The Bayesian Choice

2

> *If an event can be produced by a number of n different causes, the probabilities of theses causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given that cause, divided by the sum of all the probabilities of the event given each of these causes.*
>
> Pierre Simon, Marquis de Laplace, 1774

In this chapter, we examine the advantages of Bayesian inference using practical examples. We will see how to use Bayesian credibility intervals for inferences. We introduce new tools as the probability of relevance or the guaranteed value at a given probability. We will see the advantages of comparing treatments using ratios instead of differences. We will learn one of the main advantages of Bayesian procedures, the possibility of marginalisation. We also will see some misinterpretations of Bayesian theory and procedures.

2.1 Bayesian Inference

2.1.1 The Foundations of Bayesian Inference

Bayesian inference is based on the use of probability for expressing uncertainty. It seems more natural to express uncertainty stating that the probability of two treatments being different is of 98% than acting as if they were different hoping not to be wrong more than a 95% of times along our career. It looks more natural to find the most probable value of a parameter based on our data than finding out which value of this parameter would produce our data with highest probability if it were the true value. To examine the distribution of a combination of the data if the experiment was repeated many times is less attractive than examining the probability distribution of the parameter we want to estimate. This was recognised by the

A. Blasco, *Bayesian Data Analysis for Animal Scientists*,
DOI 10.1007/978-3-319-54274-4_2

founders of what we now call 'classical statistics' (see, for example, K. Pearson 1920; Fisher 1936; E. Pearson 1962). All of them preferred probability statements to express uncertainty, but they thought it was not possible to construct these statements. The reason being that we need some prior information for making probability statements based in our data, and it is not clear how to introduce this prior information in our analysis or how to express complete lack of information using probability statements. However, Bayesian statisticians claim they have found solutions for these problems, and they can indeed make probability statements about the parameters, making the Bayesian choice more attractive. All the controversy between both schools is centred in this point: whether the Bayesian solutions for prior information are valid or not. In this chapter, we will show why the Bayesian choice is more attractive by showing its possibilities for inference, and in the following chapters, we will see how to work with Bayesian statistics in practice. We will delay to Chap. 9 the discussion about prior information.

2.1.2 Bayes Theorem

Bayesian inference is based in what nowadays is known as 'Bayes theorem', a statement about probability universally accepted.

To see how it works, we need to define first conditional probability. Let us take a collective of men and women, British and Spanish. If we call:

A: to be man
B: to be British
N: total number of individuals
N_A: number of men
N_B: number of British people
N_{AB}: number of British men

The probability of being at the same time man and British is

$$P(A, B) = \frac{N_{AB}}{N}$$

However, if we take only British people, the probability of being a man is

$$P(A|B) = \frac{N_{AB}}{N_B}$$

This is read, 'The probability of being a man, given that this person is British, is the number of British men divided by the total number of British people'.

The relationship between both probabilities is easy to find:

$$P(A,B) = \frac{N_{AB}}{N} = \frac{N_{AB}}{N_B} \cdot \frac{N_B}{N} = P(A|B) \cdot P(B)$$

In general, if we express the probability of the joint occurrence of two events A and B, we have the same expression

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

where the bar '|' means 'given',[1] i.e. the probability of the event A is conditioned to the occurrence of event B. The probability of taking a train at 12:00 to London is the probability of arriving on time to the train station, given that there is a train to London at this time, multiplied by the probability of having indeed a train at this time.

In general, the probability of occurring two events is the probability of the first one given that the other one happened for sure, by the probability of this later event taking place; thus, we can apply this rule to both events, A and B

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

directly leading to Bayes theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Imagine we are interested in comparing how a new food affects growth rate in pigs; we have a control food and a specific food, and we call C and S the effect of these foods on growth rate, respectively. We are interested in knowing whether growth rate is higher or not with this specific food, i.e. whether $S > C$ or, in other terms, whether $S - C > 0$. This will be called 'event A', and the data of our random sample of pigs will be 'event B':

$$\begin{array}{ll} A : S - C > 0 & (\text{the unknown}) \\ B : \boldsymbol{y} & (\text{the data sample}) \end{array}$$

we have, applying Bayes theorem

$$P(S - C > 0|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|S - C > 0) \cdot P(S - C > 0)}{P(\boldsymbol{y})}$$

[1] Notice that 'l' is a vertical bar, different from '/'. The notation was introduced by Jeffreys (1931) to avoid confusion with the sign of division.

Thus, given our data, we can estimate the probability of the effect of the specific food in growth rate being higher than the effect of the control food. In order to make this inference, we need to know:

$P(\boldsymbol{y}|\, S - C > 0)$: This probability is based on the distribution of the data for a given value of the unknowns. This distribution is often known or assumed to be known from reasonable hypotheses. For example, most biological traits are originated from many causes each one having a small effect; thus, the central limit theorem says that data should be Normally distributed, which allows us to calculate this probability.

$P(\boldsymbol{y})$ is a constant, the probability of the sample. Our sample is an event, and it has a probability. Using MCMC techniques, we do not need to calculate it, as we will see in Chap. 4.

$P(S - C > 0)$ is a probability independent of any set of data. It is interpreted as the information that we have (before making the experiment) about the difference $S - C$ being positive. This *prior information* is needed to complete Bayes theorem and allow us to make a probability statement. We discuss below the different options we have for assessing this probability.

An advantage of Bayesian inference is that we have a rule for estimation in all sorts of circumstances. We know that all problems are reduced to a single pathway: we should look for a posterior distribution, given the distribution of the data and the prior distribution.

2.1.3 Prior Information

Prior information is the information about the parameters we want to estimate that exists before we perform our experiment. Normally, we are not the only people in the world working on a topic; other colleagues should have performed related experiments that would provide some prior information about our experiment. If so, it would be very interesting to blend this information with the information provided by our data. Scientists always consider prior information, even if they apply classical statistics; they compare their results with results provided by other colleagues, as we can see in the 'Discussion' section of all papers. Our conclusions are not only based in our work but also in results obtained in other publications; thus, a formal integration of all sources of information looks attractive. Unfortunately, it is almost impossible to perform this accurately, with some exceptions. We will distinguish three scenarios:

When we have exact prior information: In this case, we do not have any difficulty in integrating this prior information, as we will see with detail in Chap. 9. For example, suppose the colour of the skin in a line of mouse is determined by a single gene with two alleles (A,a). If a mouse receives the 'a' allele from both parents, its colour is brown (therefore, it is homozygous aa), but if it receives an allele 'A' from one of the parents, his colour is black (in this case it can either be

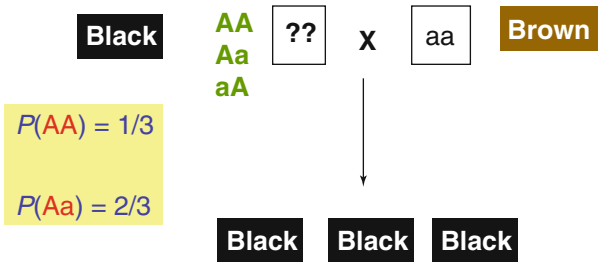

Fig. 2.1 Two heterozygous black mice have a black son, which may be homozygous or heterozygous. To test this, the son is crossed with a brown mouse, and their offspring is examined. Before performing the experiment, we have some prior information due to our knowledge of Mendel's law

homozygous AA or heterozygous Aa). We try to find out whether a black mouse, son of black heterozygous mates (Aa × Aa), is homozygous (AA) or heterozygous (Aa) (Fig. 2.1). In order to assess this, we mate this mouse with a brown (aa) mouse. If we obtain a brown son, we will be certain it is heterozygous, since it has passed an allele 'a' to the son, but if we obtain black offspring, there is still the doubt about whether our mouse is homozygous AA or heterozygous Aa. We perform the experiment, and we actually get three offspring of black mice. What is the probability, given this data, that our black mouse is heterozygous Aa?

