeras han dado lugar a una copiosa y entretenida bibliografía, el mejorador genético animal estará a menudo más interesado en las segundas.

Ventajas e inconvenientes respecto a la inferencia científica

Hasta hace algunos años por lo menos, se ha venido considerando que la diferencia esencial entre ambas escuelas residía en el

36. No podemos detallarlas aquí. Esencialmente consisten en muestrar de una función conocida que "cubre" a la desconocida, y decidir por métodos simples si el punto muestreado pertenece a una o a otra función. Para poder "cubrir" a la función desconocida es necesario conocer su máximo, lo que en el caso multivariante representa un problema y hay que acudir a otras técnicas como el muestreo de Metrópolis. TANNER (1993) trata con detalle las recetas de esta "cocina bayesiana".

uso que hacen de la información previa al experimento. La escuela frecuentista produce inferencias a partir de los datos y de su conocimiento previo de la distribución del espacio muestral. Aunque se han hecho muchas bromas respecto a la validez de realizar inferencias a partir de algo más que los datos[37], el problema del frecuentismo reside en que para poder hablar de probabilidades sin usar información previa, como querían sus fundadores (ver PEARSON, 1962), no hay más remedio que referirse a la muestra y no al parámetro que se desea estimar, lo que introduce una forma un tanto antinatural de expresar la incertidumbre asociada al resultado de un experimento. Fisher era perfectamente consciente de la pérdida que ello suponía:

"Bayes percibió la importancia fundamental de este problema y propuso un axioma que, de ser cierto, bastaría para traer al campo de la probabilidad un gran número de inferencias inductivas; de forma que después de observar una muestra se podrían hacer inferencias acerca de la población, inferencias inciertas, por supuesto, pero con la bien conocida clase de incertidumbre que proviene de las leyes de la probabilidad" (FISHER, 1936).

La escuela bayesiana realiza inferencias a partir de probabilidades asociadas a valores del parámetro sobre el que se desea establecer el grado de conocimiento existente, lo que representa un medida mucho más natural de la incertidumbre. La inquietud al usar este procedimiento surge de que todo pueda ser una ilusión; es decir, de que en realidad no sean probabilidades las densidades posteriores debido a que no se ha podido precisar el grado de información previo o a que ese grado es inexistente o vago.

Las diferencias de fondo residen en que las dos escuelas dan respuestas diferentes al problema de la inducción[38]. En términos esquemáticos, este principio niega la posibilidad de la inferencia; por ejemplo, niega la posibilidad no ya de estimar con certeza parámetros poblacionales a partir de muestras, sino de estimarlos con una probabilidad determinada (POPPER y MILLERS, 1983). La solución de HUME (1739) a este problema era de tipo psicológico: cuando observo una repetición de acontecimientos (por ejemplo, que las vacas Jersey producen menos que las Frisonas) espero que en futuro el comportamiento del fenómeno será el mismo. Obviamente esta solución no es satisfactoria y desde los primeros

37. BERGER y WOLPERT (1984) recogen una buena colección de estas bromas. Sin embargo no deja de ser razonable hacer inferencias a partir no sólo de los datos sino del conocimiento previo de cómo se distribuyen.

38. El problema de la inducción fue presentado en su forma moderna por Hume aunque como señala COPLESTON (1959, Vol. V, pp. 286) la esencia de ese problema –la dificultad de asociar efectos a causas– había sido ya planteada mucho antes. Kant buscó la justificación racional en el conocimiento innato (a priori). El *a priori* de Kant sólo tiene en común con el de Bayes el nombre. Kant dice, "Es por tanto una de las primeras y más necesarias cuestiones, y que no puede resolverse a simple vista, la de saber si hay algún conocimiento independiente de la experiencia y también de toda impresión sensible. Llámese a este conocimiento a priori, y distíngase del empírico en que las fuentes del último son a posteriori, es decir, que las tiene en la experiencia". KANT (1787, pp. 98). Naturalmente ningún científico utiliza la información a priori como algo ajeno a la experiencia. Pese a ello, Kant es aún citado por algún matemático bayesiano (ROBERT, 1992, pp. 336) que no parece haberlo leído bien. La exposición más detallada que conozco del problema de la inducción se encuentra en STOVE (1995, pp. 101-143).

