The Bayesian controversy in animal breeding

A. Blasco[1,2]

Departamento de Ciencia Animal, Universidad Politécnica de Valencia, Valencia 46071, Spain

ABSTRACT: Frequentist and Bayesian approaches to scientific inference in animal breeding are discussed. Routine 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. The Bayesian approach is discussed in cases in which prior information is available, prior information is available under certain hypotheses, prior information is vague, and there is no prior information. Bayesian prediction of genetic values and genetic parameters are presented. Finally, the frequentist and Bayesian approaches are compared from a theoretical and a practical point of view. Some problems for which Bayesian methods can be particularly useful are discussed. Both Bayesian and frequentist schools of inference are established, and now neither of them has operational difficulties, with the exception of some complex cases. There is software available to analyze a large variety of problems from either point of view. The choice of one school or the other should be related to whether there are solutions in one school that the other does not offer, to how easily the problems are solved, and to how comfortable scientists feel with the way they convey their results.

Key Words: Animal Breeding, Statistical Analysis

 J. Anim. Sci. 2001. 79:2023–2046

Introduction

Gianola and Foulley (1982) introduced Bayesian methods in animal breeding in the context of threshold traits, and soon afterward Gianola and Fernando (1986) highlighted additional possibilities exploiting Bayesian techniques. Some years before, Ronningen (1971) and Dempfle (1977) had called the attention to the fact that BLUP could be interpreted as a Bayesian estimator, and Harville (1974) offered a Bayesian interpretation of REML. However, although Bayesian methods were theoretically powerful, they usually led to formulas in which multiple integrals had to be solved in order to obtain the marginal posterior distributions used for a complete Bayesian inference. Because these integrals could not be calculated, even using approximate methods, Bayesian inference was based on the mode of posterior distributions, often giving results rather similar to the REML approaches. When Monte-Carlo Markov Chain (**MCMC**) methods were applied to estimate marginal posterior distributions, the computation problems were solved and interest in Bayesian methods was renewed. However, frequentist statisticians are reluctant to use Bayesian methods, mainly due to difficulties found in the use of prior information inherent to the Bayesian paradigm. The objective of this article is to discuss the frequentist and Bayesian approaches to scientific inference in animal breeding, showing their advantages and disadvantages, and their application to particular topics in animal breeding for which the Bayesian approach can be useful. Stuart and Ord (1991) give a frequentist viewpoint of the statistics, whereas Bernardo and Smith (1994) provide a Bayesian account. Inferences based on the likelihood can be found in Edwards (1992). There is an excellent book on comparative statistical inference by Barnett (1999).

Scientific Inference

> A constant theme in the development of statistics has been the search for justification for what statisticians do. To read the textbooks one might get the distorted idea that 'Student' proposed his t-test because it was the Uniformly Most Powerful Unbiased test for a Normal mean, but it would be more accurate to say that the concept of UMPU gains much of its appeal because it produces the t-test, and everyone knows that the t-test is a good thing.
>
> Dawid (1976), as quoted by Robinson (1991)

[1]This paper has been revised by a considerable number of colleagues. I would like to express my gratitude to all of them, especially to my friends Daniel Gianola, Daniel Sorensen, and Susie Bayarri for their detailed comments and to Luis Varona for providing me an intuitive way of explaining Gibbs sampling. I also want to express my gratitude to one of the referees for his/her detailed revision and suggestions.

[2]Correspondence: P.O. Box 22012 (phone: +34963877433; fax: +34963877439; E-mail: ablasco@dca.upv.es).

Received February 7, 2000.

March 28, 2001.

The recent irruption of Bayesian methods in animal breeding has motivated a certain perplexity among animal breeders. The supporters of these methods say they give a precise answer to some problems that are not well solved in animal breeding; for example:

> Suppose some animal model holds, and a standard analysis coupling maximum likelihood (**ML**) with BLUP methodology is carried out. With respect to the first question [What can be said about genetic variance in a base population?], we would obtain a point estimate of genetic variance and a single measure of uncertainty, which, technically speaking is only meaningful in large samples and if the data are normally distributed. To tackle the second one [How does one account for uncertainty about breeding values when location and dispersion parameters are unknown?] we would have to place ourselves conditionally on the ML estimates of dispersion parameters, ignoring the error. The issue of response to selection [Given a selection scheme or experiment, how does one assess effectiveness of selection, and attach to this a measure of uncertainty?] is even more complicated. Estimation of responses can be derived from an animal model, but their properties are unknown. . . . An alternative is to adopt the Bayesian position. For each of the questions raised above, the Bayesian answer resides in arriving at the marginal posterior distribution of the unknown of interest. This distribution provides an exact account . . . of the uncertainty about the unknown parameter.
>
> Gianola et al. (1994)

Conversely, well-known statisticians working in animal breeding show their skepticism, if not their open opposition, to the use of these methods:

> I also pointed out why some Bayesian estimators have unfortunate properties, in that they are likely to give estimates of zero for between group variance components in well designed experiments. However this type of estimator is still advocated.
>
> Thompson (1987)

> The controversy between the frequentist school and the Bayesian school of inference has huge implications. From the viewpoint of the data analysis the implication is that all of the procedures exposited in books like Snedecor and Cochran's *Statistical Methods* are not merely poor but have no logical foundation. Also, every reporting of any investigation must lead to the investigator's making statements of the form: "My probability that a parameter, θ, of my model lies between, say, 2.3 and 5.7 is 0.90.
>
> Kempthorne (1984)

Although only the Bayesian and frequentist schools are discussed, because they are the only relevant schools of inference today, inferences based on likelihood are intermediate. On one hand, the likelihood plays a central role in the Bayesian inference as the function expressing all the information derived from the data. On the other hand, the method of ML has interesting frequentist properties. The animal breeder may be interested in finding efficient solutions for estimating breeding values, and not necessarily in the underlying philosophy. Thus, it is reassuring that when a data set is sufficiently large, the results are very similar in most cases. It is also pleasing that the methods used most frequently in animal breeding can be derived under either the Bayesian or the frequentist paradigms. Harville and Carriquirry (1992) studied the properties of closely related estimators for animal breeding values, deduced from frequentist and Bayesian approaches, and stressed the similarities between results of both procedures. Robinson (1991) even states that the differences between schools are minimal for estimation of breeding values, and that most of the debate comes from the insistence of supporters of both schools on stressing their differences, rather than their similarity. We may get the impression that the differences between schools are merely "philosophical" ones, with no practical consequences. It is true that when data sets are large, problems arise more from computation difficulties and limits than from inferential reasons. However, in most studies, experimental design tries to maximize efficiency by optimizing the size of the experiment, and a large data set may not be available in many cases. Moreover, differences between schools in their approach to inference are notable, even from a practical point of view.

First, the expression of uncertainty about unknowns is completely different. The frequentist school studies the distribution of the estimator. In this framework, it is usually given a standard error of this distribution and sometimes the complete sampling distribution of the estimators using techniques such as the bootstrap. The Bayesian school aims to obtain the probability density function of the parameter for a given set of data. From this density function one can get the most probable value of the parameter, or the probability that the parameter resides within certain limits. The frequentist way of inference is based on how a large number of estimates would be distributed around the true value if a large number of samples were taken or an infinite number of repetitions of the experiment were performed, whereas Bayesians examine the probability distribution of the true value, given the data. For a frequentist, the true value is usually fixed and the sample is variable, whereas for a Bayesian the sample is fixed and the parameter of interest is a random variable. That does not mean that a true value of the parameter does not exist; for a Bayesian the true value exists, but because the Bayesian does not know which value it is, and thus the Bayesian speaks about the probability that this parameter has a particular value.

Second, some statistical concepts currently used in animal breeding do not have a Bayesian interpretation, for example, "bias" and the difference between fixed and random effects. In a Bayesian context "bias" does not exist, because conceptual repetitions of the experiment are not considered. Also, all effects are random because the Bayesian way of expressing uncertainty is to draw density functions of all unknowns, and thus all

unknowns are considered random variables. This can be surprising to an animal breeder working with BLUP (Henderson, 1973) or REML (Patterson and Thompson, 1971), but the property of unbiasedness has been discussed even within the frequentist school. For example, Henderson himself stated that biased methods could be more efficient (e.g., Henderson 1984). Fisher considered the property of unbiasedness as largely irrelevant (Fisher, 1956), mainly because it was not invariant to reparametrization. For example, the expectation of an unbiased estimator of the variance is the population variance, but the expectation of the square root of this estimator is not an unbiased estimator of the standard deviation. With respect to fixed effects, Fisher thinks that the distinction between fixed and random effects was not a clear improvement of the analysis of variance (see the introduction to the new edition of Fisher's statistical books, Yates, 1990). Henderson decided to use fixed effects in genetic evaluation, introducing mixed models, due to some technical problems related to the correlations between herd and genetic effects, as will be seen later, but noting that random effects have a lower risk of estimation.

Third, inferences obtained from both schools are not always coincident, particularly for small samples and when the Bayesian analysis uses prior information (e.g., Wang et al., 1994).

Fourth, some problems that have no solution (or only a rough approximation) in the frequentist school can be solved unambiguously with the Bayesian approach (see the previous quote of Gianola et al., 1994).

Given the apparent advantages of the Bayesian school on the first and fourth points, two obvious questions arise. Why were the Bayesian techniques abandoned in the past? Or, as Brad Efron asks in a brief paper with an interesting discussion, Why isn't everyone a Bayesian? (Efron, 1986). And why have frequentist techniques persisted in the field of animal breeding? The first question has a simple answer: Bayesian methods usually require solving complicated multiple integrals, many times requiring the use of numeric methods (e.g., Cantet et al., 1992) that may not be feasible. The recent development of MCMC techniques (Gibbs sampling and others) has given a solution to many problems that were unsolvable before, due to the impossibility of evaluating these integrals. The second question has a more difficult answer and requires a detailed comparison of how inference is made in each of the schools. Subsequent sections of this paper provide this comparison and expose reasons for and against the use of one or the other form of inference in the common statistical problems of animal breeding.

The Frequentist School

The frequentist school was developed in the 1930s and 1940s and was based on the earlier work of Karl Pearson and Ronald Fisher. The most relevant names associated with this school are Jerzy Neyman and Egon Pearson (son of Karl Pearson), who dealt with inference, and Abraham de Wald, who worked in decision theory.

Inference and Precision

Most animal breeding work relates to estimation of genetic values or parameters and to testing hypotheses; only recently has some work related to decision theory been published (e.g., Wooliams and Meuwissen, 1993). Below are some of the features of inference in the frequentist school.

Repeating an experiment conceptually an infinite number of times we can arrive at an infinite number of confidence intervals, which would include the true value of the parameter in, say, 95% of the cases. Notice that when a confidence interval is given, for example [0.15, 0.25] for h^2, this does not mean that the probability of h^2 being between 0.15 and 0.25 is 95%. What we say is that if the experiment would be repeated an infinite number of times, we would get an infinite number of confidence intervals (of which [0.15, 0.25] is just an example) that would contain the true value of h^2 in 95% of the cases. Neyman and Pearson (1933) suggested that the "scientific behavior" should be to act as if the interval obtained was the true one, being sure that, in the long run, we will be right in 95% of the cases.

Assuming the risk of considering a hypothesis to be true when it is not, or of rejecting the hypothesis when it is true, some regions of acceptance or rejection can be defined, based on the distribution of the sample under the hypothesis to be tested when the experiment is repeated an infinite number of times. The hypothesis is accepted or rejected depending on whether the actual sample falls in one of the regions or in the other. After the hypothesis is accepted, the "scientific behavior" should be to act as though this hypothesis were true.

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run, of experience, we shall not be too often wrong.
>
> Neyman and Pearson (1933)

This procedure is considered to be based on a "subjective belief" by some philosophers of science (Howson and Urbach, 1996), a statement that may be considered surprising by many members of this inference school. Note the method gives no clues about how probable is a hypothesis related to the other. The answer given by a test of hypothesis is just yes or no, because the risk and the corresponding regions are fixed *before* the experiment is performed. This is a rather weak answer from an inference point of view, although it is helpful for making decisions.

There are some problems associated with common practice. For example, a hypothesis may be that two breeds have the same mean for some trait. Note that by augmenting sample size sufficiently, this hypothesis will be always rejected. Thus, in a well-designed experi-

ment, it should be decided when (that is, how large the difference between means should be) a hypothesis should be rejected. The problem arises in poorly designed experiments, in field data, or in tests for traits not considered when designing the experiment. In all these cases, significance will not be associated with relevant differences because these differences were not considered when the experiment was designed. Thus, the answer provided by the test (to accept or reject the hypothesis) might be unsatisfactory, because we can find relevant differences that are not significant or irrelevant differences that are significant. Moreover, if we consider several traits, and the design has been made for only one trait, we will find significant differences for some of these traits even when the null hypothesis is true, just because there is always some probability of falling into the rejection region. The probability of falling into this region for at least one trait increases with the number of traits considered.

One of the main problems of scientific inference is the choice between several models that can explain the results obtained. Many times these models are nested. For example, it has to be decided whether or not a maternal effect explains a part of the variability of the data, whether or not there is a threshold for one or more traits, whether or not to include the effect of a QTL, and so on. For nested models, the significance of the ratio of the likelihoods can be approximated using the chi-square distribution, if there are enough data. Then, a hypothesis test holds and the null or the alternative hypothesis can be selected. After deciding whether the alternative hypothesis is accepted or not, the researcher will *behave* as though the model chosen was the right one, dismissing the other models as false. The main problems of this procedure are that it only can be used for nested models and that it is based on asymptotic properties. Thus, it is sometimes difficult to decide whether we have enough data to make the requisite choices among models.

Selection Indices and BLUP as Frequentist Predictors

Selection indices appeared in a paper of Karl Pearson on natural selection (Pearson, 1903), although they were used much later in animal breeding. First, Smith (1936) proposed selection indices for several traits in plant breeding. Apparently, it was Fisher who proposed the method to Smith, because he says in his article that section I [Theory] is little more than a transcription of Fisher's suggestions. Later, Hazel (1943) applied selection indices to the field of animal improvement. In 1949, BLUP appeared as a ML method, which happened not to be the case (Henderson, 1949, 1950), and it was practically forgotten for more than 20 yr. Henderson (1973) gives details of the history and development of BLUP. Here, we will examine the frequentist properties of the two methods.