Notice that *before* we perform the experiment, we have prior information due to our knowledge of Mendel's law of inheritance. We know that our mouse will receive an allele 'A' or 'a' from his father and an allele 'A' or 'a' from his mother, but it cannot receive an allele 'a' from each one at the same time, because in this case it would be brown. This means that we have only three possibilities: either it has received two alleles 'A', it received one 'A' from the father and an allele 'a' from the mother or an 'a' from the father and an 'A' from the mother. This means that the prior probability of our mouse to be heterozygous *before performing the experiment* is two-thirds, because there are two favourable possibilities in a total of three. Therefore, the probability of being homozygous AA is one-third because the sum of probabilities of all events should be 1.

We should blend this prior information with the information provided by the experiment, in our case having three black offspring when crossing this black mouse (AA or Aa) with a brown mouse (aa). We will do this in Chap. 9.

When we have vague prior information: In most cases, prior information is not as firmly established as in the example before. We have some experiments in the literature, but even if they look similar and they give their standard errors, we may not trust them or we may consider that their circumstances were only partially applicable to our case. However, they may provide information useful to us, but putting that aside, *we always need prior information in order to apply Bayes theorem*. This was not correctly perceived throughout the nineteenth century, and it was assumed that we could not integrate prior information properly. The first

solution to this problem was provided independently by the philosopher of mathematics Frank Ramsey (1931) and the Italian actuary Bruno de Finetti (1937).[2] Their solution was original but polemic. They sustained that probability, considered as ratios between events, was not sufficient to describe the use of probability we do. For example, if we say that it is probable that the Scottish nationalist party will win next elections, we are not calculating the ratio between favourable and total number of events. If we say that it is improbable that an earthquake will destroy Berlin in the next 10 years, we are not applying any ratio between events either. Both Ramsay and de Finetti proposed that probability describes beliefs. We assign a number between 0 and 1 to an event, according to our subjective evaluation of the event. This does not mean that our beliefs are arbitrary; if we are experts on a subject and we are performing an experiment, we hope to agree with our colleagues in the evaluation of previous experiments. This definition also includes other uses of probability, like the probability of obtaining 6 when throwing a dice. If we have enough data, our data will overcome the prior opinion we had, and our experiment would give approximately the same results independently on whether we used our prior opinion or the prior opinion of our colleagues. Notice that this prior belief is always based on previous data, not in unfounded guessing or in arbitrary statements.[3]

When this solution was proposed, some statisticians were scandalised by the thought of science becoming something subjective. Kempthorne expresses this point of view accurately:

> The controversy between the frequentist school and the Bayesian school of inference has huge implications ... 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."
>
> Oscar Kempthorne (1984)

Here Kempthorne confused *subjective* with *arbitrary*. As we said before, several experts can share the same subjective opinion.[4] In fields in which the expert opinion is used to make decisions, this did not represent any problem; however, in

[2]Ramsey and de Finetti formalised probability as a degree of belief, making possible to operate with it. For early discussions about probability as a degree of belief, see Howie (2002).

[3]A Bayesian statistician quotes Kant (Robert 1992, p. 336) to justify prior beliefs. Kant looked for a rational justification of the principle of induction 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 therefore at least a question requiring closer investigation, and one not to be dismissed at first glance, whether there is any such cognition independent of all experience and even of all impressions of the senses. One calls such cognitions a priori, and distinguishes them from empirical ones, which have their sources a posteriori, namely in experience' (Kant 1781, p. 136). Of course, no Bayesian scientist would use prior knowledge as something not based in previous experiences. Therefore, it seems that there is no relationship between Kant's philosophy and the Bayesian paradigm.

[4]One of the first defenders of probability as degree of belief, William Donkin, said that probability was 'not being relative to any individual mind; since, the same information being presupposed, all minds *ought* to distribute their belief in the same way' (Donkin 1851).

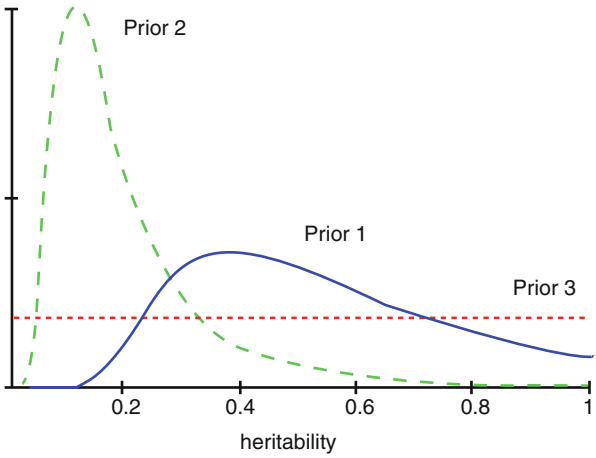

Fig. 2.2 Three different priors showing three different states of belief about the heritability of ovulation rate in French Landrace pigs (from Blasco et al. 1998)

biological sciences, it is preferred that results are based in current data more than in our prior beliefs. In this case, data should be enough to avoid dependence on prior beliefs. For example, Blasco et al. (1998) compared three different prior beliefs to estimate the heritability of ovulation rate of a pig population of French Landrace. According to literature, there is a large range of variation of this parameter, the average being of about 0.4. However, in a former experiment performed 11 years ago with the same population, the heritability was of 0.11. Then, vague states of beliefs were tested (Fig. 2.2).

The first state of beliefs considered that it was more probable that the true value of the heritability was around 0.4 and the second that it was more probable around 0.11. A third state of opinion was considered: all of the possible values would have the same probability. After performing the experiment, all analyses gave the same results approximately; thus, prior beliefs were irrelevant for the conclusions. In Chap. 5 we will see how to integrate vague beliefs in the analysis.

It can be argued that when the prior belief is very sharp with respect to the distribution of the data, it will dominate, and the results will reflect our prior belief instead of what our experiment brings. This is a correct criticism, but it is unrealistic, why should we perform an experiment if we have sharp prior beliefs? If we are sure about the value of the heritability, no experiment will change our beliefs. For example, it is known after many experiments that heritability of litter size in pigs has values between 0.05 and 0.15. This means that our prior belief around these values is very sharp, and if we perform an experiment, our results will be similar to our prior opinion, independently of the result of our experiment. However, the same will happen if we use classical statistics: if we find a heritability of 0.90 for litter size, we will not trust it, and then we will use our prior opinion to disregard our result and still believe that heritability of litter size is low. Thus, scientists using classical statistics *also* use prior beliefs, although in a different manner.

A main problem for defining subjective beliefs arises in the multivariate case. In this case, we cannot state a prior opinion because human minds cannot imagine distributions in more than three dimensions. We will deal with this problem in Chap. 9.

When we do not have any prior information. Describing ignorance: It is uncommon the lack of prior information, usually somebody else has worked before in the same subject or in a similar one. Nevertheless, even having prior information, it may be interesting to know what we will obtain ignoring prior information and basing our inferences only in our data. Unfortunately, it does not seem possible to describe ignorance using probability statements. Along the nineteenth century and the first three decades of the twentieth century, it was applied what Keynes (1921) named the *principle of indifference*, consisting in assuming that all events had the same prior probability, i.e. all possible values of the parameters to be estimated were equally probable before performing the experiment. These priors are called *flat priors* because of their shape; prior 3 of Fig. 2.2 is a flat prior. The problem is that ignorance is not the same as indifference (it is not the same to say 'I don't know' that 'I don't care'). Moreover, this principle leads to paradoxes, as we will see in Chap. 9. Other alternatives have been proposed: Jeffreys (1961) proposed priors that are invariant to transformations, and Bernardo (1979) proposed priors that have minimum information. All these priors are called 'objective' or 'noninformative' by Bayesian statisticians; however, it should be noted first that *all of them are informative* (although usually the information they provide is rather vague) and second that *the information they provide is not objective*, at least using the common meaning of this word.[5] Nevertheless, the principle of indifference is widely used because, introducing a very small amount of information, it does not affect the results unless we have very small samples. In Chap. 10, Sect. 10.3.2, we will find some support to the principle of indifference using information theory. We will examine the problem of representing ignorance in Chap. 9. Until then, in all forthcoming chapters, we will ask the reader to admit the use of flat priors, and we will use them in most examples.