intentos de KANT (1787) hasta los más recientes de RUSSELL (1948) y POPPER (1972) se ha intentado dar una justificación racional de por qué a pesar de todo creemos que la repetición de un fenómeno tiene una causa y podemos estimarla, aunque las soluciones han sido poco satisfactorias. RUSSELL (1948, pp. 491) propone cinco postulados que, de admitirse (y esto es voluntario, por lo que el problema de la inducción sigue sin resolverse) justificarían la inferencia, y POPPER (1972) se limita a demarcar mejor el problema de la inducción sin aportar nada nuevo[39]. En la actualidad se admite que el problema de la inducción no tiene una solución del tipo lógico-deductivo, y la investigación actual se centra en el examen de las "buenas razones" que hay para creer en este principio, y de la naturaleza de estas "buenas razones"[40].

Entre los estadísticos frecuentistas, Neyman prefería hablar de "comportamiento" de un científico ante los resultados del experimento antes que de inferencia científica. Fisher defendía, sin embargo, que era posible asociar una probabilidad a un suceso, y en una larga discusión con el estadístico bayesiano Jeffreys (recogida en LANE, 1980) este problema marcó las principales diferencias entre los dos planteamientos. En realidad la estadística frecuentista tampoco resuelve el problema de la inducción, por un lado porque no hay justificaciones claras para afirmar que se conoce con precisión la distribución del espacio muestral, y por otro porque cada medida de incertidumbre acaba por ir asociada a un parámetro poblacional que se desconoce (por ejemplo, el error estándar de la media depende de la varianza poblacional).

Una forma clara de comprender el problema de la inducción es examinar el teorema de Bayes. Para poder asociar una probabilidad a un acontecimiento debo saber su densidad de probabilidades *a priori*. Si tomo una muestra de vacas frisonas y observo que producen 400 litros de leche al año por término medio, debo saber que ese valor es muy improbable *a priori* para poder establecer una conclusión razonable. Sin embargo esta probabilidad a priori nos es desconocida, por lo que estoy basándome en mi grado de creencia para establecer la inferencia final. Por supuesto, en el caso de las vacas, y si soy una persona relacionada con la ganadería, tengo un convencimiento psicológicamente fuerte de que las vacas frisonas producen más de 400 litros de leche al año porque creo que son correctas mis lecturas a través de los años, los comentarios de los colegas en los Congresos, etc., pero estoy hablando de un estado de creencias, que será mucho más vago si el campo en el que estoy experimentando me es poco conocido.

Obsérvese que el uso del estado de creencias es perfectamente razonable, y no implica que no puedan ser compartidas por otros colegas de mi profesión. Un asunto diferente, y que coloca al paradigma bayesiano en una situación incómoda desde el punto de vista ontológico, es la imposibilidad psicológica que se puede dar cuando

39. Popper sólo es considerado como un gran filósofo en ambientes científicos, probablemente por ser un filósofo fácil de leer y por ocuparse de temas relacionados con la ciencia. Stove (1995) realiza una crítica demoledora de su filosofía.

40. El conocido filósofo americano Max Black ha hecho una revisión del estado actual de la investigación en torno al principio de inducción, y también de su relación con la probabilidad (BLACK, 1984).

hay que establecer un grado de creencias multivariante, como es el caso habitual de la mejora genética animal. Debido a la dificultad de precisar bien el estado de creencias a priori, o el de ignorancia, el bayesianismo moderno tiende a considerar la función de densidad a priori como un artificio matemático, como hemos dicho antes[41]. Por otro lado, se ha producido una cierta aproximación al modo de trabajar bayesiano por parte de la estadística frecuentista mediante el uso de técnicas de boot-strap, que permiten tener una representación de la distribución muestral; aunque el problema de fondo sigue siendo el mismo, puesto que sigue sin poderse establecer la incertidumbre respecto al parámetro desconocido en términos de probabilidad. Así las cosas, la discusión sobre inferencia promete continuar produciendo una cantidad considerable de literatura, no necesariamente útil a la hora de ir a seleccionar conejos a las ocho de la mañana de cada martes.