Consider the linear model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is the data vector, $\mathbf{b}$ the vector of fixed effects, $\mathbf{u}$ the vector of random effects (breeding values), $\mathbf{e}$ the vector of residuals, and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices. The predictor

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

where $\mathbf{G}$ and $\mathbf{V}$ are the known (co)variance matrices of the random effects and the data, respectively, and $\hat{\mathbf{b}}$ is the generalized least squares estimator of the fixed effects, gives the BLUP of $\mathbf{u}$. Frequentist statisticians distinguish between estimating a fixed effect, which remains constant for each conceptual repetition of the experiment, and predicting a random value, which changes in each repetition. However, this distinction seems to be rather unnecessary. When there are not fixed effects, removing $\mathbf{X}\hat{\mathbf{b}}$ from the formula we obtain the expression of the selection index, or best linear predictor.

Inverting $\mathbf{V}$ is not easy when thousands of data are available, as usually happens in animal breeding. Henderson (1963) demonstrated that solving the equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad [2]$$

resulted in the same and $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ as before. Later, he found an algorithm to calculate $\mathbf{G}^{-1}$ directly from pedigrees, which simplified the problem and made Eq. [2] practical. Since then this method has been widely used by animal breeders. These equations [2] are called *mixed-model equations* in the animal breeding literature.

It should be noted that BLUP is said to be unbiased in a sense differing somewhat from that employed in frequentist statistics. Consider one of the elements of $\hat{\mathbf{b}}$, say, $\hat{b}$ for simplicity. Taking an infinite number of samples, the values of $\hat{b}$ obtained will have an average value

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

That is, the estimates $\hat{b}$ are distributed around the true value b. Then we say that $\hat{b}$ is an unbiased estimator of b. However, this condition cannot be met for random values, because they change in each repetition of the experiment. For random values we use the expression

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

to express unbiasedness. This latter property is much less appealing than the former one. The predicted values $\hat{u}$ in infinite repetitions of the experiment are not distributed around the true value u, because in each repetition we obtain not only a different $\hat{u}$ but also a different u. The risk of the estimator is also different.

In general, when a quadratic loss function $\ell(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$ is used,

$$\begin{aligned}\text{Risk } (\hat{\theta}) &= \text{E } [\ell(\theta, \hat{\theta})] = \text{E } (\theta - \hat{\theta})^2 \\ &= \text{E } [(\theta - \hat{\theta}) - \text{E}(\theta - \hat{\theta}) + \text{E}(\theta - \hat{\theta})]^2 \\ &= \text{var } (\theta - \hat{\theta}) + [\text{E}(\theta) - \text{E}(\hat{\theta})]^2\end{aligned}$$

Thus, the risk can be deconstructed into two quantities: the variance of the error of estimation θ, $\hat{\theta}$, and the square of the bias. For fixed effects, because b is a fixed value and does not change in the conceptual repetitions of the experiment, E(b) = b, and the variance of the error of estimation becomes var(b – $\hat{\text{b}}$) = var($\hat{\text{b}}$). Then,

$$\text{Risk } (\hat{\text{b}}) = [\text{bias } (\hat{\text{b}})]^2 + \text{variance } (\hat{\text{b}})$$

However, for random effects[3], the variance of the error of estimation is as follows:

$$\begin{aligned}\text{variance } (\text{u} - \hat{\text{u}}) &= \text{variance } (\text{u}) - 2 \text{ cov } (\text{u},\hat{\text{u}}) + \text{variance } (\hat{\text{u}}) \\ &= \text{variance } (\text{u}) - 2 \text{ variance}(\hat{\text{u}}) + \text{variance } (\hat{\text{u}}) \\ &= \text{variance } (\text{u}) - \text{variance } (\hat{\text{u}}).\end{aligned}$$

Then,

$$\text{Risk } (\hat{\text{u}}) = [\text{bias } (\hat{\text{u}})]^2 + \text{variance } (\text{u}) - \text{variance}(\hat{\text{u}}).$$

The variance (u) cannot be changed because it is a population parameter that depends on the genetic determination of the trait. As information on the random effect increases, variance ($\hat{\text{u}}$) increases, tending toward variance (u), and the quadratic risk decreases. However, for fixed effects, variance ($\hat{\text{b}}$) decreases as information increases, and the quadratic risk also decreases. This different behavior of variance ($\hat{\text{u}}$) and variance ($\hat{\text{b}}$) is a typical source of confusion.

There may be an infinite number of estimators having the same risk but with different bias and variance. To solve the problem, the predictor with minimum variance could be chosen from among the unbiased ones, even if they may have a higher variance of the error of estimation and consequently a higher risk than some biased predictors, and BLUP is derived in this way. To compute BLUP, the variances of the random effects **u** and **e** in model [1] should be known. This is usually not the case, so it is often recommended to estimate these from the data by ML methods, or to use estimates from the literature.

[3]Here bias($\hat{\text{u}}$) = [E($\hat{\text{u}}$) – E(u)], which is different from that normally used in frequentist statistics for parameter estimation. Bias should be [E($\hat{\text{u}}$|u) – u], but this is different from zero and, consequently, BLUP is biased. To state that BLUP is unbiased by changing the usual definition of bias seems to be a rather liberal use of the language. Besides, the term *best* is somewhat misleading. BLUP is only best among a restricted class of predictors (the "unbiased" ones), and only when the true variances of the random effects are used.

The Method of Maximum Likelihood

The concept of likelihood and the method of ML were developed by Fisher between 1912 and 1922, although there are historical precedents attributed to Bernouilli (1782, translated by Kendall, 1961). By 1912 the theory of estimation was in an early state and the method was practically ignored. However, Fisher (1922) published a paper in which the properties of the estimators were defined and he found that this method produced estimators with good properties, at least asymptotically. The method was then accepted by the scientific community and has since been used frequently (Fisher 1912, 1922).

Suppose we know how the samples are distributed in infinite repetitions of an experiment, or how infinite samples would be distributed. We will use the following notation: f(**y**|u), which means the density function of the sample **y** for a "given" value of u. Thus, here the variable is **y** because we examine how it is distributed when repeating the experiment, and u is a fixed value that we provide in order to examine the probability density distribution of **y** for this given value. The ML method consists of finding which value of u maximizes the probability of our sample **y**. To explain the concept of likelihood and the rationale for using its maximum we put forth an example.

Consider finding the average weight of rabbits of a breed at 8 wk of age. We take a sample of one rabbit, and its weight is $y_0 = 1.6$ kg. Figure 1 shows the density functions of several possible populations from which this rabbit can come, with population means $m_1 = 1.50$ kg, $m_2 = 1.60$ kg, $m_3 = 1.80$ kg. Notice that at y_0 the probability densities of the first and third population $f(y_0|m_1)$ and $f(y_0|m_3)$ are lower than that of the second one $f(y_0|m_2)$. Therefore, it seems more *likely* that the rabbit comes from the second population. All the values $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$, . . . define a curve with a maximum in $f(y_0|m_2)$. This curve varies with m, and the sample y_0 is a fixed value for all those density functions. Each value represents an "instant probability," as Fisher (1912) first called it, in the sense that it belongs to a density function. However, it is obvious that the new function defined by these values is *not* a density function, because each value belonged to a different probability density function. Fisher (1912) proposed to take the value of m that maximized $f(y_0|m)$ because from all the populations defined by $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$, . . . it is the one for a given value of m that makes the sample y_0 most probable. Here the word *probability* can lead to some confusion, because these values belong to different density functions and the function defined with all of them is not a probability function. Thus, Fisher preferred to use the word *likelihood* for all the values considered together. Notice that the method of ML is *not* the one that makes the sample most probable. This method provides a value of the parameter such that *if this were the true value* the sample would be most probable. Here we face a problem of notation: speaking about a set of density functions

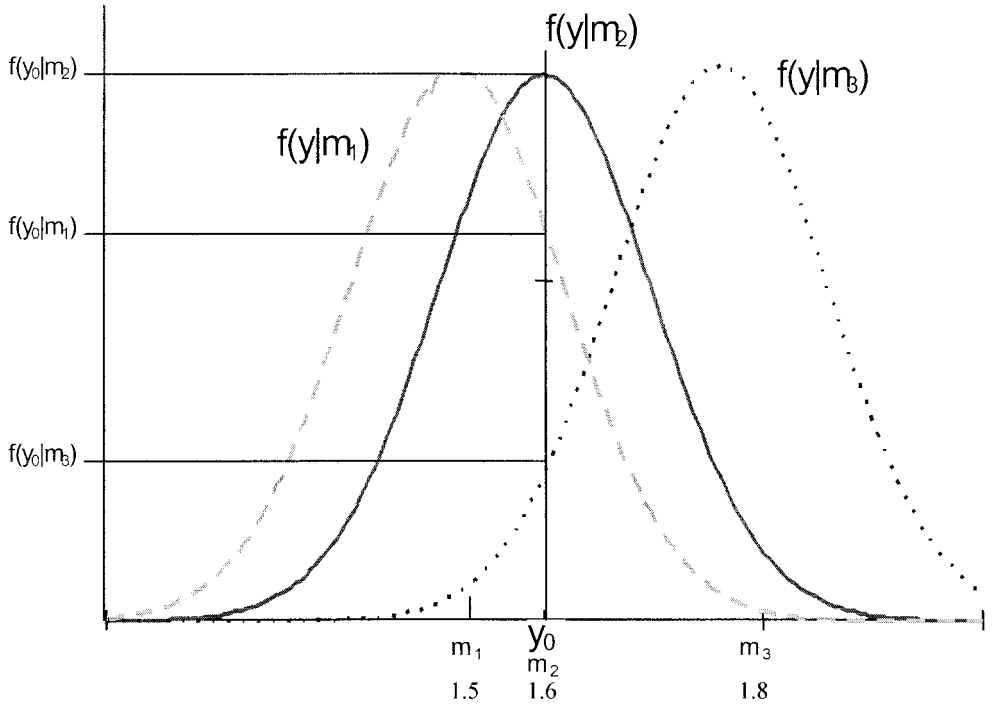

Figure 1. Construction of the likelihood curve.

$f(y|m_1)$, $f(y|m_2)$, $f(y|m_3)$ . . . for a given y is the same as speaking about a function L(m|y) that is not a density function and consequently it does not represent probabilities[4]. We will use the notation L(m|y) as the one that expresses the likelihood most clearly.

Fisher (1912, 1922) not only proposed a method of estimation, but also proposed the likelihood as a *degree of belief* different from the probability but that allowed uncertainty to be expressed in a similar manner. What Fisher proposed is to use the whole likelihood curve, not only its maximum, a practice rather unusual nowadays with the exception of QTL analyses. In these cases there is the risk of confusing likelihood and probability. For example, Von Mises (1957) accused Fisher of exposing with great care the differences between likelihood and probability, only to forget it later and use the word *likelihood* as we use *probability* in common language. Today, frequentist statisticians typically use only the maximum of the curve because it has good properties in repeated sampling. The method is asymptotically unbiased, sufficient when there are sufficient estimators, efficient, optimum asymptotically normal, and so on. Repeating the experiment an infinite number of times, the estimator will be distributed near the true value, with a variance that can also be estimated. But all those properties are asymptotic, and thus there is no guarantee with small samples about the goodness of the estimator. The animal breeder usually works with a large number of animals, but here "small samples" does not mean a small total number of data. Depending on the problem and on the distribution of the data, the information for estimating some parameters can come from a reduced number of data. Besides, the ML estimator is not necessarily the estimator that minimizes the risk[5]. Nevertheless, the method has an interesting property apart from its frequentist properties: any reparametrization leads to the same type of estimator. For example, the ML estimator of the variance is the square of the ML estimator of the standard deviation, and a function of ML estimators is also a ML estimator.

From a practical point of view, the ML estimator is an important tool for the applied researcher. The frequentist school developed a list of properties that good estimators should have but does not give rules about how to find them. Maximum likelihood is a way of obtaining estimators with (asymptotically) desirable properties. It is also possible to find a measurement

[4]Classic texts of statistics such as Kendall's (Stuart and Ord, 1991) contribute to the confusion by using the notation L(y|m) for the likelihood. Moreover, some authors distinguish between "given a parameter" (always fixed) and "giving the data" (which are random variables). They use (y|m) for the first case and (m;y) for the second. Thus, likelihood can be found in textbooks as L(m|y), L(y|m), f(y|m), and L(m;y).

[5]A well-known example is the James-Stein paradox: Suppose we have K populations normally distributed with the same variance and we try to estimate the means m_i with minimum risk $E\left[\sum(m_i - \hat{m}_i)^2\right]$. The maximum likelihood estimator takes the sample means of each population $m = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k]$. James and Stein (1961) found an estimator on lower risk for $K > 2$.

of precision from the likelihood function itself. If the likelihood function is sharp, its maximum gives a more *likely* value of the parameter than other values near it. Conversely, if the likelihood function is rather flat, other values of the parameter will be almost as *likely* as the one that gives the maximum to the function. Fisher (1922) introduced the concept of "amount of information" in order to measure the accuracy of the estimation. This "amount of information" is defined as

$$E[d\ \log L(u|\mathbf{y})/du]^2$$

Logarithms are taken to facilitate an additive measure of information. For example, if n individuals of a sample are independent,

$$L(m|y_1, \ldots, y_n) = L(m|y_1) \cdot L(m|y_2) \cdot \ldots \cdot L(m|y_n),$$

$$\text{and } \log\ [L(m|y_1, \ldots, y_n)]$$

$$= \log[L(m|y_1)] + \log[L(m|y_2)] + \ldots + \log[L(m|y_n)]$$

This quantity is an indicator of the sharpness of the likelihood function. If the function is sharp, the quantity [d logL(u|**y**)/du] will be high in absolute value, and thus the amount of information will be large (the square prevents negative values). Conversely, if the function is rather flat, this quantity will be low, meaning that some solutions far away from the maximum have almost the same likelihood or "degree of belief." In small samples, the likelihood can be asymmetrical or multimodal, giving varying chances (degrees of belief) to points that are not near the maximum.