2.1.4 Probability Density

When we make inferences about continuous variables, we have an infinite number of possible values; thus, we make probability statements as 'probability of having a response to selection higher than 10 g/day' or 'probability of differences in litter size being between 1.0 and 1.5' because the probability of a single point (like 1.4564587534... or 10.00000000...) is always zero. To make these statements, we need an auxiliary function f called 'probability density function'. As we will see in

[5]Geneticists can find this nomenclature particularly annoying; because due to their knowledge of Mendel's laws, they have real objective priors. Thus, when they are using prior knowledge of relationships between relatives, they are not using subjective priors at all.

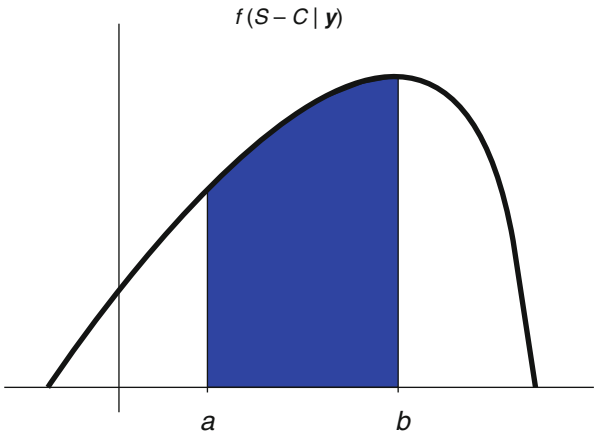

Fig. 2.3 Probability density function of the difference $S - C$ given the data. In *blue*, $P(a \leq S - C \leq b)$

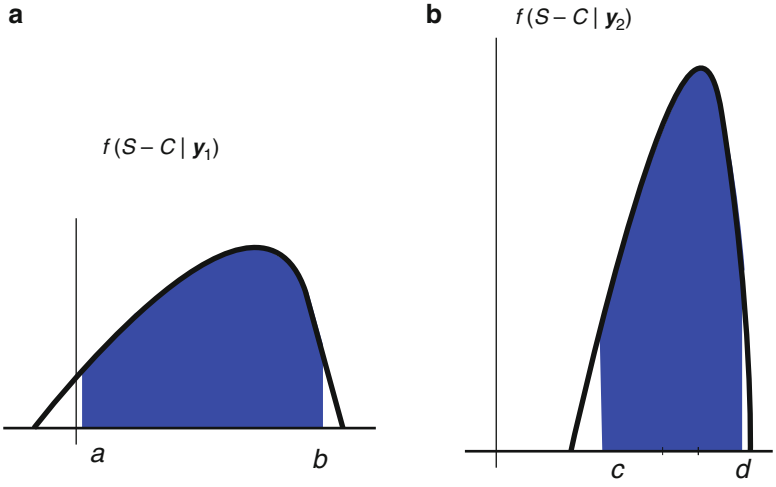

Fig. 2.4 (**a**) Probability density of the difference $S - C$ given the set of data $\boldsymbol{y}_1$. (**b**) Probability density of the difference $S - C$ given the set of data $\boldsymbol{y}_2$. In *blue*, $P(a \leq S - C \leq b) = 0.95$ in both cases

Chap. 3, the probability of $S - C$ being between the values a and b, given our data, is the area enclosed by the probability density function between both values (Fig. 2.3).

It is important to notice that we make our inferences, 'given our data', which means that if we have other data set, our inferences can change (Fig. 2.4). For example, if we have more data, the density function will be sharper. In this case, we

can make more accurate inferences about the true value $S - C$; the interval $[c, d]$ of Fig. 2.4 has the same probability as the interval $[a, b]$, but it is shorter, and the inference is more precise. Saying 'the difference in growth rate between specific and control feed is between 100 and 200 g/day' (Fig. 2.4b) is more precise than saying that this difference is between 50 and 300 g/day (Fig. 2.4a).

In a Bayesian context, our inferences depend on our unique sample, not in how samples would be distributed if we had a hypothetical large number of them. If we collected more samples, we would put all of them together and make inferences conditioned to our new, larger data set. From now, we will not consider this point again; thus, we will always show inferences 'given our sample'.

The probability density functions conditioned to the data are called *posterior distributions*. The functions of Fig. 2.2 are also probability density functions, but in this case, they are not conditioned to the data because they express our prior uncertainty about the parameters to be estimated before doing the experiment, they are *prior distributions*. In Chap. 3 we will learn how to formally use these density functions.

2.2 Features of Bayesian Inference

2.2.1 Point Estimates: Mean, Median and Mode

All information is contained in the posterior probability density function $f(S - C|\mathbf{y})$; thus, we do not really need a point estimate (i.e. a single estimated value of $S - C$) to make inferences. We can derive probabilities from the areas contained in the density function, as we have done before, establishing the limits for the true value with a determined probability of our choice. In both, classical and Bayesian statistics, it looks somewhat strange to say that our estimate is 10, just to immediately state that we do not know whether its true value is between 9 and 11 or not. However, if we need a point estimate, for example, to facilitate comparisons with other authors, we have in a Bayesian context several choices (Fig. 2.5). We can take the mean, the median or the mode of the posterior distribution as a point estimate

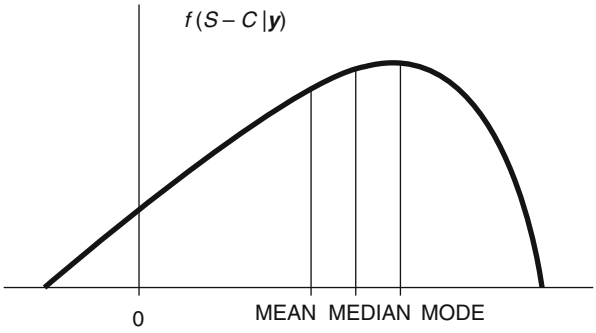

Fig. 2.5 Mean, median and mode of the posterior distribution of the difference between specific and control feed, given the information provided by our data

and say that the difference between S and C is, say, 160 g/day, 150 g/day or 140 g/day. Each point estimate has its advantages, and we will discuss them below.

As we have seen in Chap. 1, a good estimator should minimise the RISK of the estimation, i.e. the mean of the loss function. The problem here is that each point estimate is based in a different loss function; thus, each one minimises a different risk. Calling u the unknown and $\widehat{u}$ its estimate, the risk minimised by each point estimator is (Appendix 2.1):

	RISK minimised
MEAN	$E_u(\widehat{u}-u)^2$
MEDIAN	$E_u\lvert\widehat{u}-u\rvert$
MODE	$RISK = 0$ if $\widehat{u} = u$, $RISK = 1$ otherwise

MEAN: It is quite common to use the mean because it minimises the risk that is more familiar to us. However, the risk function of the mean has two inconveniences. Firstly, it penalises high errors, since we work with the square of the error, and it is not clear why should we do this. Secondly, this risk function is not invariant to transformations, i.e. the risk of u^2 is not the square of the risk of u. For example, if we estimate the variance as the mean of its posterior distribution, and we do the same with the standard deviation, one estimate is not the square of the other (the same happens in a frequentist context when estimating parameters by least squares: a transformation of a least square estimate is not necessarily a least square estimate itself).