Ventajas e inconvenientes respecto a la aplicación

Hasta hace muy pocos años, la crítica fundamental que desde un punto de vista práctico se hacía a los estadísticos bayesianos era su esterilidad a la hora de producir resultados útiles. La metodología implicaba la resolución de integrales complicadas o el uso de aproximaciones más o menos afortunadas. La aparición de técnicas de Monte-Carlo en el muestreo de cadenas de Markov, de los que el popular "Gibbs sampling" es sólo un ejemplo[42], ha resuelto la mayor parte de problemas operativos, y aunque ha generado otros nuevos, particularmente cuando los modelos no son Normales, la entidad de estos problemas operativos es ahora mucho menor. Las ventajas de disponer con facilidad de muestras realizadas de la densidad posterior vuelve a la estadística bayesiana particularmente atractiva, puesto que no es menester conocer la forma de la función para realizar inferencias: la media se estima directamente de los individuos muestreados, así como los intervalos de confianza u otros estimadores puntuales. El principal inconveniente se centra en que no se puede tener, al menos por el momento, la certeza matemática de que las cadenas de Markov utilizadas tengan las propiedades requeridas para garantizar que los puntos muestreados pertenecen, efectivamente, a la densidad posterior. Este tema se discute con más detalle en el Apéndice.

Las dificultades de las técnicas frecuentistas en mejora animal son de dos tipos: por un lado las derivadas del uso de grandes bases de datos y por otro las que derivan de la dificultad del frecuentismo para precisar el error estándar de un parámetro en situaciones medianamente complejas. Un ejemplo de lo primero es la dificultad de obtener estimas REML multivariantes

41. Aunque esto es relativamente moderno. En un trabajo clásico, LINDLEY y SMITH (1972) arguyen "que es típicamente cierto que hay información a priori disponible acerca de los parámetros, y que puede ser explotada para encontrar mejores y a veces sustancialmente mejores estimas [respecto a las de mínimos cuadrados]". Compárese esta afirmación con la de BERNARDO y SMITH, 1994, pp. 298 que figura al final del apartado 3.2.

42. Tan popular, que hay quien lo confunde con un método de estimación. JOSIAH WILLARD GIBBS (1839-1903), célebre físico americano fundador de la termodinámica estadística, tiene poco que ver con estas técnicas. El nombre proviene de que estas cadenas de Markov generadas por muestreo aleatorio se aplicaron por primera vez a distribuciones inventadas por Gibbs, aunque más de 80 años después de la muerte de éste.

cuando se dispone de algunas decenas de miles de datos, y un ejemplo de lo segundo es la dificultad de estimar con precisión el error de la respuesta a la selección cuando ésta se calcula mediante medias genéticas estimadas por BLUP usando parámetros genéticos estimados mediante REML. Sin embargo hay casos en los que los métodos frecuentistas son rápidos y precisos (por ejemplo en el uso habitual de los índices de selección o el BLUP para estimar los valores aditivos de los individuos con objeto de realizar selección truncada) y no es fácil que técnicas más sofisticadas los desplacen[43]. El abordaje bayesiano presenta, sin embargo ciertas ventajas en algunos casos:

1 *En el abordaje de nuevos problemas:* La forma de trabajar es clara y siempre la misma, obtener densidades posteriores de las variables a estimar. En esto se parece en cierto modo al método de máxima verosimilitud.