The frequentist school has reduced the problem by making inferences only at the maximum of this function. With infinite repetitions of the experiment the likelihood function would converge asymptotically to a normal with mean the true value and variance the inverse of the information quantity (e.g., Stuart and Ord, 1991). The main problem is that this approximation is based on the central limit theorem, needing large samples, but it is not possible to determine how large the samples should be to ensure normality.

Because likelihood means "rational degree of belief," it would be expected that ratios of likelihoods would indicate differences among models in hypothesis tests, and some statisticians (Edwards, 1992) recommend the use of the likelihood ratio reasoning in this way. Unfortunately, likelihoods are not quantities that can be treated as probabilities; a likelihood four times higher than another one does not lead to a "degree of rational belief" four times higher. Nevertheless, the likelihood ratio has good asymptotical frequentist properties and is currently used in testing hypotheses.

REML as a Frequentist Estimator

Although Fisher proposed both the method of ML and the analysis of variance, he never intended to estimate variance components by ML. It was Hartley and Rao (1967) who first proposed this. The ML procedure leads to complex equations that should be solved approximately and using iterative algorithms. This causes computing difficulties, and because of them ML procedures were not used in the field of animal breeding until recently. A conceptual difficulty also appeared when mixed models were used. The ML estimation of variance components does not take into account the loss of degrees of freedom caused by estimation of the fixed effects. This can be worrying when the model includes fixed effects with many levels, as in national genetic evaluations of dairy cattle. The solution, proposed by a simple model by Thompson (1962) and generalized by Patterson and R. Thompson (1971), was called residual or restricted ML (**REML**). In recent years, REML has been the preferred method of animal breeders for variance component estimation. Following the Harville (1977) presentation of REML, this method is based on projecting the data in a subspace free of fixed effects and maximizing the likelihood in this subspace. The REML method has the advantage of giving the same results of the ANOVA in balanced designs (ANOVA is unbiased with minimum variance for such designs). Alternatively, ML estimation produces different results from ANOVA in balanced designs, which is somewhat disturbing. To better explain the differences between ML and REML, consider a very simple model:

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

where $\mathbf{X} = \mathbf{1}$ is a vector of 1's. The (co)variance matrix of the residuals is $\mathbf{V} = \mathbf{I}\sigma^2$, and its determinant $|\mathbf{V}| = (\sigma^2)^n$. The likelihood function, when $y \sim N(\mu,\sigma^2)$ and the sample is of size n, is

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

$$(\mathbf{y} - \mathbf{1}\mu)/2\sigma^2].$$

The value of σ^2 that maximizes $L(\sigma^2|\mathbf{y})$ is

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

Notice that μ must be known to obtain the ML estimate of the variance. Because we do not know μ, we substitute the ML estimate of μ:

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

The ML estimate of the variance is:

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

Although this estimate is a function of another estimate, it is still a ML estimate, and thus it has all the asymptotic properties from before, and there is not a formal reason to reject it.

To calculate the REML estimates the data are projected in a subspace without fixed effects. A projection matrix **K**, that follows

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

is found (i.e., $\mathbf{K}$ has been chosen to follow the condition $\mathbf{K}'\,\mathbf{1} = \mathbf{0}$). For example,

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

satisfies this condition.

The variance of $\mathbf{K}'\mathbf{y}$ is $\mathbf{K}'\mathbf{V}\mathbf{K} = \mathbf{K}'\mathbf{K}\sigma^2$ and the likelihood is as follows:

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

where

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

The value of σ^2 that maximizes $L(\sigma^2|\mathbf{K}'\mathbf{y})$ is:

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

It can be shown that, in general,

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

In this case, $\mathbf{X} = \mathbf{1}$, $\mathbf{V} = \mathbf{I}\sigma^2$. Thus,

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

and

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

Thus, the REML estimate of the variance is

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

which is similar to the ML estimate, but dividing by n – 1 instead of by n. Despite the similarity of both estimates, here there is no substitution of a true value by its estimate. Rather, as a result of the estimation process an expression in the formula $(1/n)\Sigma y_i$ is coincident with $\hat{\mu}$. The degree of freedom lost when estimating μ is reflected in dividing by (n – 1) instead of n. When there are many fixed effects, this distinction is important because the ML estimate is

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

whereas the REML estimate is

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

where $r(\mathbf{X})$ is the rank of $\mathbf{X}$ and $\hat{\mathbf{b}}$ is the ML estimate of $\mathbf{b}$.

With the proposed matrix $\mathbf{K}$, the analysis of variance is made on the vector

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

in order to avoid the estimation of the mean μ. By using differences between data,

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

μ is removed. Several matrices follow the same condition $\mathbf{K}'\mathbf{1} = \mathbf{0}$. This analysis could be also made on

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

and the same result obtained. This does not mean that every $\mathbf{K}$ would be useful; for example, a $\mathbf{K}$ with half of the rows null will follow the condition $\mathbf{K}'\mathbf{1} = \mathbf{0}$, as will the matrix $\mathbf{K} = \mathbf{0}$. It is intended to find matrices that will not lose information relative to the dispersion parameters, which implies introducing in $\mathbf{K}$ a maximum number of independent linear contrasts of *residuals*. It is not important which $\mathbf{K}$ is used. In fact, the usual way in which REML is found in the literature is in terms of solutions to the mixed-model equations [2] and components of the mixed-model equations, an expression in which $\mathbf{K}$ does not appear explicitly. Notice that in the example $\mathbf{K}$ has dimensions (n – 1) × n (there are only n – 1 couples of data to calculate their difference) and a degree of freedom is lost. Thus, summarized information was used because the new data vector $\mathbf{K}'\mathbf{y}$ has only n – 1 elements. To obtain a geometrical representation of this sample a n-dimensional space is needed, but using REML only a n – 1 dimensional space is required to represent the sample that will be used in estimation, thus the reference to working in a *restricted* space, having lost a degree of freedom. When there are many fixed effects or many levels of one effect, this is much more notorious.

As Searle et al. (1992) stated, there is not a clear (frequentist) argument for preferring REML to ML. In the former example, the REML estimate was unbiased but had a higher risk than the ML estimate, although this may not happen in more complex cases. In other circumstances, the risk of REML will be higher or lower than the risk of ML, depending on the true values of

the parameters and the structure of the data. Restricted ML became the preferred method among animal breeders, particularly among dairy cattle breeders, for indirect reasons rather than for clear statistical advantages such as minimizing risk. One reason to prefer REML is that in practical problems such as genetic evaluation of dairy cattle there may be many levels of herd-year-season effect and other effects, and ML does not take into account the loss in degrees of freedom due to the estimation of these effects. However, REML is used in the field of animal breeding as the method of choice independent of whether there are few or many levels of fixed effects, perhaps because, as Dawid (1976, cited by Robinson, 1991) stated for the *t*-test in the quote that initiates this paper, "everyone knows that REML is a good thing."

The Bayesian School

The Bayesian school was, in practice, founded by Count Laplace through several works published from 1774 to 1812, and it had a preponderant role in scientific inference during the 19th century (Stigler, 1986). Some years before Laplace's first paper on the matter, the same principle was formalized in a posthumous paper presented at the Royal Society of London and attributed to a rather obscure priest, rev. Thomas Bayes (who never published a mathematical paper during his life). Apparently, the principle upon which Bayesian inference is based was formulated before. Stigler (1983) attributes it to Saunderson (1683–1739), a blind professor of optics who published a large number of papers on several fields of mathematics. Fisher's work on likelihood in the 1920s and the work of the frequentist school in the 1930s and 1940s eclipsed Bayesian inference until the 1960s, when a "revival" was started that has increased since. The Bayesian paradigm was introduced in animal breeding by Daniel Gianola, first in work on threshold traits published with J. L. Foulley, and later in series of papers in which virtually all topics about animal breeding were addressed (see Gianola and Fernando, 1986; Gianola et al., 1990 for reviews).

Inference and Precision

In a Bayesian context, the objective is, given the data, to describe the uncertainty about the true value of some parameter, using probability as a measurement of this uncertainty. For example, if the parameter of interest is the heritability of some trait, the aim of Bayesian inference is to find a probability density of the heritability given the data, $f(h^2|\mathbf{y})$, where $\mathbf{y}$ is the vector of observations (Figure 2). When this distribution is obtained, inferences can be made in multiple manners: for example, we can calculate the probability of h^2 to be between 0.1 and 0.3, by integrating the function between these values. We can also find which is the shortest interval in which the probability of finding h^2 is more than 95%. If we are interested in point estimation, we can give several values of h^2 calculated from the distribution $f(h^2|\mathbf{y})$. The mode is the value that maximizes $f(h^2|\mathbf{y})$, i.e., the most probable value of h^2 that we can infer given the data. It can be shown that this is the value that minimizes risk when the loss function has a null value for $h^2 = \hat{h}^2$ and is 1 in other cases. The probability that the true value of h^2 is greater than the median is the same as being less. Thus, the median minimizes risk when the loss function is $|h^2 - \hat{h}^2|$. Finally, the mean minimizes the risk when the loss function is the quadratic one, $(h^2 - \hat{h}^2)^2$.

In order to make all of these inferences, the density function $f(h^2|\mathbf{y})$ should be obtained. To do this, we will use the Bayes theorem. The probability of two events happening together is

$$P(A,B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \qquad [3]$$

Thus,

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

In the present case:

$$f(h^2|\mathbf{y}) = f(\mathbf{y}|h^2)\ f(h^2)/f(\mathbf{y}) \propto f(\mathbf{y}|h^2)\ f(h^2) \qquad [4]$$

where, $\propto$ means "proportional to."

Notice $f(h^2|\mathbf{y})$ is a function of h^2, but not of $\mathbf{y}$, which is given (it is "fixed"); therefore, $f(\mathbf{y}|h^2)$ is a function of h^2 but not of $\mathbf{y}$, which means that $f(\mathbf{y}|h^2)$ is the likelihood, as we defined it before. Moreover, $f(\mathbf{y})$ is a constant for the same reason: it does not depend on h^2, and $\mathbf{y}$ is fixed. Finally, $f(h^2)$ is the probability of h^2 that does not depend on our data. The criticism of the Bayesian inference is focused on this last probability, called probability *a priori* because it can be defined independently of the data and before the start of the experiment (e.g., Glymour, 1981). In defining likelihood, we stressed that it was the probability of the sample *if the value of the parameter were the true value*. Now we weigh this by the probability that the parameter is indeed the true value. Sometimes this prior probability is well-defined (e.g., the *prior probability* of obtaining a recessive individual when crossing two individuals with a heterozygous trait is 1/4), but for heritability the value of this prior probability may not be obvious. The density function $f(h^2|\mathbf{y})$ is called *posterior* probability, posterior density, or, more commonly, posterior distribution. A common expression of this posterior density is also

$$f(h^2|\mathbf{y}) \propto L(h^2|\mathbf{y})\ f(h^2)$$

containing $L(h^2|\mathbf{y})$, which is the density function of the data $f(\mathbf{y}|h^2)$, but when considered as a function of h^2 is the likelihood. We will use this notation because it is more coherent with the use of the bar "|" that means "given." Thus, $L(h^2|\mathbf{y})$ is a function of h^2 given the data.

Strictly speaking, all probabilities of a parameter are based on a certain set of assumptions. These probabili-

ties cannot be constructed without them (e.g., the assumption that the sample results from a random process, that the sample space is uniform, that the observer does not influence the observation, etc.). If we call H the set of hypothesis used to describe $f(h^2)$, the right notation should be $f(h^2|H)$, and therefore the theorem should be expressed as $f(h^2|\mathbf{y},H) \propto L(h^2|\mathbf{y},H)\ f(h^2|H)$, but we will suppose this implicitly in order to simplify the notation.

In the Bayesian school, there is not a unique possibility for point estimation, and the point estimate of choice can be based on the risk function or other reasons. For example, if the prior density is bimodal, the mean, which minimizes the quadratic risk, can have a much lower probability than the modes. In this case, the point estimate is not consistent with the uncertainty about the parameter h^2. Another example of the risk of using point estimation is the case in which an asymmetric posterior density of a genetic correlation gives a negative mode but a much higher probability of being positive (Figure 3).

Confidence intervals for the point estimates can be asymmetric. In the case of asymmetric posterior densities, if we prefer the shortest confidence interval, it is necessarily asymmetric. There is a practical difficulty for the estimation of confidence intervals, or "credible regions" as Bayesians prefer to call them. They require knowledge of the posterior density, and it is necessary to integrate this density between the limits of the interval in order to know the probability of the interval. Moreover, to calculate the posterior density it is necessary to know $f(\mathbf{y})$, a value difficult to calculate, because

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

Because **y** is a vector, all these integrals are multidimensional and difficult, if not impossible, to calculate, even by approximate methods. Until recently this has been a weak point in application of Bayesian methods, and it has been the main hindrance to their development in animal breeding. Now, MCMC methods (Gibbs sampling, Metropolis-Hastings, etc.) have solved the problem.

Because there are not infinite repetitions of the experiment, there is nothing that can be considered as "bias." Moreover, there are no "fixed effects," because all parameters are random variables in a Bayesian context, including those that are considered "fixed" in a frequentist context. Uncertainty is expressed as the probability of a parameter or an effect having a particular true value, so they must be random variables.