MODE: It is quite popular for two reasons—one reason is that it is the most probable value and the second one is that, in the era before MCMC, it was easier to calculate than the other estimates. Unfortunately, mode has a terrible loss function. To understand what this function means, see the (rather artificial) distribution shown in Fig. 2.6, representing the posterior distribution of a correlation coefficient given our data.

This distribution has a negative mode, but although the most probable value is negative, the coefficient is probably positive, because the area of probability in the positive side is much higher. Only if we are right and the true value is exactly the mode, we will not have losses.

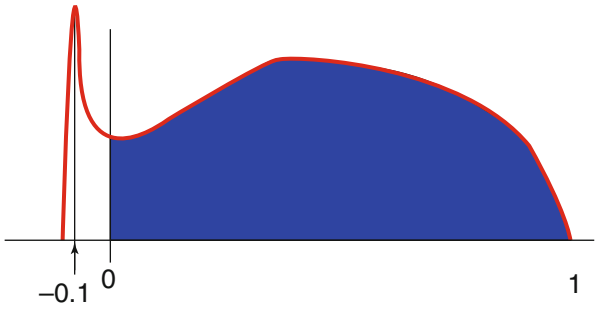

Fig. 2.6 Posterior probability distribution of a correlation coefficient given the data. The mode is negative, but the coefficient of correlation is probably positive

MEDIAN: The true value has a 50% of probability of being higher or lower than the median. The median has an attractive loss function in which the errors are considered according to their value (not to their square or any other transformation). The median has also the interesting property of being invariant to one-to-one transformations (e.g. if we have five values and we calculate the square of them, the median is still in the middle, and the median of the set of squared values is the square of the former median). A short demonstration is in Appendix 2.2. Statisticians tend to prefer the median as a point estimator when using posterior distributions (this should not be confused with the use of the median of a sample when we want to estimate the population mean; in this case the median is less accurate than the mean).[6]

When the amount of data increases, the distributions tend to become Normal (see Appendix 2.3), and the mean, median and mode tend to become coincident. Nevertheless, some parameters like the correlation coefficient show asymmetric distributions, particularly near the limits of the parametric space (near -1 or $+1$) even with samples that are not small. The same happens with heritabilities that are near zero, variances and other parameters. In these cases, it is not trivial to choose the mean, mode or median as a point estimator.

Notice that *Risk* has the same definition as it did in the frequentist case we saw in Chap. 1, Sect. 1.4.1 (the mean of the loss function), but here the variable is not $\widehat{u}$. Our estimate is a combination of the data, and in a Bayesian context, the data are fixed; we do not repeat the experiment conceptually in an infinite number of times. Here the variable is u because we make probability statements about the unknown value; thus, we use a random variable u that has the same name as the constant unknown true value. This is a frequent source of confusion (see 'misinterpretations' below).

2.2.2 Credibility Intervals

2.2.2.1 Highest Posterior Density Interval

Bayesian inference provides probability intervals. Now, the confidence intervals (Bayesians prefer to call them credibility intervals) contain the true value with a probability of 95% or with other probabilities defined by the user. An advantage of the Bayesian approach, mainly through MCMC procedures that we will see in Chap. 4, is the possibility of construction of all kinds of intervals easily. This allows us to ask questions that we could not ask within the classical inference approach. For example, if we provide the median and the mode and we ask for the precision of

[6]Please do not confuse the median of the distribution with the median of a sample when we want to estimate the population mean. In the latter case, the median has less information than the arithmetic mean (in a frequentist context, a large s.e.), and the mean should be preferred. We are considering here probability distributions in the continuous case, with an infinite number of points; we are not using a sampling estimator.

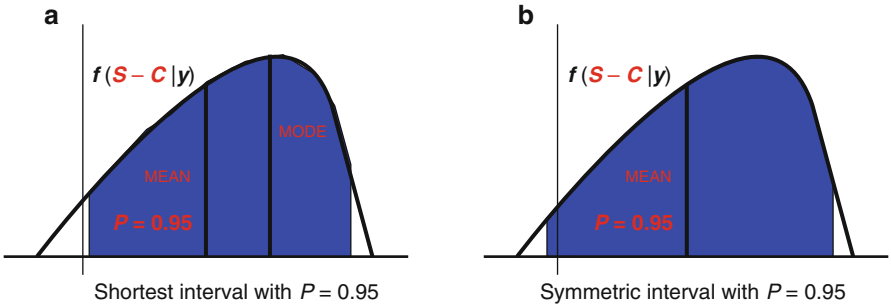

Fig. 2.7 Credibility intervals containing the true value with a probability of 0.95. (**a**) Shortest interval (not symmetric around the mean or the mode). (**b**) Symmetric interval around the mean

our estimation, we can find the *shortest interval* containing the true value with a 95% probability (what is called the *highest posterior density interval* at 95%, HPD95%). We like short intervals because this means that the value we are trying to estimate is between two close values. Notice (and this is important) that this interval is independent from the estimate we give, and it can be asymmetric around the mean or the mode (Fig. 2.7a). Of course, we can also obtain the symmetric interval about the mean or the mode containing 95% of the probability (Fig. 2.7b) if this is what we wish, although it would be larger. Notice that zero is included in the symmetric interval in this example, but it is not in the shortest interval. Including zero is relevant in classical statistics because it is not possible to state whether S is larger than C or not, it is related to nonsignificant differences. However, as we will see below, including zero in the Bayesian confidence interval is irrelevant, because we have other intervals to determine the probability of the difference $S - C$ being higher than zero. It may happen, for example, that zero is included in the interval HPD95% and simultaneously the probability $S - C > 0$ being 96%.

2.2.2.2 Probability of Being Positive (or Negative)

We can also calculate the probability of the difference between S and C being higher than 0 (Fig. 2.8a), which is the same as the probability of S being greater than C. In the case in which S is less than C, we can calculate the probability of $S - C$ being negative, i.e. the probability of S being less than C (Fig. 2.8b). Notice that *this is not a hypothesis test*, since we are not comparing two hypotheses as in Chap. 1, Sect. 1.2; this is the *estimation* of the actual probability of $S > C$ (or that of $S < C$ if the difference is negative). This can be more practical than a test of hypothesis; as we will argue later, we do not need hypothesis tests for many biological problems. Notice that we can estimate the probability of $S - C$ being positive or negative, but we do not estimate the probability of $S - C = 0$ because this probability is always zero. We know that $S - C$ is not going to be exactly 0.0000000.... Later we will define the probability of similitude, for the cases in which $S - C$ is irrelevant.