2. *Cuando se dispone de pocos datos:* Aunque la información usada en programas de mejora animal es habitualmente masiva, en ciertos experimentos la cantidad de datos está limitada por el coste o las facilidades disponibles. Incluso si el total de datos es grande pero la cantidad de datos por "efecto fijo" es pequeña –como es frecuente con el uso de animales en bandas, o en datos de campo tomados en granjas pequeñas– las estimas de los efectos fijos no son buenas y esto repercute en las estimas de los efectos aleatorios. Los métodos máximo verosímiles tienen propiedades asintóticas y podrían no ser útiles en estos casos, por lo que las densidades posteriores bayesianas sistematizan mejor el grado de ignorancia. Por otra parte, encontrar un máximo multivariante cuando hay pocos datos es poco seguro, la función de verosimilitud es muy plana y hay riesgos de caer en máximos locales; sin embargo las densidades bayesianas son habitualmente densidades marginales, por lo que su representación ilustra mucho mejor el grado de imprecisión del resultado que el dato del máximo y el error típico asociado (los perfiles de verosimilitud no son fiables con pocos datos, debido a las propiedades asintóticas de la verosimilitud que acabamos de comentar). Además, si no se desea confiar mucho en la información producida por los datos para las componentes de varianza, puede usarse información subjetiva (un ejemplo de esta forma de proceder se encuentra en BLASCO *et al.*, 1997); de una forma similar, los usuarios frecuentistas de BLUP usan a menudo componentes de varianza basadas en la bibliografía, sólo que el estadístico bayesiano puede precisar mejor el grado de ignorancia a priori.

3. *Cuando se desconoce la distribución en el espacio muestral del parámetro que se desea estimar:* Este es un caso frecuente, particularmente cuando el carácter analizado no tiene una distribución Normal. Por ejemplo, si se comparan dos curvas de Gompertz –lo que es habitual en estudios de crecimiento con mamíferos– no hay fórmulas que precisen el error típico de las diferencias entre parámetros de las curvas. La ventaja de la aproximación bayesiana,

43. Aunque la afición de los mejoradores animales al cómputo es tan notable, que ya aparecen BLUPs calculados mediante cadenas de muestreo de Gibbs (ver, por ejemplo, las últimas Jornadas de Producción Animal de Zaragoza (1997). Esto haría válida la ironía de uno de los autores del REML, Robin Thompson, cuando en el 6° Congreso mundial de genética aplicada a la producción ganadera (GUELPH, 1994) comentaba: "¿Qué es el Gibbs sampling? ¡El BLUP con ruido!

unida a las nuevas técnicas de muestreo de Monte-Carlo en cadenas de Markov, es que se pueden obtener muestras de una densidad posterior, y establecer inferencias bayesianas sin necesidad de realizar las operaciones matemáticas, frecuentemente irresolubles incluso por métodos numéricos, que implica la utilización de la fórmula matemática de la densidad posterior para hallar medias, modas o intervalos de confianza. Una solución de tipo general aplicable a cualquier curva ha sido propuesta por *Varona et al.* (1997), y dos aplicaciones, aún sin publicar en el momento de escribir estas líneas, han sido propuestas por REKAYA *et al.* (1998) y BLASCO y VARONA (1998) para curvas de lactación y de crecimiento respectivamente. Hay ocasiones en que el problema planteado tiene una solución bayesiana trivial, por ejemplo cuando el parámetro a estimar es combinación de otros parámetros estimados (por ejemplo un coeficiente de correlación). En estos casos, como se dispone de valores muestreados de la densidad posterior conjunta, basta con tomar el valor de la covarianza y de las desviaciones típicas de cada individuo muestreado y calcular un coeficiente de correlación. Todos ellos serán una muestra de la densidad posterior conjunta del coeficiente de correlación, con lo que se puede ya hacer inferencias sobre este coeficiente y calcular su precisión.

4. *En problemas de estimación jerarquizados:* Cuando se necesita un parámetro para realizar una estimación de otro parámetro es complicado o imposible hallar un error estándar que incluya ambos procesos de estimación. Por ejemplo, cuando se desea estimar el valor genético de los animales mediante un BLUP es necesario estimar las componentes de varianza de los efectos aleatorios. Es frecuente estimarlas mediante REML a partir de los mismos datos, pero cuando se calculan los errores estándar de los valores genéticos no se tiene en cuenta el error cometido al estimar las componentes de varianza, sino que se toman éstas como si fueran valores verdaderos (o estimados sin error). Una forma de abordar el problema desde la perspectiva bayesiana consiste, como hemos visto, en calcular la densidad posterior conjunta $f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e | \mathbf{y})$ e integrar respecto a las componentes de varianza y al vector **b** para obtener la densidad posterior $f(\mathbf{u}|\mathbf{y})$. Esto equivale a dar a las componentes integradas todos los valores posibles y ponderarlos por sus probabilidades respectivas. SORENSEN *et al.* (1994) han mostrado cómo hacerlo de forma sencilla utilizando técnicas de muestreo de Gibbs y ha sido aplicado primero por WANG *et al.* (1994) en porcino y posteriormente por un número creciente de autores.