The result of a test of hypothesis is different in the Bayesian school. For example, if the hypothesis to be tested (H_0) is that two breeds have the same mean for a trait, and the alternative (H_1) is that the means are different, in a Bayesian context it is calculated as

$$P(H_0|\mathbf{y})/P(H_1|\mathbf{y})$$

Thus, if this ratio is, for example 7, the hypothesis of having equal means is seven times more probable than the hypothesis of having different means. These posterior distributions can be expressed as

$$\frac{P(H_0|\mathbf{y})}{P(H_1|\mathbf{y})} = \frac{P(\mathbf{y}|H_0)\cdot P(H_0)/P(\mathbf{y})}{P(\mathbf{y}|H_1)\cdot P(H_1)/P(\mathbf{y})}$$

$$= BF\cdot\frac{P(H_0)}{P(H_1)}$$

wherein the ratio of the likelihoods is called the Bayes factor (**BF**). Because the prior probabilities of both hypotheses are often considered equal (many times in order to give a false impression of objectivity), Bayes

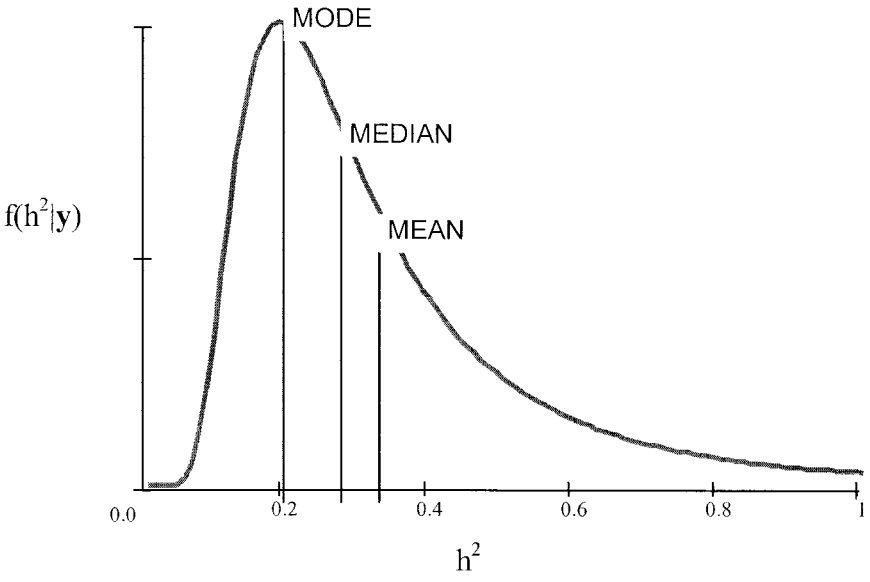

Figure 2. Example of a probability density for heritability (h^2) given the data (y) and estimates from posterior distribution.

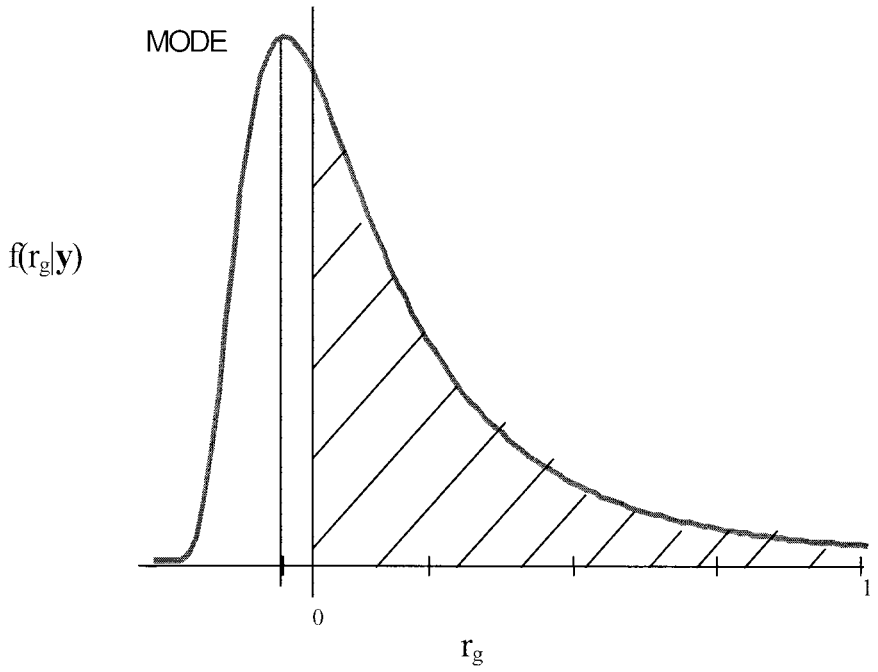

Figure 3. Asymmetrical posterior distribution for a genetic correlation (rg) illustrating an inconsistency in sign between the mode and the majority of estimates.

factors coincide with the ratio of posterior probabilities and are used to decide which model is chosen.

Notice that for nested models the frequentist school also uses likelihood ratio; thus, sometimes the numerical result of a frequentist test and a Bayesian test can be the same. However, the interpretation is completely different. In a frequentist context, this ratio is used because under infinite repetitions of the experiment it will be approximately distributed as a chi-square. Thus, areas of acceptation and rejection of the hypothesis are established and a decision is made based on this distribution. In the Bayesian context, the ratio of likelihoods is used because (under equal prior probabilities) it is equal to the ratio of posterior probabilities and gives an exact account of the evidence of one hypothesis over the other. This is neither restricted to nested hypotheses nor is it an approximated result. The procedure is also not restricted to either of only two hypotheses. The ratio of the likelihoods is the part of the inference that is supported by the data and shows how the data modify the prior state of knowledge to obtain the posterior one. Lavine and Schervish (1997) give a detailed comparison of Bayes factors and *P*-values from likelihood-ratio tests.

Although Bayes factors summarize the information provided by the data, it is important to notice that, in a Bayesian context, they do not have any inference property without considering the prior information. Notice that $P(\mathbf{y}|H_0)$ is the probability of obtaining the current data *only if the hypothesis* H_0 *is true*. We need to multiply $P(\mathbf{y}|H_0)$ by the probability that this hypothesis is true, $P(H_0)$, and to do the same with $P(\mathbf{y}|H_1)$ and $P(H_1)$, in order to make proper inferences[6]. When a model is chosen only under the information provided by the Bayes factor, equal prior probabilities are implicit.

An interesting procedure for inferences that has no counterpart in frequentist statistics is the Bayesian model averaging. It consists of simultaneously using several models for inferences, weighted according to their posterior probabilities. For example, if we are interested in estimating the heritability of a trait and we have two models constructed under the hypotheses H_0 and H_1 (for example, with and without a particular effect), then

$$\begin{aligned} P(h^2|\mathbf{y}) &= P(h^2, H_0|\mathbf{y}) + P(h^2, H_1|\mathbf{y}) \\ &= P(h^2|H_0, \mathbf{y})\, P(H_0|\mathbf{y}) + P(h^2|H_1, \mathbf{y})\, P(H_1|\mathbf{y}) \end{aligned}$$

where $P(h^2|H_0, \mathbf{y})$ is the posterior distribution of h^2 under the first model and $P(H_0|\mathbf{y})$ is the probability that this model is true, and the same holds for the second model. This can be generalized to several models simultaneously. A tutorial for Bayesian model averaging has been published by Hoeting et al. (1999).

[6]For example, if I want to infer the height of the typical Scotsman and I take a sample of one of them, obtaining a height of 1.70 m, the hypothesis that maximizes the probability of my sample is not that the average Scotsman measures 1.70 m, but that *all* Scotsmen measure 1.70 m. Because I have prior information about them, I know they are humans, and that the height of humans follows an approximately normal distribution, and so on, then I give a null prior probability to this hypothesis and I infer on other grounds.

The Use of Prior Information

Sometimes there is well-defined prior information on the parameters to be estimated. In these cases the use of prior information is not controversial, as we will now see taking an example by Fisher. Let us take a trait controlled by a single gene with two alleles with dominance, for example, black skin is dominant over brown skin in some mice. We try to know whether a black mouse, son of a heterozygous mates (Aa × Aa), is homozygous (AA) or heterozygous (Aa). In order to assess this, we mate this mouse with a brown (recessive) mouse, and we obtain seven offspring, all of them black. What is the probability that the black mouse is heterozygous, given this data? According to the Bayes theorem shown in Eq. [4],

$$\begin{aligned}&P(AA|y = 7 \text{ black descendants})\\ &\quad = L(AA|y = 7 \text{ black descendants}) \cdot\\ &P(AA)/P(y = 7 \text{ black descendants})\\ &P(Aa|y = 7 \text{ black descendants})\\ &\quad = L(Aa|y = 7 \text{ black descendants}) \cdot\\ &P(Aa)/P(y = 7 \text{ black descendants})\end{aligned}$$

We are trying to test whether the black mouse we have is or is not homozygous. The expectation from a mating between two heterozygotes is ¼ homozygous black, ½ heterozygotes, and ¼ homozygous brown. A black mouse from such a mating has, prior to any test-mating, a probability of 1/3 of being homozygous and of 2/3 of being heterozygous, because mating Aa with Aa gives three types of black animals, AA, Aa, and aA, with the same probability. We represent as Aa both of the heterozygotes Aa and aA. Thus, we have P(AA) = 1/3, and P(Aa) = 2/3.

After performing the experiment, the likelihood of the black mouse being homozygous (AA) is 1; that is, given seven black offspring, because *if the black father's genotype is AA*, all descendants from crossing an AA mouse with brown (aa) mice are expected to be black (Aa). Thus, L(AA|y = 7 black descendants) = 1. Further, the likelihood of the black mouse being heterozygous is $(1/2)^7$, given seven black offspring, because *if the black father's genotype is Aa*, the probability of obtaining one black descendent would be 1/2, because the descendants of a cross between Aa and a brown (aa) mouse are expected to be Aa (black) and aa (brown) in equal proportion. Thus, the probability of obtaining seven black mice is $(1/2)^7$ and L(Aa|y = 7 black descendants) = $(1/2)^7$.

Now, there are two possibilities of having seven black descendants. Either the individual is homozygous and had seven black descendants, or it is heterozygous and had seven black descendants. Thus, the probability of having seven black descendants is:

$$\begin{aligned}&P(y = 7 \text{ black descendants})\\ &\quad = \text{Prob (AA and } y = 7 \text{ black descendants)}\\ &\quad + \text{Prob (Aa and } y = 7 \text{ black descendants)}\end{aligned}$$

According to the laws of probability shown in Eq. [3],

$$\begin{aligned}&\text{Prob } (y = 7 \text{ black descendants, and AA})\\ &\quad = P(y = 7 \text{ black descendants}|AA) \cdot P(AA)\\ &\quad = L(AA|y = 7 \text{ black descendants}) \cdot P(AA) = 1 \cdot 1/3 = 1/3\\ &\text{Prob } (y = 7 \text{ black descendants, and Aa})\\ &\quad = P(y = 7 \text{ black descendants}|Aa) \cdot P(Aa)\\ &\quad = L(Aa|y = 7 \text{ black descendants}) \cdot\\ &P(Aa) = (1/2)^7 \cdot (2/3) = 1/192\end{aligned}$$

and thus,

$$P(y = 7 \text{ black descendants}) = 1/3 + 1/192$$

Thus, the probability that our black mouse is homozygous (AA), given that seven descendants were black, is

$$P(AA|y = 7 \text{ black descendants}) = (1/3) \cdot 1/(1/3 + 1/192)$$

and the probability that our black mouse is heterozygous (Aa) and had seven black descendants when crossing it with a brown (aa) mouse is

$$\begin{aligned}&P(Aa|y = 7 \text{ black descendants})\\ &\quad = (1/2)^7 \cdot (2/3)/(1/3 + 1/192)\end{aligned}$$

This is the desired answer.

Selection indexes and other related methods (BLUP or combined selection) use genetic parameters that are presumed to be known. This is a common practice in animal breeding and its use is not controversial. Let us consider the model

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

where $\mathbf{y}$ is the data vector, $\mathbf{u}$ the genetic values, $\mathbf{e}$ the residuals, and $\mathbf{Z}$ the incidence matrix. Our objective is to find the expression of the posterior density of the genetic values $f(\mathbf{u}|\mathbf{y})$. We suppose

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2) \qquad \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$$

thus,

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

$$L(\mathbf{u}|\mathbf{y}) = \text{constant} \cdot \exp\{(-1/2\sigma_e^2)\ (\mathbf{y} - \mathbf{Z}\mathbf{u})'\ (\mathbf{y} - \mathbf{Z}\mathbf{u})\}, \text{ and}$$

$$\begin{aligned}f(\mathbf{u}|\mathbf{y}) \propto L(\mathbf{u}|\mathbf{y}) \cdot f(\mathbf{u}) &= \text{constant} \cdot \exp\{(-1/2\sigma_e^2)\\ &\quad [(\mathbf{y} - \mathbf{Z}\mathbf{u})'\ (\mathbf{y} - \mathbf{Z}\mathbf{u}) + \mathbf{u}'\ \mathbf{A}^{-1}\ \mathbf{u}\ \alpha]\}\end{aligned}$$

where $\alpha = \sigma_e^2/\sigma_u^2$.

This posterior density $f(\mathbf{u}|\mathbf{y})$ has its maximum in

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

which is a well-known result of the theory of selection indexes; it is the best linear prediction of the genetic values. From a Bayesian point of view, because the posterior density is symmetric, this point estimator is at the same time the mean, median, and mode. Confidence intervals can be obtained from f(**u**|**y**) as we have seen before.

In most cases, it is difficult to quantify the prior information pertaining to a parameter. For example, how should the prior information about the heritability of litter size be quantified? In these cases, the non-Bayesian statisticians contend it is not possible to apply Bayes theorem and the problem has no solution if we want to find probabilities about the true value of the parameter. Within the Bayesian school, several solutions have been proposed.

For many cases, probability defined as a frequency is too restrictive a concept to help in solving the inference problem. Thus, expressing the probability as a subjective value between 0 and 1 that indicates the degree of belief of the researcher about the value of the parameter is preferred. Notice that *subjective* does not mean *arbitrary*. As Keynes (1921) stressed, this subjective probability should be well-founded, and a scientist should be able to persuade other scientists about the value of this subjective probability. For example, let us consider the market price of a rabbit carcass. Before taking samples from different markets, we can easily agree that it will be less than 100 euro/kg and more than 0 euro/kg. We can be more precise and think that it will be unlikely that the price will be more than 10 euro/kg or less than 1 euro/kg. Following in the same way, we can assign probabilities to different prices until a probability distribution would be of general agreement. If not, we can compare the results given by two or three possible prior densities, and base our decision on the one we believe in more. This is quite easy to do for univariate problems but becomes factually impossible for multivariate problems. For example, in estimating genetic parameters for three traits we have to determine the prior probability that the first trait has a heritability of 0.1, that the second trait has a heritability of 0.3, the third one has a heritability of 0.3, the genetic correlation of the first and second is –0.7, that of the first and the third is 0.1, and that of the first and second is 0.4, and so on. And this should be done for every possible case! Trying to determine all possibilities can lead the researcher to a psychological crisis; therefore, this way of working is not useful for many animal breeding problems.