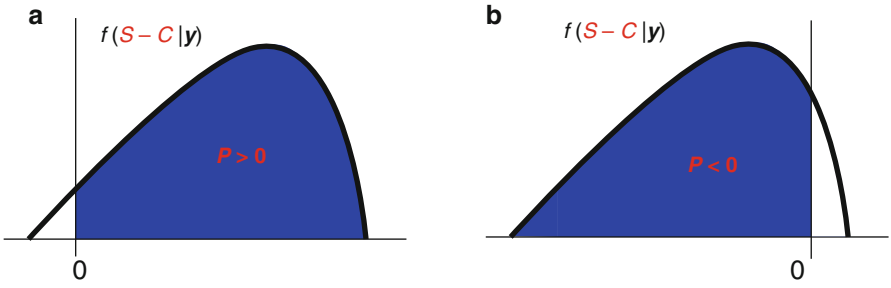

Fig. 2.8 Credibility intervals. (**a**) Interval $(0, +\infty)$ showing the probability of $S - C$ being higher than zero when $S > C$. (**b**) Interval $(-\infty, 0)$ showing the probability of $S - C$ being lower than zero when $S < C$

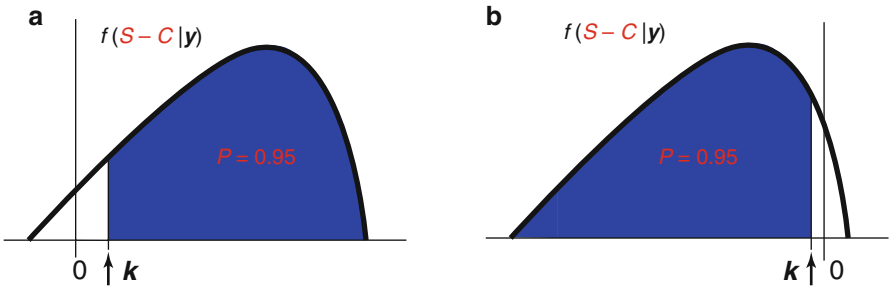

Fig. 2.9 Credibility intervals. (**a**) Interval $[k, +\infty)$ showing the minimum guaranteed value with a probability of 95%. (**b**) Interval $(-\infty, k]$ showing the maximum guaranteed value with a probability of 95%

2.2.2.3 Guaranteed Value

In some cases, it may be important to know how high this difference is with a chosen probability, for example, 95%. By calculating the interval $[k, +\infty)$ containing 95% of the probability (Fig. 2.9a), we can state that the probability of $S - C$ being less than this value k is only a 5%, i.e. we can state that $S - C$ takes *at least* a value k with a probability of 95% (or the probability we decide to take).

If S is lower than C, we can calculate the interval $(-\infty, k]$ and state that the probability of $S - C$ being higher than k is only 5% (Fig. 2.9b).[7] We call k the *guaranteed value* at a determined probability. For many problems, we will be satisfied with a guaranteed value of lower probability, say 80%.[8]

[7] These intervals have the advantage of being invariant to transformations. HPD 95% intervals are not invariant to transformations, i.e. the HPD 95% interval for the variance is not the square of the HPD 95% interval for the standard deviation. The same happens with frequentist confidence intervals.

[8] This looks smaller than the common use of 95%, but for a *guaranteed* value, it is not. It is quite common to discuss results based in a point estimation without looking at the bounds of the confidence interval. Here a guaranteed provides some safety in the discussion.

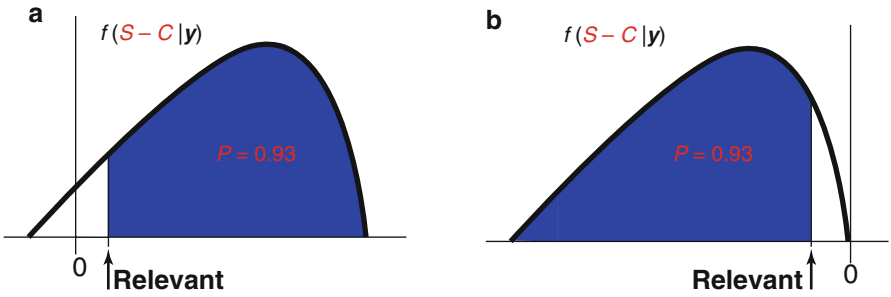

Fig. 2.10 Credibility intervals. (**a**) Interval from a relevant quantity to $+\infty$, showing the probability of the difference $S - C$ of being relevant. (**b**) Intervals from $-\infty$ to a relevant quantity, showing the probability of the difference $S - C$ of being relevant

2.2.2.4 Probability of Relevance

In practice, we are interested not only in finding whether S is higher than C or not, but in whether this difference is *relevant*. S may be higher than C, but this difference may be irrelevant. A *relevant* value is what we consider the *minimum* difference between S and C having an economical or biological meaning; of course, higher values are also relevant, but we define 'relevant value' as the *minimum relevant value*. Relevant values are defined *before* performing the experiment and do not depend on it but on economical or biological considerations (see Chap. 1, Sect. 1.2.2 and Appendix 1.1 for finding relevant values). In fact, we can define 'relevant' values even if we are not performing any experiment. Relevant values are commonly used in experimental design to determine the sample size; in classical statistics, they are the values of a trait from which differences between treatments should appear as 'significant' when calculating the sample size of the experiment.

We can calculate the probability of the difference $S - C$ being relevant, what we call *probability of relevance*. For example, if we are measuring lean content in pigs, we can consider that 1 mm of back fat is a relevant difference between S and C groups and calculate the probability of $S - C$ being more than 1 mm (Fig. 2.10a).

When S is lower than C, the relevant value is negative, and we calculate the probability of it being lower than this value (Fig. 2.10b).

2.2.2.5 Probability of Similitude

We can be interested in finding whether S is different from C. When we mean 'different', we mean higher or lower than a *relevant* value, defined as before: the minimum value with economical or biological meaning. For most biological problems, we are certain that S is different from C because they cannot be *exactly equal*. When comparing adult weight of beef cattle breeds, we are rather sure that they will differ at least in one kg, or one g or one mg, but they will never have exactly the same weight. Nevertheless, if we define a relevant value of, say 100 kg, we can obtain the probability of the absolute value of the difference between these breeds being lower than 100 kg, i.e. the probability of this difference being

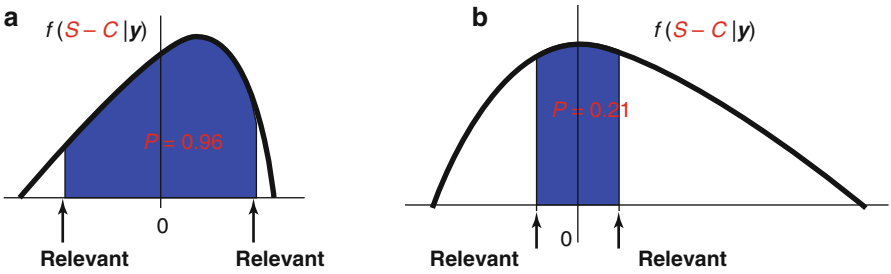

Fig. 2.11 Probability of similitude between S and C. (**a**) S and C do not differ in practice. (**b**) We do not have data enough to determine whether S is higher, lower or similar to C

irrelevant, null for practical purposes. If this probability is high (Fig. 2.11a), we can say that both breeds are equal for practical purposes.

However, it may happen that our sample is not high enough to establish conclusions, and we have, for example, a probability of similitude of 21%, a probability of relevance for $S > C$ of 49% and a probability of relevance for $C > S$ of 30% (Fig. 2.11b). In this case, we cannot establish any conclusion. Therefore, using the probability of similitude, we can differentiate between 'there is no difference' (for practical purposes) or 'I do not know whether there is a difference or not'. Comparing with the 'n.s.' result of a classical test of hypothesis, 'nonsignificance' cannot state that there are no differences between both breeds; we do not know. As before, the probability of similitude is not a hypothesis test but the actual probability of the breeds being equal (for our purposes).

We can also apply the concepts of relevance and similarity to parameters; for example, we can say that a correlation between two variables is zero for us when it is between -0.1 and 0.1; if so, our decisions concerning these two variables will be the same as if they were completely independent. If the probability of a correlation coefficient being between -0.1 and 0.1 is high, we can consider that this coefficient is null in practice (it will never be exactly 0.00000...).

It is important to notice that we are talking about relevant differences, not about infinitesimal differences. If we try to draw conclusions from figures like Fig. 2.11 when the relevant quantity is extremely small, we can have problems related to the prior distribution. Figure 2.11 shows posterior distributions, and they have been derived using a prior distribution. For most problems, we will have data enough and we can use vague priors that will be dominated by the data distribution, as we will see in Chap. 9, but if the area in blue of Fig. 2.11b is extremely small, even in these cases the prior can have an influence in the conclusions. Therefore, the probability of similitude and probabilities for other confidence intervals should be applied for relevant values, which in practice will *never* be infinitesimal.