5. *Cuando se imponen condiciones a los datos o a los parámetros a estimar:* Por ejemplo, en el caso de la durabilidad de las vacas en la explotación, se dispone de individuos con distinto número de lactaciones pero se desea estimar el número máximo posible. O por ejemplo en el caso de establecer el orden correcto de los sementales de un catálogo, y en general en cualquier problema de estadísticos ordenados. En ambos casos la solución bayesiana es trivial, en el primero usando técnicas de datos faltantes[44] y en el segundo la solución es

44. No puedo aquí entrar en el detalle de la técnica. Consiste esencialmente en hallar la densidad posterior suponiendo que se conocen los datos faltantes e integrarla luego poniendo los límites que describen la situación real.

inmediata a partir de las densidades posteriores obtenidas.

En cualquier caso, el mejorador que no quiera entrar en problemas de la filosofía de la inferencia, sino más bien usar útiles establecidos para resolver sus problemas, debe saber que las escuelas mencionadas están perfectamente establecidas y que no tiene sentido el tener que justificar el uso de un tipo de inferencia u otro al presentar sus resultados. Ninguna de las escuelas presenta hoy dificultades operativas salvo para casos especialmente complejos, y hay software disponible para analizar una gran variedad de problemas genéticos desde el punto de vista de ambas. Escoger una u otra es un acto que para un mejorador aplicado depende esencialmente de qué problemas resuelve una que la otra no haya resuelto.

Ambas escuelas presentan problemas formales y paradojas. Suele ocurrir que se admiten con más facilidad las paradojas a las que el investigador se encuentra acostumbrado. También –por qué no decirlo– las revistas científicas encuentran "socialmente admisibles" los problemas mal resueltos de la escuela frecuentista (por ejemplo, aplicar un método máximo verosímil con pocos datos) antes que los de la bayesiana, aunque esto está cambiando ya. Cualquier refinamiento estadístico no debe hacer olvidar al investigador que el problema esencial de su trabajo reside en disponer de suficiente cantidad de información experimental para que sus conclusiones puedan expresarse de forma lo bastante precisa. La angustia del técnico o del científico que desearía expresar con certeza sus resultados, y no sólo con un grado de probabilidad, no es nueva; ya Virgilio en su Egloga IX, y en un contexto agrícola, hace exclamar a su protagonista: "¡Feliz aquél que puede conocer las causas de las cosas!".

Agradecimientos

Este artículo fue revisado por un número considerable de colegas antes de su envío para publicación. Deseo expresar mi agradecimiento a todos ellos, y especialmente a mis amigos Daniel Gianola, Daniel Sorensen y Susie Bayarri por sus detallados comentarios, y a José Vicente Bonet por su ayuda en la parte filosófica.

Apéndice I. Muestreo de Gibbs

Se pretenden muestrear valores al azar de la función de densidad posterior

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

Para ello se procede de la siguiente forma:

Se toman valores arbitrarios de inicio para $\mathbf{u}$, σ_u^2, σ_e^2, y se muestrea un valor al azar de la función condicional

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

Con el valor de $\mathbf{b}$ así muestreado, y con los valores anteriores de σ_u^2, σ_e^2, se muestrea un valor al azar de la función condicional

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

Con el valor de $\mathbf{u}$ así muestreado, con el de $\mathbf{b}$ muestreado antes y con el de σ_e^2 inicial, se muestrea un valor al azar de la función condicional

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

Con el valor de σ_u^2 así muestreado y con los de **b** y **u** muestreados antes, se muestrea un valor al azar de la función condicional

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

con los valores muestreados **u**, σ_u^2, σ_e^2, se reinicia el proceso. Finalmente, al cabo de un cierto número de iteraciones, los valores que se muestrean pertenecen a la densidad conjunta posterior $f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2|y)$. Cada valor es, además, miembro de cada una de las distribuciones marginales $f(\mathbf{b}|y)$, $f(\mathbf{u}|y)$, $f(\sigma_u^2|\ y)$, $f(\sigma_e^2|\mathbf{y})$, que permiten hacer inferencias por separado en cada parámetro.