Having enough data, the prior probability has little influence on the final result (the posterior density). This argument is based on the practical fact that experiments are often performed when the information about the topic of the experiment is scarce, then prior opinions are vague and it should not be too expensive in most cases to have data enough to override the prior opinion. In these cases, prior opinion is not particularly well defined and it is usual to take prior density functions that facilitate computing the posterior density, avoiding paradoxes and inadmissible results. This argument was introduced in the literature by Jeffreys (1931) and has become popular given the difficulties of using prior information accurately (e.g., Blasco et al., 1998). Although it is somewhat deceptive to use a theory based on the harmonic use of prior information and information provided by the data, this solution is practically acceptable. Analyzing small experiments with expensive data remains problematic because samples are necessarily small and prior information will affect the results.

Thus far, even though the procedure of associating degrees of belief to probabilities is controversial no theoretical difficulties have appeared. There will be researchers that prefer not to subjectively establish the prior probability. However, even if this position is accepted the resulting inference does not lead to paradoxes or contradictions. The problem arises when it is intended to represent the *prior ignorance*. The first Bayesians (Bayes included) used a technique that Keynes (1921) called "Principle of indifference" to describe a situation in which there was no prior information and all possible events have the same probability. To admit this principle implies that in the familiar example of a bag of white and black balls that is often used to teach probability, all possible drawings have the same prior probability. For continuous variables this means that the prior density is a line parallel to the x-axis in a determined interval, and because of that they are known as "flat priors." This terrible postulate is the origin not only of philosophical, but also of mathematical, problems. For example, if we want to represent absolute ignorance about the value of the heritability and we say that all possible prior values have the same probability in the interval [0, 1], then $P(h^2 < 0.5) = ½$. But the event $h^2 < 0.5$ is the same as the event $h^4 < 0.25$, and therefore their probability should be the same $P(h^4 < 0.25) = ½$. This shows that h^4 is more proximal to 0 than to 1. Consequently, we shall say that we have no information about h^2 but we do have information about h^4, which is absurd. The final consequence is that, as Bernardo (1997) remarked, *non-informative priors do not exist*. There is no easy answer to this problem.

The admission of validity of flat priors for representing ignorance is a popular solution, but it does not seem to be well founded. Flat priors are frequently called "non-informative priors." This is not true, because all priors are informative, as we have seen. We can maintain that we have no information about h^2 but we do about h^4. For example, if we admit that h^{16} has a flat prior between 0 and 1, this implies $P(h^{16} > 0.5) = ½$, thus $P(h^2 > 0.5^{1/8}) = P(h^2 > 0.92) = ½$, which few people would find plausible. Box and Tiao (1973) sustain this rather unconvincing proposal. As Edwards (1992) remarks, ignorance about a parameter implies ignorance about any transformation. To say we ignore the prior values of a parameter is not the same as saying we think all of them have the same probability.

Alternatively, ignorance may be represented with respect to the likelihood. The supporters of this solution

think ignorance is only a scale problem and that by changing the scale prior densities can be found that do not change with reparametrization. First, we have to define what we mean by absence of information. From the Bayes theorem we see that all information coming from the data is contained in the likelihood, whereas the prior information is independent of the data. Prior ignorance should be expressed in a scale in which the likelihood will not change. Jeffreys (1961) first proposed this solution, and it can be found fully developed by Box and Tiao (1973). Because it is not always possible to find such scale, approximations may be required, and they work well with large samples (Box and Tiao say "moderate sample sizes" to avoid obvious criticism). However, when prior ignorance should be expressed about two parameters at the same time, such as mean and variance, these functions either do not exist or they lead to paradoxes. Thus, Jeffrey's priors have not been applied successfully in animal breeding, despite being currently used in other fields of scientific inference. This is most likely because animal breeding problems are multivariate in one way or another (we usually have environmental effects, permanent or litter effects, multitrait problems, etc.).

Many times the suggested prior functions are not proper probabilities (they do not integrate to 1). For example, a flat prior gives the same probability values from $-\infty$ to $+\infty$. These functions can lead to posterior densities that do not integrate to 1 and thus they are not probabilities, making the inference impossible. It is not always easy to detect an improper posterior density, particularly when they are estimated using MCMC methods such as Gibbs sampling. For example, Hoeschele and Tier (1995) found improper posteriors for herd-year-season effect in a threshold model when they used an improper flat prior for this effect. Due to this difficulty, Hobert and Casella (1996) recommend always using proper prior densities, unless the improper prior being used belongs to a class of improper priors that leads to proper posterior probabilities. However, in using improper priors the rather philosophical problem of using a description of prior uncertainty that is not a probability still persists. Moreover, it is difficult, from a theoretical point of view, to sustain that prior ignorance depends on the true value of an unknown population parameter.

With prior ignorance, or in checking the robustness of results by comparing them with the case in which there is no prior information, Bernardo (1979) proposed calculating posterior densities that give maximum weight to the data. The advantage of this solution is that the posterior density is directly calculated, ensuring that it is a proper probability density. For these authors:

> The problem of characterizing a "*non-informative*" or "*objective*" prior distribution, representing "*prior ignorance*", "*vague prior knowledge*" and "*letting the data speak for themselves*" is far more complex that the apparent intuitive immediacy of these words would suggest . . . "vague" is itself much too vague an idea to be useful. There is no "objective" prior that represents ignorance . . . the *reference prior* component of the analysis is simply a mathematical tool.
>
> Bernardo and Smith (1994)

The main problem with this solution is an operative one: all of these priors are difficult to calculate, particularly in the case of selected data, and they change with the model used. In the multivariate case, there is also a distinction between the parameter of interest and nuisance parameters. Depending on the order in which the nuisance parameters are disposed for conditionalization, the reference prior obtained for the parameter of interest can be different (e.g., Robert, 1992). This is somewhat puzzling, because it means that the "minimum amount of information" changes with the order in which the analysis is conducted. Of course, if it is admitted that priors are "simply mathematical tools," then the problem vanishes.

What can be done when there is no prior information? The easiest solution is to use several priors and observe any effect on the results. For example, in estimating the heritability of a new trait, use of vague prior information around several values of h^2 may result in different posterior distributions. If the results are similar regardless of the prior distribution used, the information in the data has overwhelmed the prior information and the lack of prior information is of no consequence. Although the posterior distribution was obtained under the false hypothesis of the existence of prior information, this hypothesis would be irrelevant to the results obtained.

BLUP and Selection Indexes as Bayesian Estimators

Ronningen (1971) and Dempfle (1977) remarked that BLUP can be considered as a Bayesian estimator. Giving the linear model [1]

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

where $\mathbf{t}' = [\mathbf{b}'\ \mathbf{u}']$ and $\mathbf{W} = [\mathbf{X}\ \mathbf{Z}]$, the Bayesian objective consists of finding the function $f(\mathbf{u}|\mathbf{y})$ to be able to make inferences. In a Bayesian context there are no fixed effects. Thus, selection indexes are still a particular case without the **b** effects. The objective is to estimate the vector **t** from the data **y**. First, determine $f(\mathbf{t}|\mathbf{y})$. Applying Bayes theorem [3] results in

$$f(\mathbf{t}|\mathbf{y}) = \text{constant} \cdot f(\mathbf{y}|\mathbf{t})\ f(\mathbf{t})$$

If $f(\mathbf{y}|\mathbf{t}) = N(\mathbf{Wt}, \mathbf{I}\sigma_e^2)$, then

$$f(\mathbf{y}|\mathbf{t}) \propto \exp[(\mathbf{y} - \mathbf{Wt})'\ (\mathbf{y} - \mathbf{Wt})] \qquad [4]$$

Suppose that the prior density of **t** is also normal. This may be reasonable for the genetic values, but it may not be reasonable for the environmental values. Then,

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

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

where **m** and **S** are the prior mean and variance of the environmental effects. Also suppose that environmental and genetic values are uncorrelated, although this is not always true. For example, in dairy cattle it is well known that the "best" farms tend to buy the "best" semen. Avoiding this correlation between environmental and genetic effects was one of the reasons Henderson (1973) proposed considering farm effects as fixed. With small farms the herd effects are not well estimated, which has consequences for the estimation of genetic effects and creates a problem with no easy solution.

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

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

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

There are three Bayesian estimators that can be taken from the posterior distribution $f(\mathbf{t}|\mathbf{y})$: mean, median, and mode. Because this is a normal distribution, all of them are coincident. We will calculate the mode, which is the maximum of the distribution. Deriving with respect to **t** and equating to zero, we obtain the equation

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

or, in expanded form

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

These are similar to the mixed-model equations [2], but here we include the mean of the environmental effects $\mathbf{m}_b$, which is not null, and the matrix $\mathbf{S}^{-1}$, which for the environmental effects is equivalent to the $\mathbf{G}^{-1}$ matrix. If we apply the *principle of indifference* and the fixed effects can have any value in the interval $]-\infty, +\infty[$, their variance *a priori* tends to infinity and consequently $\mathbf{S}^{-1}$ tends to zero. If $\mathbf{S}^{-1}$ is null, we obtain the mixed-model equations [2]. Therefore, BLUP is a Bayesian estimator (mode, median, and mean of a normal posterior distribution) constructed using flat priors for the environmental effects and a normal prior with zero mean and variance **G** for the genetic effects. This does not mean that BLUP is an optimal Bayesian estimator. It is not reasonable to suppose that the environmental effects can take any value with the same probability. Usually, prior information is available, at least for the maximum and minimum values that an environmental effect can have could be expressed by defining a prior normal distribution. The use of flat priors has its only justification as reference priors and because it facilitates calculation of the posterior density.

ML and REML as Bayesian Estimators

If any value of h^2 has the same prior probability[7], $f(h^2) = \text{constant}$ (which is only admissible in the interval [0,1], then Eq. [3] becomes

$$f(h^2|\mathbf{y}) \propto f(\mathbf{y}|h^2)$$

and thus the posterior distribution is calculated directly from the likelihood. This distribution is usually easy to calculate. Often the data are thought to be distributed either normally or according to another known function. Notice that, although using the same function, the Bayesian inference is based on a completely different principle from ML. Here, the likelihood is used because it is proportional to $f(h^2|\mathbf{y})$, which is a distribution of probabilities, whereas frequentist inferences based on the likelihood result from the good properties that its maximum has in conceptually infinite repetitions of the experiment.

Harville (1974) was the first to realize that REML could be considered a Bayesian estimator. Taking the mixed model [1], Bayesian inferences are made from the complete posterior density $f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e, |\mathbf{y})$. The ML estimators of **b**, σ^2_u and σ^2_e are coincident with the mode of the marginal posterior density

$$f(\mathbf{b}, \sigma^2_u, \sigma^2_e|\mathbf{y}) = \int f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e|\mathbf{y})\ d\mathbf{u}$$

whereas the REML estimators of σ^2_u and σ^2_e are coincident with the mode of the marginal posterior density

$$f(\sigma^2_u, \sigma^2_e|\mathbf{y}) = \iint f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e|\mathbf{y})\ d\mathbf{b}\ d\mathbf{u}$$

when prior values are assumed to be flat for **b** and normal for **u**, as in the case of BLUP. Note that there is no REML estimation of the "fixed" effects **b**, because it is made in a space free of these effects.

The mode of the multivariant distribution $f(\mathbf{b}, \sigma^2_u, \sigma^2_e|\mathbf{y})$ should not be the same, in general, as the mode of the bivariant distribution $f(\sigma^2_u, \sigma^2_e|\mathbf{y})$. Integrating out the environmental effects means considering all possible values for **b**, weighting them by their probabilities, and summing all these possible weighed values. This is a Bayesian justification for using REML instead of ML, and it is clearer than the frequentist one. Also notice

[7]Strictly speaking, all the intervals of the same length have the same probability; the probability of a concrete value is zero, because there are infinite points in the interval [0, 1]. The same can be said for the flat priors of the previous section.

that from a Bayesian point of view the weighing argument can be continued. Thus, it can be proposed to estimate the variance components from the mode of the marginal posterior densities

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

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

This is what Gianola and Foulley (1990) propose. Originally, their reason to take the mode was a practical one: it could be calculated by finding the maximum of the marginal posterior densities. Now, the modern MCMC techniques (Gibbs sampling, Metropolis-Hastings, etc.) make finding the mode unnecessary because all of the posterior density can be inferred.

Bayesian Estimators of Genetic Values and Dispersion Parameters

Independent of any Bayesian interpretation of the methods used in animal breeding that have been derived from a frequentist perspective, there is a Bayesian way of estimating breeding values and genetic parameters. We will maintain the notation of the mixed model [1], although there are not fixed effects in a Bayesian context. Thus, here **b** is a random vector of the environmental effects. The most general problem is to find the posterior density $f(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma_e^2|\mathbf{y})$ and to make inferences from it or from the marginal densities $f(\mathbf{b}|\mathbf{y})$, $f(\mathbf{u}|\mathbf{y})$, $f(\sigma_u^2|\mathbf{y})$, and $f(\sigma_e^2|\mathbf{y})$. The likelihood and the prior distributions are needed.

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

It is usually assumed that priors for dispersion parameters are mutually independent between them and independent of priors for genetic and environmental values. Moreover, because we do not know the joint prior distribution $f(\mathbf{b}, \mathbf{u})$, we will assume that priors for **b** and **u** are independent. Thus,

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

In some cases **b** and **u** are not independent *a priori*, as in the example of the "best" dairy farms buying the "best" semen to generate replacements for their cows. However, $f(\mathbf{b})$ can be assumed constant to override the problem, as in the Bayesian interpretation of BLUP and REML, with similar advantages and disadvantages. If **b** and **u** are assumed independent, it seems appropriate to give informative prior values for **b**. This will often lead to a symmetric function (e.g., a normal distribution). In the case of very vague prior opinions, a flat prior may be used within the limits of an interval that fixes the maximum and minimum prior values for the effect. However, our prior opinion about the genetic determination of a trait is usually based on our prior opinion about the heritability of the trait and is affected by σ_u^2 and σ_e^2 at the same time. Thus, it is not the best solution to suppose the independence of both variance components. One way of facilitating the expression of this prior opinion is to reparameterize the prior density and to express it as a function of heritability, as Sorensen (1999) proposed. Nevertheless, in an experiment it is expected that prior opinions should be vague and most of the information should come from the data. If prior opinions about the parameters of interest are very sharp then there is no reason to perform the experiment.