2.2.2.6 Credibility Intervals of Ratios

The use of the probabilities of relevance and similitude, although attractive, has the problem of defining a relevant difference or a relevant value of a parameter. For many traits, it is difficult to state what is a relevant difference or a relevant value.

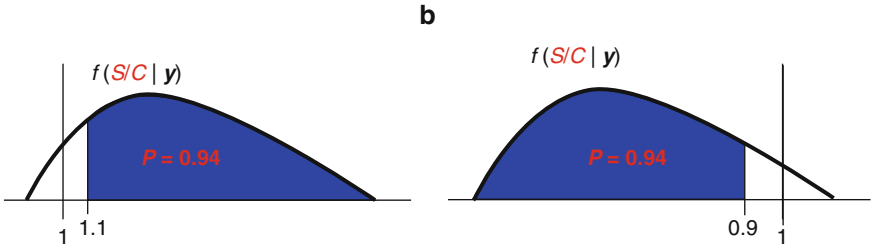

Fig. 2.12 Credibility interval for the ratio of levels of a treatment. (**a**) The probability of S being 10% higher than C is 0.94. (**b**) The probability of S being 90% of C is 0.94

For example, if we measure the effect of a treatment on enzymes activities, it is not clear which difference in enzyme activity can be considered as 'relevant'. We have proposed a procedure for these cases in Appendix 1.1 in Chap. 1, but we have here another solution. For these cases, we can express our results as ratios instead of differences. We will make the same inferences as before, but now using the marginal posterior distribution of the ratio S/C instead of the posterior distribution of $S - C$. We can apply all the credibility intervals exposed before, now to the ratio S/C instead of applying them to the difference $S - C$. For example, we can calculate the probability of the treatment S being a 10% higher than the control C for a particular trait (Fig. 2.12). If S is lower than C, we can be interested in the probability of the selected population being, for example, lower than a 90% of the control population (Fig. 2.12b).

We have doubts about what can be a relevant difference for enzyme activities, but we know what it means that a treatment produces a 10% more of some enzyme activity. This is possible to do in classical statistics by computing the ratios of least square means, but on one the hand, the ratio does not have the statistical properties of least square estimation any more, and on the other hand, obviously, the s.e. of the ratio is not the ratio of the s.e. If we want the s.e. of the ratio, we should use approximations using Taylor series, it is not immediate. The use of MCMC techniques easily provides the posterior distribution of the ratios and all the confidence intervals.

2.2.3 Marginalisation

One of the main advantages of Bayesian inference is the possibility of marginalisation. This implies that in a model with several variables and parameters, we can examine the uncertainty of each variable and parameter one by one, taking into accounts the errors of estimation of all the other unknowns. Take, for example, a mixed model: in classical statistics, we estimate the variance components and use them for the estimation of the fixed effects we have in the model. We use point estimates of the variances taking their values as true values, with no error. Another example is the estimation of growth curves or lactation curves: we estimate the

curve for each animal, take the curve's parameters as true values and then apply a fixed effect or a mixed model to these 'true' values estimated with no error. The same can be said for the estimation of residual food intake and, in general, for any 'nested' model in which some parameters are considered as 'true values' in order to apply another statistical mode on them. Classical statistics has no solution for 'nested' models, but Bayesian statistics has a solution, easy to apply considering the MCMC methods that we will see in Chap. 4. This is possible because Bayesian inference is based in probabilities. Inferences in a Bayesian context are made from the *marginal* posterior distribution of an effect or a parameter, having integrated out all other unknowns, i.e. giving to them all their possible values, multiplying by the probability of each value and summing up. We will see how it works in a simple example.

Let us come back to the group of people composed by British and Spanish men and women. We have made a contract with this group of people, and we should now pay them. We have the average salary they earn in the following table, in thousands of euros per year, divided by sex and nationality.[9] The percentage of individuals of each type in our group of people is between brackets.

	British	Spanish
Men	36 (40%)	26 (10%)
Women	30 (20%)	20 (30%)

Now we would like to know how much we should pay to the British people and to the Spanish people, independently on whether they are men or women. We need to *marginalise* our data. We see that among British people, 2/3 are men and1/3 are women,[10] so

$$\text{Salary for British} = 36 \cdot \frac{2}{3} + 30 \cdot \frac{1}{3} = 34$$

We also see that 1/4 of Spanish people are men and 3/4 are women, so

$$\text{Salary for Spanish} = 26 \cdot \frac{1}{4} + 20 \cdot \frac{3}{4} = 21.5$$

Thus, we have

British	Spanish
34 (60%)	21.5 (40%)

[9]These are approximate figures, but near the real ones, according to Eurostat.

[10]Doing that properly, the proportion of British men is $\frac{0.4}{0.4+0.2}$ and the same holds for the other proportions.

Now instead of two variables, sex and nationality, we have only nationality; we have *integrated out* the variable sex that has now disappeared.

In our example about treatments S and C, we do not know the residual variance of the model y = Treatment + e, which has to be estimated from the data. Suppose that this residual variance can only take two values, $\sigma^2 = 0.5$ and $\sigma^2 = 1$. The marginal posterior distribution of the difference between treatments will be the sum of $f(S - C$ given the data and given that $\sigma^2 = 0.5)$ and $f(S - C$ given the data and given that $\sigma^2 = 1)$ multiplied by the respective probabilities of σ^2 taking these values.

$$f(S - C|\boldsymbol{y}) = f\left(S - C|\boldsymbol{y}, \sigma^2 = 0.5\right) \cdot P\left(\sigma^2 = 0.5\right) + f\left(S - C|\boldsymbol{y}, \sigma^2 = 1\right) \cdot P\left(\sigma^2 = 1\right)$$

When σ^2 can take all possible values from 0 to ∞, instead of summing up, we calculate the integral of $f(S - C|\,\boldsymbol{y}, \sigma^2)$ for all possible values of σ^2 from 0 to ∞.

$$f(S - C|\boldsymbol{y}) = \int_0^\infty f\left(S - C, \sigma^2|\boldsymbol{y}\right) d\sigma^2 = \int_0^\infty f\left(S - C|\boldsymbol{y}, \sigma^2\right) f\left(\sigma^2\right) d\sigma^2$$

Thus, when we marginalise, we take all possible values of the unknowns, we multiply by their probability and we sum up. This has two main consequences:

1. We concentrate our efforts of estimation only in the posterior probability of the unknown of interest. All multivariate problems are converted in a set of univariate problems of estimation.
2. We take into account the uncertainty of all other parameters when we are estimating the parameter of interest.

2.3 Test of Hypothesis

2.3.1 Model Choice

Sometimes, we face real hypothesis testing, for example, when it should be decided in court a verdict of 'guilty' or 'innocent', but as we have stressed before, there is no need of hypothesis tests for most biological experiments. This has been emphasised by Gelman et al. (2013), Howson and Urbach (1996) and Robert (1992) among many others for both, frequentist and Bayesian statistics. Nevertheless, there are circumstances in which we may need to select one among several statistical models. For example, when we are trying to eliminate some noise effects of a model that have too many levels to be tested two by two, or when we have many noise effects and we would like to know whether we could be freed from them. Even in these cases, our opinion is that the researcher should know when to add noise effects, from her biological knowledge of the problem; statistics is a tool, and it cannot substitute thinking. We will come back on this problem on Chap. 10.