Lo mismo se haría para realizar inferencias sobre combinaciones lineales de efectos "fijos" y aleatorios. Si no se sabe cómo extraer muestras de una función directamente, puesto que la función condicional puede tener una forma complicada, se utilizan muestreos en "caja negra", por lo que casi siempre es posible realizar un proceso de muestreo de Gibbs. Los problemas asociados a este procedimiento son tres:

1) *No se puede demostrar que se está muestreando de la función de densidad posterior.* Para que una cadena de Markov converja a una densidad posterior debe cumplir ciertas propiedades y ser una cadena *irreducible.* Aunque se puede demostrar que ciertas cadenas no son irreducibles, no hay una demostración general que asegure que la cadena que estamos usando tiene esta propiedad (los detalles pueden consultarse en los apuntes del curso que impartió Daniel Sorensen en la Escuela de Agrónomos de Valencia; SORENSEN, 1997).

2) *No se puede saber con certeza a partir de qué iteración los puntos extraídos pertenecen a la función de densidad posterior.* Es decir, aunque supiéramos que nuestra cadena es irreducible, no sabríamos a partir de qué iteración hemos llegado a la convergencia.

3) *No se puede asegurar que la cadena de Markov haya convergido a una distribución estacionaria.* A veces se encuentran zonas de tránsito que dan la impresión de que pertenecen ya a la distribución de convergencia, pero continuando el muestreo un número suficiente de veces la cadena evoluciona hacia otra distribución.

Estas dificultades se podrían resumir diciendo que no sabemos cuándo ni a qué hemos convergido, en el caso de que hayamos convergido a algo. Como se ve, el problema no es trivial y es en la actualidad objeto de una investigación intensiva por matemáticos especializados en cadenas de Markov. Los científicos aplicados se conforman con iniciar varias cadenas desde puntos de partida diferentes y comprobar que aparentemente se converge a la misma distribución. Este convencimiento es de tipo psicológico, pero suficiente para la mayor parte de usuarios. No es diferente del convencimiento psicológico que tiene quien aplica el método de máxima verosimilitud a un problema multivariante y cree que ha obtenido un máximo global en lugar de uno local. O del convencimiento de quien estima una tendencia genética y cree que los parámetros genéticos que ha usado (y que van a determinar sus resultados) son próximos a los verdaderos. Ningún método operativo de cualquiera de las dos escuelas, frecuentista o bayesiana, está hoy por hoy libre de problemas, aunque hay esperanzas razonables de que en ambos casos se vayan encontrando soluciones aceptables para todos.

Bibliografía

ANDERSON R.L., BANCROFT, T.A., 1952. Statistical theory in Research. Mc-Graw Hill.

BLASCO A., SORENSEN D., BIDANEL J.P., 1997. A bayesian analysis of genetic parameters and selection response for litter size components in pigs. Genetics, en prensa.

BLASCO A., VARONA L. 1998., Ajuste y comparacion de curvas de crecimiento. ITEA. Remitido.

BARNETT V., 1973/1982. Comparative Statistical Inference. Second edition in 1982. Chichester, Wiley.

BERGER J.O., WOLPERT R.L. 1984., The likelihood principle. Institute of Mathematical Statistics. Lecture Notes - Monograph Series. Series Editor, Shanti S. Gupta, Purdue Universityp

BERNARDO J.M., 1979. Reference posterior distributions for bayesian inference. J. Roy. Statist. Soc. B 41, 113-147.

BERNARDO J.M., SMITH A.F.M., 1994. Bayesian theory. Wiley .

BLACK M., 1984. Inducción y probabilidad. Cátedra.

BOX G.E.P., TIAO G.C., 1973. Bayesian inference in statistical analysis. Reading, MA, Addison-Wesley.

CANTET R.J.C., FERNANDO R.L., GIANOLA D., 1992. Bayesian inference about dispersion parameters of univariate mixed models with maternal effects, theoretical considerations. Genet. Sel. Evol. 24, 107-135.