In the past, it was important to find prior functions that "conjugate" with the likelihood, in order to find known functions and simplify the problem of deriving inferences from a posterior density. The problem was that if the posterior density was not a function of some type that had been studied before, then it was complicated to derive inferences. For example, if the posterior density was normal, about 95% of its elements were known to be within approximately 2 SD of its mean. However, if the posterior density was not a function of some known type then it was difficult to find probability areas, because it was necessary to integrate the function within limits. Another difficulty was calculation of $f(\mathbf{y})$, for which multiple integrals should be taken. The Bayesian procedures were limited to the use of asymptotic properties, and calculating the mode of the distributions. Nowadays, the use of MCMC simulation allows deriving inferences without regard for the functional form of the posterior density. Nevertheless, conjugate functions are used because they facilitate the numeric task of Gibbs sampling (i.e., sampling from conditional functions). Thus, inverted chi-square functions are used for variance component priors instead of using the reparameterization that allows expressing prior opinions of heritabilities. Inverted chi-square functions are used not only for mathematical commodity, but also because these functions depend on two parameters that allow expression of almost any kind of prior opinion (see, for example, the graphs of Lee, 1997). Animal breeders tend to call these parameters "degree of belief ν" and "dispersion parameter S," thinking the first leads to drawing more or less sharp densities and the second controls dispersion of the density prior, but this is incorrect. The shape of the function depends on both, and a prior inverted chi-square can be made sharp by changing S. Moreover, ν and S do not have to be natural numbers, because they do not represent degrees of freedom. They should be taken only as two parameters that describe the shape of a prior belief, and nothing else. The values of these parameters should not be established by appealing to external "objective" arguments, such as the degrees of freedom of the sampling distribution of the variance or other reasons not related to the real prior opinion of the researcher. These values are completely arbitrary insofar as they lead to a prior function that describes a prior state of opinion about the variance.

Bayesian estimators of the breeding values based on the marginal density $f(\mathbf{u}|\mathbf{y})$ take into account the uncertainty about the variance components and the uncertainty about environmental effects. All possible values of variance components and environmental effects are weighted according to their probability and considered in the estimation of the breeding values, because

$$f(\mathbf{u}|\mathbf{y}) = \iiint f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e|\mathbf{y})\, d\mathbf{b}\, d\sigma^2_u d\sigma^2_e$$

This is particularly useful for traits that are difficult or expensive to measure (for example, meat quality traits or traits that require surgical interventions), because in these cases the database may not be large enough to estimate the variance components with precision and the literature may be scarce. It is also useful when the database is large and variance components are estimated using subsets of it.

Bayesian estimators of variance components based on the marginal densities of the variance components have the advantage of taking into account the uncertainty associated with the other parameters. When they are estimated with MCMC techniques, there is a supplementary advantage: the posterior density of the heritability and genetic correlations $f(h^2|\mathbf{y})$ and $f(r_g|\mathbf{y})$ can be calculated immediately (see Appendix I).

Discussion

Bayesian Methods and Scientific Inference

Until recently, it was thought that the main difference between the two schools of inference lay in the use of prior information. The frequentist school produces inferences based on the data and prior knowledge of the distribution of estimators in the sampling space. The problem of the frequentist school exists in the use of probabilities without using prior information, as their founders wanted (see Pearson, 1966). In this case, the distribution of the estimator is used for inferences, instead of the distribution of the parameter, which leads to a rather unnatural form of expressing uncertainty about the results of an experiment. Fisher was perfectly conscious of the loss of clarity that this method implied, and wrote:

> Bayes perceived the fundamental importance of this problem and framed an axiom, which, if its truth were granted, would suffice to bring this large class of inductive inferences within the domain of the theory probability; so that, after a sample had been observed, statements about the population could be made, uncertain inferences, indeed, but having the well-defined type of uncertainty characteristic of statements of probability
>
> Fisher (1936)

The Bayesian school makes inferences from probabilities associated with values of the parameter of interest, which is a more natural way of expressing uncertainty. The main problem of this school is to be sure that these probabilities are indeed probabilities, because the prior information used to derive them can be nonexistent or difficult to quantify precisely.

Frequentist and Bayesian methods produce different answers to the problem of induction[8]. Basically, the principle of induction consists of admitting that it is possible to make inferences about the population parameters from samples. The problem of induction appears when this possibility is denied, even with some degree of probability (Popper and Miller, 1983). The solution proposed by Hume (1738) was psychological: when I observe the repetition of an event (for example, Friesian cows produce more than Jersey cows), I expect that in the future the behavior of the event would be the same. This is an unsatisfactory solution. From the early attempts of Kant (1781) until the more modern results of Russell (1948) and Popper (1972) there has been a search for a rational justification for believing causation can be associated with repeated phenomena and general conclusions can be derived from samples. Russell (1948) proposes five postulates that, if admitted (and this is arbitrary), would justify inductive inference. Popper (1972) simply helps in a better demarcation of the problem without bringing anything new to bear. It is currently admitted that the induction problem does not have a logical-deductive solution, and research is focused on examining the "good reasons" for believing in the induction principle, and the nature of these reasons (Bird, 1998). Frequentist statisticians, such as Neyman and Pearson, preferred to speak about "the scientist behavior" instead of induction. However, Fisher believed that we could associate probabilities with events, and he and Sir Harold Jeffreys remarked on the principal differences between both schools facing this problem (see Lane, 1980, for a summary of the discussion). The main problem of the frequentist solution is that inference is associated with distributional assumptions and often with unknown parameters of the population. The main problem of the Bayesian school is that probabilities often reflect states of belief instead of "objective" probabilities.

Nevertheless, a good way of understanding the induction problem is to look at the Bayes theorem. It is necessary, in order to associate a probability with an event,

[8]The problem of induction was presented in its modern form by the Scottish philosopher Hume, although as Copleston (1985) remarks the essence of the problem (the difficulty of associating effects to causes) was outlined at least from the second century. Kant looked for a rational justification of the principle in the prior knowledge. But the prior knowledge of Kant only has its name in common with the Bayesian prior knowledge. Kant says: "It is one of the first and most important questions to know whether there is some knowledge of experience. Call this knowledge 'a priori' and make distinction from the empiric knowledge in that the sources of this last one are 'a posteriori', based in the experience" (Kant, 1983). Of course, no Bayesian scientist would use priori knowledge as something not based in previous experiences. Although there is a Bayesian statistician that quotes Kant (Robert, 1992), I do not find any relationship between Kant's philosophy and the Bayesian paradigm.

to know its *a priori* probability. For example, to know how Spanish Landrace performs for litter size, a sample of 20 sows can be evaluated for litter size in their first parity. Say an average of five piglets born is observed. This seems a very unlikely outcome *a priori*. Because no external *a priori* probability is available, previous beliefs are used to express this *a priori* knowledge. These beliefs are not arbitrary, but are based on knowledge of swine performance and the performance of Landrace in other countries. Notice that these beliefs can be easily shared by many other colleagues, and they should not represent any problem in the analysis, insofar as they are vague (if they are not, there is no reason to perform the experiment). Frequentist scientists also use subjective beliefs in discussing results and comparing them with other published results. If a frequentist scientist finds a very unlikely value for a parameter then the tone of the discussion will cast doubt about its validity, if it is published at all. Frequentists and Bayesians do not differ in the use of subjective prior information, but in how they use the information.

The prior density is constructed from a given set of assumptions (for example, that previous experiments are reliable, the samples are taken at random, the sampling space does not contain irregularities, etc.). Because the truthfulness of these assumptions cannot be known (and this is the heart of the induction problem), strictly speaking, inferences about the population parameters cannot be made. In the future, it may be discovered that the samples were not taken at random or that the sampling space was bimodal and only one region of it was sampled. In this case, the process should be restarted and inferences made based on the new state of knowledge. The induction problem does not make science impossible; theories or alternatives can still be evaluated based on reasonable sets of assumptions.

A different matter, which places Bayesians in an uncomfortable situation, is the impossibility of fixing these previous beliefs in the multivariate case. Due to this difficulty, the modern Bayesian tends to ignore prior information, considering the prior density as a mathematical artifact that allows him to make inferences based on probabilities. In a classic Bayesian paper, Lindley and Smith (1972) argued that "it is typically true that there is available prior information about the parameters and that this may be exploited to find improved, and sometimes substantially improved, estimates [with respect to the least squares ones]."

Compare this with Bernardo and Smith (1994), quoted at the end of the paragraph on prior information. The problem in the multivariate case is that, because even flat priors are somewhat informative, how they affect the posterior distributions should be ascertained. However, subjective priors cannot be prepared as in the univariate case. One way of dealing with this difficulty is to try several priors constructed under different hypotheses (for example, independence of the parameters, dependence described from values taken from the literature, flat priors) and check whether the results are robust with respect to the change of prior information.

Bayesian Methods Applied

Until recently, the main criticism made to Bayesian methods was their practical sterility: they did not produce results. The methodology implies the resolution of complicated multidimensional integrals or the use of more or less accurate approximations. The recent use of MCMC techniques has solved most of these problems, although it has generated new ones. Most of these new problems are related to the convergence of these chains. Fortunately, these new problems are easier to handle, particularly when the distribution of the data is normal. These MCMC techniques provide random samples of the joint and marginal posterior distributions, and thus the mean, standard deviation, and confidence intervals can be directly estimated from the samples without the need of integration. New variables such as ratios, squares, and so on can be derived from the components of the joint posterior distribution samples obtained, and confidence intervals can be derived from these distributions without the need of any approximation.

There are many cases in animal breeding in which the frequentist approach gives an accurate and rapid answer, and Bayesian methods are not needed (for example, the estimation of breeding values by BLUP or indexes). The main problems of the frequentist techniques when they are applied to animal breeding are of two sorts. One type is derived from the use of large databases, and the other one is related to the difficulties in obtaining accurate estimates of the standard errors in many complex situations. An example of the first type is the difficulties with obtaining multivariate REML estimation of the variance components when the database is large (e.g., in dairy cattle). An example of the second type is the problem of taking into account the error of estimation of variance components in the prediction of breeding values. This is a major problem when the database is not large and references about variance components cannot be found in the literature, as often happens in meat quality traits or litter size components, for example. For large databases, MCMC techniques permit drawing posterior distributions of genetic parameters that may be easier to compute than the corresponding REML solutions. The Bayesian approach also has advantages in some complex situations.

The Bayesian way of attacking an estimation problem is always the same: to derive the posterior distribution of the parameters given the data. After calculating the likelihood and determining the prior distributions, MCMC techniques can be used to obtain samples of the marginal posterior distributions. New problems can be solved, in principle, just by defining them correctly. This is in some ways similar to ML techniques. Although many genetic programs have a massive amount of data, there are still experiments in which the database is limited by the cost or the available facilities.

Even when the amount of data is substantial, it can happen that some levels of fixed effects contain only a small amount of information. An example of large data sets with lack of information is the case of the heterogeneity of variance with respect to herds in dairy cattle. The model may incorrectly suppose the same additive and residual variances for all herds. However, particularly when herds are managed in many different environments, there may not be enough information to estimate the genetic variance within each herd (see Gianola et al., 1992, for a proposal of Bayesian estimation of heterogeneous variances in dairy cattle). Maximum likelihood methods have good frequentist properties, but only asymptotically. When the data are scarce it is unknown whether these good properties are conserved. Posterior densities are more precise in these cases for describing the state of ignorance, and by examining different priors it is possible to determine how much the results are affected by the lack of information. Moreover, with few data, likelihoods are rather flat, making it difficult to find the absolute maximum, and a local maximum may be found instead. Posterior marginal densities take into account this imprecision and have the advantage of being univariate functions. Samples of these functions can be found by MCMC techniques and their representation better illustrates the accuracy of the estimation. Finally, when the amount of data is small, REML estimates of dispersion parameters are not reliable. This may be a serious problem, because breeding values depend on them. Marginal posterior densities weight the state of ignorance about variance components when breeding values are estimated. If some estimates of variance components can be found in the literature, the frequentist way of estimating breeding values implies the use of a single value based on them (the average of them, for example). The Bayesian way weights every possible value by its posterior probability (e.g., see Blasco et al. [1998] for a prior weighting of estimates of variance components based on the literature and Piles et al. [2000b] for a comparison of meat quality traits between a selected and a control line).

Standard errors of the heritability, genetic correlation, or other functions of the variance components are estimated after Taylor series approximations (see, for example, Bulmer, 1985). Using modern MCMC techniques it is not necessary to use any approximation to draw the marginal posterior distribution of a function of the variance components or of other estimated parameters (an example is given in Appendix I). On other occasions, for example in nonlinear problems, the sampling distribution of the parameters is not known. Here again, MCMC techniques permit drawing a marginal posterior distribution of the parameters of interest, expressing the uncertainty about these parameters (e.g., see Mignon-Gastreau et al. [2000] and Piles et al. [2000a] for examples with nonlinear growth curves).

When two lactation curves or two growth curves are to be compared, the frequentist may use transformations or approximations (for example, to work in logarithmic scale or to approximate to linear functions by using Taylor series). In order to apply these transformations, more hypotheses are needed; for example, if a logarithmic scale is used, the errors are assumed multiplicative instead of additive. Tests of hypotheses to compare curve parameters become difficult to apply and are often based on approximated likelihoods used for likelihood ratios. Moreover, the standard error of a parameter in the original scale cannot be properly obtained from the standard error calculated in the transformed scale. Bayesian solutions for these problems are straightforward and have been developed by Varona et al. (1997) in a general form, Varona et al. (1998) for lactation curves, and Piles et al. (2000a,b) for growth curves. Gianola et al. (1999) derived a general solution for nonlinear fitting of longitudinal data when data are selected for other traits (for example, growth curves in populations selected for growth rate, or lactation curves in populations selected for milk, fat, and protein). They always consist, again, of deriving the posterior distribution of the parameters given the data and no transformations or linear approximations are then needed.