Now suppose we have two models to be compared, one model has an effect; the other model does not have this effect. This is known as 'nested' model. In the Bayesian case, we have a wider scope. We can compare models that are not nested. For example, we can try to fit different growth curves to the data based on different functions. We can also compare several models simultaneously, based on different hypotheses. Suppose we have Hypothesis 1, 2, 3..., we can calculate $P(H_1|\boldsymbol{y})$, $P(H_2|\boldsymbol{y})$, $P(H_3|\boldsymbol{y})$, ... and choose the most probable one. Here we are not assuming risks at 95% as in frequentist statistics, the probabilities we obtain are the actual probabilities of these hypotheses, thus if we say that, when comparing two hypotheses, H_1 has a probability of 90% and H_2 has a 10%, we can say that H_1 is 9 times more probable than H_2. Notice that we are giving the probabilities relative to the hypotheses that are being tested; for example, if we test two hypotheses and one of them has a probability of 60%, the other one will have a 40%, but if we test a third hypothesis and it has a 10% probability, the probability of the other two ones is modified.

To calculate the probability of each hypothesis, we have to perform marginalisation as we have seen before. We give all possible values to the parameters that need to be estimated $\boldsymbol{\theta}$, we multiply by their probability and we sum up. In the continuous case, we integrate instead of summing. For each hypothesis H, we have:

$$P(H|y) = \int f(\boldsymbol{\theta}, H|\boldsymbol{y}) d\boldsymbol{\theta}$$

As we will see in Chap. 10, Sect. 10.2.2, these integrals are highly dependent on the prior information $f(\boldsymbol{\theta})$, which makes Bayesian model choice extremely difficult.

Model choice is a difficult area that is still under development in statistical science. We have first to decide which will be our criterion for preferring a model to others. Test of hypothesis is only one of the possible ways of model choice. We can use criteria based in amount of information, or criteria based on probability, or heuristic criteria that we have derived by simulation or according to our experience. We can also choose a model based in its predictive properties for new data. In the last chapter, we will discuss the different approaches to model selection.

2.3.2 Bayes Factors

A common case is to have only two hypotheses to be tested, then

$$\frac{P(H_1|\boldsymbol{y})}{P(H_2|\boldsymbol{y})} = \frac{\frac{P(\boldsymbol{y}|H_1)\cdot P(H_1)}{P(\boldsymbol{y})}}{\frac{P(\boldsymbol{y}|H_2)\cdot P(H_2)}{P(\boldsymbol{y})}} = \frac{P(\boldsymbol{y}|H_1)\cdot P(H_1)}{P(\boldsymbol{y}|H_2)\cdot P(H_2)} = \mathrm{BF}\cdot\frac{P(H_1)}{P(H_2)}$$

where

$$\text{BF} = \frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_2)}$$

is called 'Bayes factor' (although Bayes never used it, this was actually proposed by Laplace). In practice most people consider that 'a priori' both hypotheses to be tested have the same probability, then if $P(H_1) = P(H_2)$ we have

$$\text{BF} = \frac{P(\boldsymbol{y}|H_1)}{P(\boldsymbol{y}|H_2)} = \frac{P(H_1|\boldsymbol{y})}{P(H_2|\boldsymbol{y})}$$

and we can use Bayes factors to compare the posterior probabilities of two hypotheses. The main problem with Bayes factors is that the probabilities of the hypotheses are sensitive to the prior distributions of the unknowns $f(\boldsymbol{\theta})$. Moreover, if we have complex models, Bayes factors are difficult to calculate. Notice that the justification of the use of Bayes factors in a Bayesian context is that they express the ratio of posterior probabilities when prior probabilities are the same, they are not justified as 'the support that the hypothesis gives to the observed data' or other informal justifications.

2.3.3 Model Averaging

Another possibility of Bayesian inference is to do model averaging. This interesting procedure for inferences has no counterpart in frequentist statistics. It consists in using simultaneously several models for inferences, weighted according to their posterior probabilities. For example, if we are interested in estimating a parameter θ that appears in both models and has the same meaning in both models (this is important!), we can find that, given the data, H_1 has a probability of 70% and H_2 has of 30%. This is unsatisfactory, because when choosing H_1 as the true model and estimate θ with it, there is still a considerable amount of evidence in favour of H_2. Here we face the problem we saw in Chap. 1 when having insufficient data to choose one model; our data does not support either model 1 or model 2. In a classical context, the problem has no solution because the risks are fixed before the analysis is performed, and they do not represent the probability of the model being true, as we explained in Chap. 1. In a Bayesian context, we have the actual probabilities of each model, and we can make inferences *from both hypotheses*, weighing each one by its probability.

$$P(\theta|\boldsymbol{y}) = P(\theta, H_0|\boldsymbol{y}) + P(\theta, H_1|\boldsymbol{y}) = P(\theta|H_0, \boldsymbol{y}) \cdot P(H_0|\boldsymbol{y}) + P(\theta|H_1, \boldsymbol{y}) \cdot P(H_1|\boldsymbol{y})$$

We should be careful, in that θ should be the same parameter in both models; for example, the parameters b, k of the logistic growth curve have different meanings than the same parameters in the Gompertz growth curve.

2.4 Common Misinterpretations

The main advantage of Bayesian inference is the use of prior information: This would be true if prior information was easy to integrate in the inference. Unfortunately, this is not the case, and most modern Bayesians do not use prior information except as a tool that allows them to work with probabilities. The real main advantage of Bayesian inference is the possibility of working with probabilities, which allows making inferences about the unknowns based in probabilities and permits marginalisation.

Bayesian statistics is subjective; thus, researchers find what they want: Sometimes Bayesian statistics can be subjective (when there is vague prior information), but *subjective* does not mean *arbitrary*, as we have discussed before. It is true that we can always define a prior probability that will dominate the results, but if we really believe in a highly informative prior, why should we perform any experiment? Subjective priors should always be vague, and data should usually dominate the results. We have an additional difficulty, which is that in multivariate cases subjective priors are almost impossible to be defined properly (in Chap. 9 we will come back on this topic).

Bayesian results are a blend of the information provided by the data and by the prior: This should be the ideal scenario, but as said before, it is difficult to integrate prior information; thus, modern Bayesians try to minimise the effect of the prior. They do this using enough data in order to be sure that data will dominate the results, so that after checking several vague priors or after using minimum informative priors, the results stay the same.

The posterior of today is the prior of tomorrow: This is usually untrue, at least in biological and agricultural experiments. When we are analysing a new experiment, other people have been working in the field, so our last posterior should be integrated subjectively with this new information. Moreover, we will normally try to avoid the effect of any prior by having enough data. We will not normally use our previous posterior as a new prior.

In Bayesian statistics, the true value is a random variable: We can find statements like 'in frequentist statistics, the sample is a variable and the true value is fixed, whereas in Bayesian statistics the sample is fixed, and the true value is a random variable'. This is nonsense. The true value θ_{TRUE} is a constant that we do not know. We use the random variable θ (which is not the true value) to make probability statements about this unknown true value θ_{TRUE}. Unfortunately, frequentist statisticians use θ as the true value; thus, this is a source of confusion; what is worse, some Bayesian statisticians use θ for both the true value and the variable used to express uncertainty about the true value. Perhaps Bayesian statisticians should use another way of representing the random variable used to make statements about the unknown true value, but the common practice is to use σ^2 to represent the random variable used to express uncertainty about the true value σ^2_{TRUE}.

Bayesian statistics ignores what would happen if the experiment were repeated: We can be interested in which would be the distribution of a Bayesian estimator if the experiment were repeated. In this case, we are not using frequentist statistics because our estimator was derived under other basis, but we would like to know what would happen when repeating our experiment or we would like to examine the frequentist properties of our estimator. To know what will happen when repeating an experiment is a sensible question and Bayesian statistics often examine this.[11]

Credibility intervals should be symmetric around the mean: This is not necessary. These intervals do not represent the accuracy of the mean or the accuracy of the mode but another way of estimating our unknown quantity, *interval estimation* instead of *point estimation*.