COPLESTON F., 1963/1985. A History of Philosophy. Vol V. Hobbes to Hume. Bantam Doubleday Dell Publ. Inc.

DEMPFLE L., 1977. Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayesiens. Ann. Genet. Sel. Anim. 9, 27-32.

EDWARDS A.W.F., 1992. Likelihood. Cambridge, University Press.

EFRON B., 1986. Why isn't everyone a bayesian? The American Statistician 40, 1-11.

FISHER R.A. 1912. On an absolute criterion for fitting frequency curves. Messeng. Math. 41, 155-160.

FISHER R.A. 1922. On the mathematical foundations of theoretical statistics. Phil. Trans., A. 222. 309-368.

FISHER R.A. 1935. The logic of inductive inference. J. Roy. Stat. Soc. 98, 39-54.

FISHER R.A., 1936. Uncertain inference. Proc. Am. Acad. Arts Sci. 71, 245-258.

FISHER R.A. 1956. Statistical methods and scientific inference. Hafner Press.

GIANOLA D., FOULLEY J.L., 1990. Variance estimation from integrated likelihoods. Genet. Sel. Evol. 22, 403-417.

GIANOLA D., RODRÍGUEZ-ZAS S., SHOOK G.E., 1994. The Gibbs sampler in the animal model, a primer. Séminaire Modele Animal. INRA Departament de Genetique Animale. Ed. J.L. Foulley y M. Molenat. pp. 47-56.

HARTLEY H.O., RAO J.N.K., 1967. Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54, 93-108.

HARVEY W.R., 1960. Least squares of data with unequal subclass numbers. U.S. Department of Agriculture, A.R.S. 20-28, 1-157.

HARVILLED., 1974. Bayesian inference for variance components using only error contrasts. Biometrika 61, 383-385.

HENDERSON C.R., 1949. Estimation of changes in herd environment. J. Dairy Sci. 32, 706

HENDERSO C.R., 1950. Estimation of genetic parameters. Ann Math. Stat. 21, 309-310.

HENDERSON C.R., 1953. Estimation of variance and covariance components. Biom. 9, 226-252.

HENDERSON C.R., 1963. Selection index and expected genetic advance. En. Statistical Genetics and Plant Breeding, W.D. Hanson y H.F. Robinson (Ed.). National Academy of Sciences-National Research Council, Washington, pp. 141-153.

HENDERSON C.R., 1973. Sire evaluation and genetic trends. En. Proceedings of the animal breeding and genetics symposium in honor of Dr. L. Lush. Blacksburg, Virginia. pp. 10-41.

HENDERSON C.R., 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biom. 32, 69-84.

HENDERSON C.R., 1984. Applications of linear models in animal breeding. University of Guelph

HUME D., 1739/1983. Tratado de la naturaleza humana. Orbis S.A.

JAMESW., Stein C., 1961. Estimation with quadratic loss. Proc. Fourth Berkeley Symp., 1 (J. Neyman and E.L. Scott, eds.) Berkeley. Univ. California Press, pp. 361-380.

JEFFREYS H., 1931. Scientific inference. Cambridge University Press.

JEFFREYS H., 1961. Theory of probability. Oxford University Press.

KANT E. 1787/1983. Crítica de la razón pura. Orbis.

KEMPTHORNE O., 1984. Revisiting the past and anticipating the future. En, Statistics, an appraisal. Proceedings 50th Anniversary Iowa State Statistical Laboratory (ed, H.A. David and H.T. David). The Iowa State University Press. pp. 31-52.

KENDALL M.G., 1961. Daniel Bernoulli on maximum likelihood. Biometrika 48,1-8.

KEYNES J.M., 1921. A treatise on probability. London. Macmillan.

KOTZ S., JOHNSON N.L., 1982. Encyclopedia of statistical sciences. Wiley .

LAMOTTE L.R., 1970. A class of estimators of variance components. Tech. Report 10. Univ. of Kentucky.

LANE D.A., 1980. Fisher, Jeffreys, and the nature of probability. En, Lecture Notes in Statistics (ed. S.E. Fienberg and D.V. Hinkley), Springer-Verlag Berlin Heidelberg., pp. 148-160.