When a parameter is needed for estimating another one, for example, when variance components are needed to estimate breeding values, Bayesian techniques allow consideration of the error of estimation of the nested parameter (the variance, in this case). Inferences are made on the marginal posterior distributions, after integrating out variance components (Sorensen et al., 1994), that is, after having considered all possible values of variance components, weighted by their probability. Frequentist techniques are usually based on tests of robustness, which are complicated to perform in multivariate cases. There are other cases in which we can find more levels of hierarchy. For example, in nonlinear growth or lactation curves, the parameters of the curve may be affected by environmental and genetic effects. Here there are three levels of hierarchy: variance components are needed to estimate the effects and the effects are needed to estimate the growth curve parameters (see Gianola et al., 1999; Varona et al., 1998; and Piles et al., 2000a,b).

There is a technique called *data augmentation* that permits simplifying the computation for the models used for inferences. For example, when we have censored data, it would be easier to use a complete normal distribution rather than a truncated one. In multivariate analyses with different design matrices (for example, growth traits and reproductive traits), it would be easier to solve a case in which all traits have the same matrix designs. In these cases, the technique consists of *augmenting* the data vector by generating data to fill the gaps. If we call $\mathbf{z}$ the vector of generated data, the inferences on the heritability (for example) are made using MCMC techniques to draw the posterior distribution $P(h^2, \mathbf{z}|\mathbf{y})$, because $P(h^2, \mathbf{z}|\mathbf{y}) \propto P(\mathbf{z}, \mathbf{y}|h^2)\, P(h^2)$ and it is easier to use MCMC techniques (to derive conditionals, as it is explained in Appendix I) with the com-

plete set of data (**z**, **y**). After obtaining samples for $P(h^2, \mathbf{z}|\mathbf{y})$, just ignore the samples generated for **z** and consider only the samples for h^2, obtaining a draw from the marginal posterior distribution $P(h^2|\mathbf{y})$, as shown in Appendix I (notice the similarity to the EM-algorithm used for ML estimations). A detailed account of this procedure can be found in Tanner (1995). It has been used in a variety of problems, for example, for threshold models (Sorensen et al., 1995), or in QTL mapping (Satagopan et al., 1996).

A common need in QTL detection is to choose between models with no QTL, one QTL, or two or more QTL, and so forth. All the advantages of the Bayesian approach for model choice hold here. Scheler et al. (1998) have compared the frequentist and Bayesian approaches in closely related models and find the Bayesian estimates to be more accurate. Because Bayes factors are not easy to calculate, alternative MCMC techniques (such as "reversible jump") for model choice have been proposed. Uimari and Hoeschele (1997) compared several approaches, and Sillanpaa and Arjas (1998, 1999) used reversible jump for mapping multiple QTL in a model in which the number of QTL is an unobserved random variable. New research is done in the field to overcome the numerical difficulties. A tutorial about reversible jump can be found in Waagepetersen and Sorensen (2000).

Threshold traits are an example of three levels of hierarchy, where the thresholds depend on fixed effects and breeding values, and these effects depend on the variance components to be estimated. The first work done using Bayesian techniques in animal breeding was just for threshold traits, by Gianola and Foulley (1982), proposing to calculate the mode of a joint posterior density. Now, with MCMC techniques, marginal posterior distributions can be drawn. Bayesian techniques also allow the joint analysis of threshold and continuous traits (Janss and Foulley, 1993). A detailed description of how to arrive at the marginal posterior distributions using MCMC techniques can be found in Sorensen et al. (1995) and Sorensen (1999).

Survival analyses need the use of complex proportional hazard models (Weibull and Cox models). These models can be extended to include genetic effects. Mixed survival models are called "frailty" models. There are frequentist approaches to these models using the EM algorithm, and also MCMC techniques have been derived to estimate posterior distributions, but they are computationally difficult. A Bayesian solution proposed by Ducroq and Casella (1996) using an approximate method common in Bayesian analyses (Laplacian integration) results in a more straightforward solution.

Consider one of the main problems of dairy cattle genetic evaluation: the undeclared preferential treatment. Some farmers provide more feed, for example, to the cows obtained from good sires, but they do not declare this. Because the normal distribution has thin tails (i.e., individuals that are placed one, two, or three standard deviations greater than the mean become much more rare than in the case of other distributions), this preferential treatment can have a substantial effect on the evaluation of sires. Other distributions, such as Student's *t*, have thicker tails and this overvaluation has a much smaller effect; the analysis becomes more "robust." The "thickness" of the tail depends on the degrees of freedom, and it is a parameter to be estimated. Bayesian techniques permit handling problems related to robustness by using *t*-distributions (Stranden and Gianola, 1999). The inconvenience is that multiple *t*-distributions do not have the nice properties of the multinormal world (e.g., independence when the covariance is null, regression of one variable on another is no longer linear, etc.), and analyses become more complicated.

Conclusions

If the animal breeder is not interested in the philosophical problems associated with induction, but in tools to solve problems, both Bayesian and frequentist schools of inference are well established and it is not necessary to justify why one or the other school is preferred. Neither of them now has operational difficulties, with the exception of some complex cases. There is software available to analyze a large variety of problems from both points of view. To choose one school or the other should be related to whether there are solutions in one school that the other does not offer, to how easily the problems are solved, and to how comfortable the scientist feels with the particular way of expressing the results. Both schools present formal problems and paradoxes, although problems and paradoxes with methods of greater familiarity are often better tolerated. For example, a REML analysis with few data may be more easily accepted than a Bayesian analysis in which prior information affects the results. Statistical refinements should not lead the scientist to forget the necessity of having sufficient experimental information to achieve an appropriate level of precision in expression of the results. The anguish of the scientist for finding exact results instead of probable inferences is not new. We can find in the *Eglogae IX* of Virgilius, in an agricultural context, the exclamation "*Felix qui potuit rerum cognoscere causas*": Happy the man that can know the causes of the things!"

Implications

If the animal breeder is not interested in the philosophical problems associated with induction, but in tools to solve problems, both Bayesian and frequentist schools of inference are well established and it is not necessary to justify why one or the other school is preferred. Neither of them now has operational difficulties, with the exception of some complex cases. There is software available to analyze a large variety of problems from both points of view. To choose one school or the other should be related to whether there are solutions

in one school that the other does not offer, to how easily the problems are solved, and to how comfortable the scientist feels with the particular way of expressing the results.

Literature Cited

Barnett V. 1999. Comparative Statistical Inference. 3rd ed. John Wiley & Sons, Chichester, U.K.

Bernardo J. M. 1979. Reference posterior distributions for Bayesian inference. J. R. Stat. Soc. B 41:113–147.

Bernardo, J. M. 1997. Noninformative priors do not exist: A discussion. J. Stat. Planning Inf. 65:159–189.

Bernardo, J. M., and A. F. M. Smith. 1994. Bayesian theory. John Wiley & Sons, Chichester, U.K.

Bernouilli, D. 1778. Dijudicatio maxime probabilis plurium observatonium discrepantium atque verisimillia inductio inde formanda. Acta Acad. Scien. Imp. Petropolitanae. pp 3–23. Reprinted in translation with Kendall, 1961.

Bird, A. 1998. Philosophy of Science. UCL Press, London.

Blasco, A., D. Sorensen, and J. P. Bidanel. 1998. A Bayesian analysis of genetic parameters and selection response for litter size components in pigs. Genetics 149:301–306.

Box, G. E. P., and G. C. Tiao. 1973. Bayesian Inference in Statistical Analysis. John Wiley & Sons, New York.

Bulmer, M. G. 1985. The Mathematical Theory of Quantitative Genetics. Clarendon, Press, Oxford, U.K.

Cantet, R. J. C., R. L. Fernando, and D. Gianola. 1992. Bayesian inference about dispersion parameters of univariate mixed models with maternal effects: Theoretical considerations. Genet. Sel. Evol. 24:107–135.

Copleston, F. 1959. A History of Philosophy. Vol. V: Hobbes to Hume. Reprinted. p 286. Bantam Doubleday Dell Publishing, New York.

Dawid, D. 1976. Discussion of the paper of O. Bandorff-Nielsen "Plausibility inference." J. R. Stat. Soc. B 38:123–125.

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

Ducroq, V., and G. Casella. 1996. A Bayesian analysis of mixed survival models. Genet. Sel. Evol. 28:505–529.

Edwards, A. W. F. 1992. Likelihood. Cambridge University Press, Cambridge, U.K.

Efron, B. 1986. Why isn't everyone a Bayesian? Am. Stat. 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. 1936. Uncertain inference. Proc. Am. Acad. Arts Sci. 71:245–258.

Fisher, R. A. 1956. Statistical methods and scientific inference. Reprinted. pp 146–147. Oxford University Press, New York.

Gelman, A., J. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian Data Analysis. Chapman and Hall, London.

Gianola, D., and R. L. Fernando. 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63:217–244.

Gianola, D., and J. L. Foulley. 1982. Non linear prediction of latent genetic liability with binary expression: An empirical Bayes approach. In: Proc. 2nd World Congr. Genet. Appl. Livest. Prod., Madrid, Spain. 7:293–303.

Gianola, D., and J. L. Foulley. 1990. Variance estimation from integrated likelihoods. Genet. Sel. Evol. 22:403–417.

Gianola, D., J. L. Foulley, R. L. Fernando, C. R. Henderson, and K. A. Weigel. 1992. Estimation of heterogeneous variances using empirical Bayes methods: Theoretical considerations. J. Dairy Sci. 75:2805–2823.

Gianola, D., S. Im, and F. W. Macedo. 1990. A framework for prediction of breeding value. In: D. Gianola and K. Hammond (ed.) Statistical Methods for Genetic Improvement of Livestock. Springer-Verlag, Berlin, Germany.

Gianola, D., M. Piles, and A. Blasco. 1999. Bayesian inference about parameters of a longitudinal trajectory when selection operates on a correlated trait. In: Proc. Int. Symp. Anim. Breed. Genet. Univ. Federal de Vicosa, Brazil. pp 101–132.

Gianola, D., S. Rodriguez-Zas, and G. E. Shook. 1994. The Gibbs sampler in the animal model: A primer. In: J. L. Foulley and M. Molenat (ed.) Séminaire Modele Animal. INRA Departament de Genetique Animale, La Colle sur Loup, France. pp 47–56.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov Chain Monte Carlo in practice. Chapman & Hall, London.

Glymour, C. 1981. Why I am not a Bayesian. Reprinted in: D. Papineau (ed.) 1996. The Philosophy of Science. Oxford University Press, New York.

Hartley, H. O., and J. N. K. Rao. 1967. Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54:93–108.

Harville, D. 1974. Bayesian inference for variance components using only error contrasts. Biometrika 61:383–385.

Harville, D. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72:320–342.

Harville, D., and A. Carriquirry. 1992. Classical and Bayesian predictions as applied to an unbalanced mixed linear model. Biometrics 48:987–1003.

Hazel, L. N. 1943. The genetic basis for constructing selection indices. Genetics 38:476–490.

Henderson, C. R. 1949. Estimation of changes in herd environment. J. Dairy Sci. 32:706–715.

Henderson, C. R. 1950. Estimation of genetic parameters. Ann Math. Stat. 21:309–310.

Henderson, C. R. 1963. Selection index and expected genetic advance. In: W. D. Hanson and H. F. Robinson (ed.) Statistical Genetics and Plant Breeding. pp 141–153. National Academy of Sciences—National Research Council, Washington, DC.

Henderson, C. R. 1973. Sire evaluation and genetic trends. In: Proc. Anim. Breed. and Genet. Symp. in Honor of Dr. J. L. Lush. Am. Soc. Anim. Sci., Champaign, IL. pp 10–41.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Ontario, Canada.

Hobert, J. P., and G. Casella. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. J. Am. Stat. Assoc. 436:1461–1473.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. Stat. Sci. 14:382–417.

Hoeschele, I., and B. Tier. 1995. Estimation of variance components of threshold characters by marginal posterior modes and means via Gibbs sampling. Genet. Sel. Evol. 27:519–540.

Hume, D. 1738. Treatise of Human Nature. Reprinted. Orbis S.A., Madrid, Spain.

James, W., and C. Stein. 1961. Estimation with quadratic loss. In: Proc. 4th Berkeley Symp., University of California Press, Berkeley. pp 361–380.

Jans, L. L. G., and J. L. Foulley. 1993. Bivariate analysis for one continuous and one threshold dichotomous trait with unequal design matrices and an application to birth weight and calving difficulty. Livest. Prod. Sci. 33:183–198.

Jeffreys, H. 1931. Scientific Inference. Cambridge University Press, Cambridge, U.K.

Jeffreys, H. 1961. Theory of Probability. Oxford University Press, Oxford, U.K.

Kant, E. 1781. Critique of Pure Reason. p 98. Reprinted in translation. Orbis S.A., Madrid, Spain.

Kempthorne, O. 1984. Revisiting the past and anticipating the future. In: Statistics: An Appraisal. Proc. 50th Anniversary Iowa State Stat. Lab. The Iowa State University Press, Ames. pp 31–52.

Kendall, M. G. 1961. Daniel Bernoulli on maximum likelihood. Biometrika 48:1–8.

Keynes, J. M. 1921. A Treatise on Probability. Macmillan Publishing Co., London.

Lane, D. A. 1980. Fisher, Jeffreys, and the Nature of Probability. In: S. E. Fienberg and D. V. Hinkley (ed.) Lecture Notes in Statistics. pp 148–160. Springer-Verlag, Berlin, Germany.

Lavine, M., and J. Schervish. 1997. Bayes Factors: What they are and what they are not. Am. Stat.
Lee, P. M. 1997. Bayesian Statistics. pp 78–79. Arnold, London.
Lindley, D. V., and A. F. M. Smith. 1972. Bayes estimates for the linear model. J. R. Stat. Soc. B 34:1–41.
Mignon-Grasteau, S., M. Piles, L. Varona, H. Rochambeau, J. P. Poivey, A. Blasco, and C. Beaumont. 2000. Genetic analysis of growth curve parameters for male and female chickens resulting from selection based on juvenile and adult body weights simultaneously. J. Anim. Sci. 78:2515–2524.
Neyman, J., and E. Pearson. 1933. On the problem of the most efficient test of statistical hypotheses. Phil. Trans. R. Soc. 231A:289–337.
Patterson, H. D., and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. Biometrika 58:545–554.
Pearson, E. S. 1966. Some thoughts on statistical inference. In: The Selected Papers of E. S. Pearson. pp 276–283. Cambridge University Press, Cambridge, U.K.
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. R. Soc. 200A:1–66.
Piles, M., A. Blasco, D. Gianola, and L. Varona. 2000a. Bayesian inference about parameters of growth curves in rabbits when selection operates on growth rate. In: Proc. 51st Ann. Mtg. Europ. Assoc. Anim. Prod., The Hague, The Netherlands.
Piles, M., A. Blasco, and M. Pla. 2000b. The effect of selection for growth rate on carcass composition and meat characteristics of rabbit. Meat Sci. 54:347–355.
Popper, K. 1972. Objective Knowledge. Tecnos, Madrid, Spain.
Popper, K., and D. Miller. 1983. A proof of the impossibility of inductive probability. Nature (Lond.) 302:687–688.
Robinson, G. K. 1991. That BLUP is a good thing: The estimation of random effects. Stat. Sci. 6:15–51.
Robert, C. P. 1992. L'Analyse Statistique Bayesienne. pp 109–111, 336. Economica, Paris, France.
Robert, C. P., and G. Casella. 1999. Monte Carlo Statistical Methods. Springer, New York.
Ronningen, K. 1971. Some properties of the selection index derived by "Henderson's Mixed Model Method." Z. Tierz. Zuechtbiol. 83:186–193.
Russell, B. 1948. Human Knowledge: Its scope and limits. p 491. Orbis S.A., Madrid, Spain.
Satagopan, J. M., B. S. Yandell, M. A. Newton, and T. C. Osborn. 1996. A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. Genetics 144:805–816.
Scheler, P., B. Mangin, B. Goffinet, P. LeRoy, D. Boichard, and J. M. Elsen. 1998. Properties of a Bayesian approach to detect QTL compared to the flanking markers regression method. J. Anim. Breed. Genet. 115:87–95.
Searle, S. R., G. Casella, and C. E. Mcculloch. 1992. Variance Components. John Wiley & Sons, Chichester, U.K.
Sillanpaa, M. J., and E. Arjas. 1998. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373–1388.
Sillanpaa, M. J., and E. Arjas. 1999. Bayesian mapping of multiple quantitative trait loci from incomplete outbred line cross data. Genetics 151:1605–1619.
Smith, H. F. 1936. A discriminant function for plant selection. Ann. Eugen. 7:240–256.
Sorensen, D. 1999. Gibbs sampling in quantitative genetics. Natl. Inst. Anim. Sci. Internal Rep. No. 82. Tjele, Denmark.
Sorensen, D. A., C. S. Wang, J. Jensen, and D. Gianola. 1994. Bayesian analysis of genetic change due to selection using Gibbs sampling. Genet. Sel. Evol. 26:333–360.
Sorensen, D. A., S. Andersen, D. Gianola, and I. Korsgaard. 1995. Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. 27:229–249.
Stigler, S. M. 1983. Who discovered Bayes's theorem? Am. Stat. 37:290–296.
Stigler, S. M. 1986. The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press, Cambridge, MA.
Stranden, I. J., and D. Gianola. 1999. Mixed effects linear models with t-distributions for quantitative genetic analysis: A Bayesian approach. Genet. Sel. Evol. 31:25–42.
Stuart, A., and J. K. Ord. 1991. Kendall's Advanced Theory of Statistics vol. 2. Arnold, London.
Tanner, M. A., 1996. Tools for Statistical Inference. 3rd ed. Springer-Verlag, New York.
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. Ph.D. dissertation. University of Edinburgh, U.K.
Uimari, P., and I. Hoeschele. 1997. Mapping-linked quantitative trait loci using Bayesian analysis and Markov Chain Monte Carlo algorithms. Genetics 146:735–743.
Von Mises, R. 1928/1957. Probability, Statistics and Truth. Macmillan Publishing Co., London.
Varona, L., C. Moreno, L. A. Garcia-Cortes, and J. Altarriba. 1997. Multiple trait analysis of underlying biological variables of production functions. Livest. Prod. Sci. 47:201–209.
Varona, L., C. Moreno, L. A. García Cortés, and J. Altarriba. 1998. Bayesian analysis of the Wood's lactation curve for Spanish dairy cows. J. Dairy Sci. 81:1469–1478.
Waagepetersen, R., and D. Sorensen. 2000. A tutorial on Reversible Jump MCMC with a view toward applications in QTL mapping. Stat. Rev. (In press).
Wang, C. S., D. Gianola, D. A. Sorensen, J. Jensen, A. Christensen, and J. J. Rutledge. 1994. Response to selection for litter size in Danish Landrace Pigs: A Bayesian analysis. Theor. Appl. Genet. 88:220–230.
Woolliams, J. A., and T. H. E. Meuwissen. 1993. Decision rules and variance of response in breeding schemes. Anim. Prod. 56:179–186.
Yates, F. 1990. Foreword. In: R. A. Fisher (ed.) Statistical Methods, Experimental Design and Scientific Inference. Oxford University Press, Oxford, U.K.

Appendix I. Gibbs sampling

Consider the model [1]

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

The objective is to obtain random samples from the marginal posterior distributions of all unknowns; i.e., from

$$f(b_1|\mathbf{y}),\ f(b_2|\mathbf{y}),\ \ldots,\ f(u_1|\mathbf{y}),\ f(u_2|\mathbf{y}),\ \ldots,\ f(\sigma^2_u|\mathbf{y}),\ f(\sigma^2_e|\mathbf{y})$$

and also for any combination of unknowns, for example

$$f[\sigma^2_u/(\sigma^2_u + \sigma^2_e)|\ \mathbf{y}] = f(h^2|\mathbf{y})$$

These samples will be used for inferences. They can be used for drawing histograms that will give an approximate idea about the shape of the posterior distribution, and they can provide estimates of the mean, mode, and median of the distribution and can also be used for estimating confidence intervals (called "credibility intervals" by Bayesians). To obtain these samples, first obtain random samples of the complete posterior distribution $f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e|\mathbf{y})$. This will lead to a matrix in which each row will be one sample of each unknown, and each column a set of random sampled points of the marginal posterior distribution of each unknown.

$$\begin{bmatrix} b_{11} & b_{21} & \ldots & u_{11} & u_{21} & \ldots & \sigma^2_{ul} & \sigma^2_{el} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{1i} & b_{2i} & \ldots & u_{1i} & u_{2i} & \ldots & \sigma^2_{ui} & \sigma^2_{ei} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

Thus, the set of values (b_{11}, . . ., b_{1i}, . . .) is a random sample of the posterior distribution $f(b_1|\mathbf{y})$ and will be used to make inferences about b_1, and the same can be said about the other columns, the set (σ^2_{ul}, . . ., σ^2_{ui}, . . .) is a sample of the posterior distribution $f(\sigma^2_{ul}|\mathbf{y})$, and so on. To make inferences about the heritability, create a new column using σ^2_u and σ^2_e of each row. Then the set $[\sigma^2_{u1}/(\sigma^2_{u1}+\sigma^2_{e1}), \ldots, \sigma^2_{ui}/(\sigma^2_{ui}+\sigma^2_{ei}), \ldots]$ is a random sample of the posterior distribution $f(h^2|\mathbf{y})$.

It is not possible to directly obtain a random sample of the complete posterior distribution $f(\mathbf{b}, \mathbf{u}, \sigma^2_u, \sigma^2_e|\mathbf{y})$, because this is a high-dimensional function, the product of several multidimensional functions; for example, if we take normal priors for **b** and **u**, and inverted chi-square priors for the variance components, the posterior distribution is a multidimensional function, a product of normal and inverted chi-square distributions, and it is divided by a constant $f(\mathbf{y})$ that is virtually impossible to calculate. To reduce the dimensionality of the problem and make it possible to obtain random samples of the complete posterior distribution, Monte-Carlo Markov Chain (MCMC) techniques are used. The most used among them is called "Gibbs sampling" because originally it was used to sample the type of distributions called "Gibbs distributions." The procedure consists of sampling from conditional distributions in an iterative way, as follows. Write the conditional function

$$f(b_1|b_2, b_3, \ldots, b_i, \ldots, u_1, u_2, u_3, \ldots, u_i, \ldots, \sigma^2_u, \sigma^2_e, \mathbf{y}),$$

then give arbitrary values to b_2, b_3, . . ., b_i, . . ., u_1, u_2, u_3, . . ., u_i, . . ., σ^2_u, σ^2_e, and sample a random value for b_1 from this conditional distribution. To do this, the conditional distribution should be of a type for which sampling algorithms are available, otherwise other MCMC techniques (for example, Metropolis-Hastings) should be used to sample a random number from this conditional distribution. With the randomly sampled value for b_1, and the other former arbitrary values, take a random sample of b_2 from the conditional distribution

$$f(b_2|b_1, b_3, \ldots, b_i, \ldots, u_1, u_2, u_3, \ldots u_i, \ldots, \sigma^2_u, \sigma^2_e, \mathbf{y})$$

With the sampled values for b_1 and b_2, and the other former values, take a random sample from the conditional distribution of b_3, and proceed this way until a sample of each conditional distribution of each unknown has been taken. After this, begin again and repeat the cycle. After some cycles of iteration, the random sample numbers extracted from the conditionals are also random samples from the posterior distribution. These first iterations (the "burn in") are disregarded.

Consider a simple example in order to understand this intuitively: to obtain a posterior distribution of f(x, y) sampling from the conditionals f(x|y) and f(y|x). Take an arbitrary value for x, say x = 0. Figure 4 shows f(x,y) represented as lines of equal probability (as level curves in a map). Our arbitrary value x = 0 permits sampling a random number from f(y|x = 0), which is the y-axis. Because the probability between a and b is higher than in other parts of f(y|x = 0), the number sampled will be found between a and b more probably than in other parts of the conditional function. Suppose y = 2: sample on f(x|y = 2). The line y = 2 cuts f(x,y) and gives as an interval (c,d) in which we will find the next sampled extraction with more probability, thus x will probably be found between c and d: say, x = 4. Sample any form f(y|x = 4). This number will probably be between the interval (e,f). Observe, the tendency to sample from the highest areas of probability more often than from the lowest areas. At the beginning, x = 0 and y = 2 were points of the posterior distribution, but they were not random extractions, and thus we were not interested in them. However, after many iterations, we will find more samples in the highest areas of probability than in the lowest areas, and thus we find random samples from the posterior distribution. This explains why the first points sampled should be discarded, and only after some cycles of iteration are samples taken at random.

The main problems associated to this procedure are the following.

First, strictly speaking, it cannot be demonstrated that we are finally sampling from a posterior distribution. A Markov chain must be *irreducible* to converge to a posterior distribution. Although it can be demonstrated that some chains are not irreducible, there is no general procedure to ensure irreducibility.

Second, even in the case in which the chain is irreducible, it is not known when sampling from the posterior distribution begins. By using several chains a point is arrived at where the variability among chains may be attributed to Monte-Carlo sampling error, and thus support the belief that samples are being drawn from the posterior distribution. There are some tests to check whether this is the situation. Good practical textbooks are Gelman et al. (1995), Gilks et al. (1996), and Robert and Casella (1999).

Third, even having a irreducible chain and when the tests ensure convergence, the converged distribution may not be stationary. Sometimes there are large sequences of sampling that give the impression of stability, and after many iterations the chains move to another area of stability.

The above problems are not trivial, and they occupy a part of the research in MCMC methods. Practically speaking, what people do is to launch several chains with different starting values and to observe their behavior. No pathologies are expected for a large set of

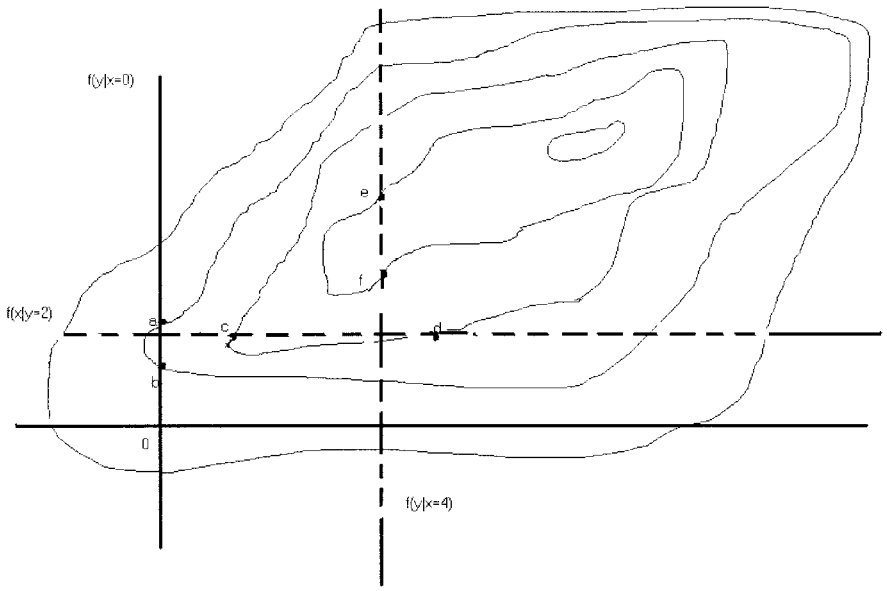

Figure 4. Posterior distribution and conditionals.

problems (for example, when using multivariate distributions), but some more complicated models (for example, threshold models with environmental effects in which no positives are observed in some level of one of the effects) should be examined with care.

It should be noted that these difficulties are similar to finding a global maximum in multivariate likelihood with several fixed and random effects. With a large database, second derivative algorithms cannot be used, and thus there is a formal incertitude about whether the maximum found is global or local. Here again, people use several starting values and examine the behavior of their results. Maximum likelihood methods present additional problems in estimation when the model becomes more complex. Thus, the difficulties of MCMC methods should not be considered a major problem in using Bayesian techniques, at least not any more difficult than the ones found when using advanced frequentist techniques, for which MCMC methods are sometimes also used.