LINDLEY D.V., SMITH A.F.M., 1972. Bayes estimates for the linear model. J. Royal Statist. Soc. Ser. B 34, 1-41.

LUSH J.L., 1947. Family merit and individual merit as bases for selection. Amer. Nat. 81, 241-261.

PATTERSON H.D., THOMPSON R., 1971. Recovery of interblock information when block sizes are unequal. Biometrika 58, 545-554.

PEARSON E.S., 1962 Some thoughts on statistical inference. 275-283.

PEARSON K., 1903. Mathematical contributions to the theory of evolution. XI on the influence of natural selection on the variability and correlation of organs. Phil. Trans. of the Royal Soc. of London A, 200, 1-66.

POPPER K.R., 1972. Conocimiento objetivo. Tecnos.

POPPER K., MILLER D., 1983. A proof of the impossibility of inductive probability. Nature 302, 687-688. PRESS S.J., 1989. Bayesian statistics. Wiley. RAO C.R., 1971. Minimum variance quadratic unbiased estimation of variance components. J. Mult. Anal., 1, 445-456.

RAO C.R., KLEFFE J., 1988. Estimation of variance components and applications. North-Holland.

REKAYA R., CARABAÑO M.J., TORO M.A., 1998. Bayesian analysis of lactation curves with a non-linear model in lactation curves. J. of Dairy Sci. Remitido.

ROBINSON G.K., 1991. That BLUP is a good thing, the estimation of random effects. Statistical Science 6, 15-51.

ROBERT C.P., 1992. L'Analyse statistique bayesienne. Economica.

RONNINGER K., 1971. Some properties of the selection index derived by "Henderson's Mixed Model Method". Z. Tierz. Züchtbiol. 83, 186-193.

RUSSELL B., 1948/1983. El conocimiento humano. Orbis S.A.

SCHAEFFER L.R., 1986. Pseudo expectation approach to variance component estimation. J. Dairy Sci. 69, 2884-2889.

SCHEEFFE H., 1967. The analysis of variance. Wiley .

SORENSEN D.A., WANG C.S., JENSEN J., GIANOLA D., 1994. Bayesian analysis of genetic change due to selection using Gibbs sampling. Genet. Sel. Evol. 26, 333-360.

SORENSEN D., 1997. Gibbs sampling in quantitative genetics. National Institute of Animal Sciences. Internal rapport No. 82. Tjele, Dinamarca, 188pp.

STIGLER S.M., 1986. The history of statistics. The measurement of uncertainty before 1900. Harvard University Press.

STIGLER S.M., 1983. Who discovered Bayes's theorem?. The American Statiscian 37, 290-296.

STOVE D.C., 1995. Popper y después. Cuatro irracionalistas contemporáneos. Tecnos.

STUART A., ORD J.K., 1991. Kendall's avanced theory of statistics Vol.2. Arnold

TANNER M.A., 1993. Tools for Statistical Inference. Springer-Verlag.

THOMPSON W.A., 1962. The problem of negative estimates of variance components. Ann. Math. Stat. 33, 273-289.

THOMPSON R., 1987. Statistical methods and applications to animal breeding. Thesis presented for degree of Doctor of Science. University of Edinburgh.

VON MISES R., 1928/1957. Probability statistics and truth. London. Macmillan.

VARONA L., MORENO C., GARCÍA-CORTES L.A. ALTARRIBA J., 1994. Multiple trait analysis of underlying biological variables of production functions. Livest. Prod. Sci. 47.201-209.

WANG C.S., GIANOLA D., SORENSEN D.A., JENSEN J., CHRISTENSEN A., RUTLEDGE. J.J., 1994. Response to selection for litter size in Danish Landrace pigs, a bayesian analysis. Theor. Appl. Genet. 88. 220-230.

WOOLLIAMS J.A., MEUWISSEN T.H.E., 1993. Decision rules and variance of response in breeding schemes. Anim. Prod. 56. 179-186.

YATES F., 1990. Foreword. En, R.A. Fisher, Statistical methods, experimental design and scientific inference. Oxford Univ. Press.

(Aceptado para publicación el 2 de enero de 1998)