Credibility intervals should always contain a 95% probability: The choice of the 95% was made by Fisher (1925a, p. 46) because approximately two standard deviations of the Normal function included 95% of its values. Here we are not working with significance levels; thus, we obtain actual probabilities. If we obtain 89% probability, we should ask ourselves whether this uncertainty is enough for the trait that we are examining. For example, we may never play lottery, but if a magician told us that if playing tomorrow we would have an 80% probability to win, we can seriously consider playing. However, if the magician says that if driving tomorrow we would have a 95% probabilities to survive, we can consider that the risk is too high and not to drive.

When 0 is included in the credibility interval 95%, there are no significant differences: Firstly, there is no such thing as 'significant differences' in a Bayesian context. We do not have significance levels, since we can measure the actual probability of a difference being greater than zero. Secondly, in a frequentist context, 'significant differences' is the result of a hypothesis test, and we are not performing any test by using a credibility interval; the result of a frequentist test is 'yes' or 'no', but we are here estimating the precision of an unknown. Finally, the Bayesian answer to assess whether S is greater than C is not to offer an HPD95% but to calculate the probability of $S > C$.

We can calculate the probability of $S > C$ and the probability of $S < C$, but my interest is which is the probability of $S = C$, how can I calculate this? It is not necessary to calculate it if you are working with a continuous variable, as usual in biology or agriculture. This probability is always zero because there are infinite numbers in the real line. The question is not correctly formulated. We are not interested in knowing whether the difference between S and C is $0.0000000000\ldots$ but in whether if it is lower than a value small enough to consider this difference to be irrelevant. Then we can find probabilities of similitude as in Fig. 2.11.

[11] This does not mean that Bayesian estimators have good frequentist properties. For example, they are usually biased due to the prior. However, this does not mean that they are not good estimators, what happens is that their 'good properties' are different.

We need hypothesis tests for checking whether there is or not an effect: Let us assume, for example, that there is a sex effect in the comparison of treatments. We can compare two models for growth rate at a determined age, one with a sex effect and other model without sex effect, and choose the most probable one. However, if we chose the model with no sex effect, this does not mean that the difference between males and females for this trait is zero. As we said before, the question is not properly formulated. We are not interested in knowing whether the sex difference is 0.0000000 ... but if it is lower than a quantity that would be irrelevant to us. Even when performing a hypothesis test, a negative answer does not mean that we are sure about the absence of this effect, only that *our data is compatible with the absence of this effect*, which is not the same. Moreover, if we have few data, our sample will always be compatible with the absence of the effect we are testing; we have seen in Chap. 1 that differences are always significant if the sample is large enough. Because of the difficulties of performing tests, that we mentioned before, in practice, it is much easier and more informative to find the posterior probability of the difference between males and females in order to check whether this difference is relevant or not. If the probability of similitude between males and females is high, i.e. if this difference in absolute value has a high probability of being lower than a relevant value, the sex effect can be ignored.

Bayes factors contain all information provided by the data; thus we can make inferences with no prior probabilities: Inferences are made in a Bayesian context from posterior distributions. A ratio of posterior distributions is the product of a Bayes factor by the ratio of prior probabilities of the hypotheses; thus, it is true that all information coming from the data is contained in the Bayes factor. The problem is that *we need the prior probabilities to make inferences*; we cannot make inferences without them because in this case we cannot apply Bayes theorem. In a Bayesian context, when we make inferences from Bayes factors, it is *always* assumed that prior probabilities of both hypotheses are the same.

Bayes factors show which hypothesis makes the data more probable: Again, as in the case of maximum likelihood we discussed in Chap. 1, Bayes factors show which hypotheses, *if it was the true hypothesis and not otherwise*, would make the sample more probable. This is not enough to make inferences because it does not lead to learning which hypothesis is the most probable one, and this is the only way for drawing inferences in a Bayesian context. Our interest is not to know which is the hypothesis that, if it was the true one, will make our data more probable but to find, given our data, which is the most probable hypothesis. In the first case, we do not have any measure of evidence; in the second case, we can compare the probabilities of both hypotheses.

Bayes factors are equivalent to the maximum likelihood ratio: The maximum likelihood ratio is a technique that shows how to construct hypothesis tests in the frequentist world, leading to chi-square distributions that can be used for drawing rejection areas for nested hypothesis tests. The interpretation is thus completely different. Moreover, Bayes factors use the average likelihoods, not the likelihoods at their maximum, which can lead to different results when likelihoods are not symmetric. Finally, remember that Bayes factors can only be used for inferences

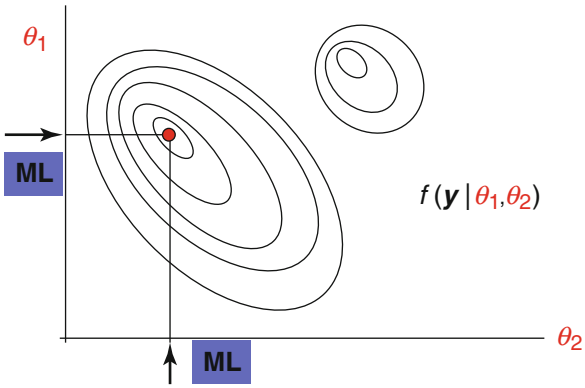

Fig. 2.13 Probability density of the data for two parameters. Lines should be interpreted as level curves of a map

when prior probabilities of both hypotheses are the same. When used for inferences, Bayes factors show ratios of probabilities, a much more informative way of dealing with uncertainty than likelihood ratio tests.

Bayesian statistics give the same results as likelihood when the prior is flat: The shape of the function can be the same, but the way of making inferences is completely different. We have seen that likelihoods cannot be integrated because they are not probabilities; thus, no credibility intervals can be constructed with the likelihood function and no marginalisation of the parameters can be made.

Marginal posterior distributions are like maximum likelihood profiles: In classical statistics, for example, in genetics when searching major genes, maximum likelihood profiles are used. They consist in finding the maximum likelihood estimate for all parameters but one and examine the likelihood curve substituting all unknowns, with the exception of this one, by their maximum likelihood estimates. In Fig. 2.13, we represent the likelihood of two parameters θ_1 and θ_2. The lines should be taken as level curves in a map; we have two 'hills', one higher than the other one. The objective in classical analysis is to find the maximum of this figure, which would be the top of the high 'hill' forgetting the rest of the hill although it contains some information of interest. When a maximum likelihood profile is made by 'cutting' the hill along the maximum likelihood of one of the parameters in order to draw a maximum likelihood profile, the smaller 'hill' is still forgotten. In the Bayesian case, if these 'hills' represent a posterior distribution of both parameters, marginalisation will take into account that there is a small 'hill' of probability, and all the values of θ_2 in this area will be multiplied by their probability and summed up in order to construct the marginal posterior distribution of θ_1.

2.5 Bayesian Inference in Practice

In this section, we will follow the examples given by Blasco (2005) with small modifications. Bayesian inference modifies the approach to the discussion of the results. Classically, we have point estimation, usually a least square mean, and its

standard error, accompanied by a hypothesis test indicating whether there are differences between treatments according to a significance level previously defined. Then we discuss the results based upon these features. Now, in a Bayesian context, the procedure is inverted. We should first ask which question is relevant for us and then go to the marginal posterior distribution to find an answer.

Example 1 We take an example from Blasco et al. (1994). They were interested in finding the differences in percentage of ham from a pig final cross using Belgian Landrace or Duroc as terminal sire. They offer least square means of 25.1 ± 0.2 kg and 24.5 ± 0.2 kg, respectively, and find that they are significantly different. Now, in order to present the Bayesian results, we have estimated the marginal posterior distribution of the difference between both crosses. Now, we should ask some questions:

1. *What is the difference between both crosses?*

We can offer the mean, the mode or the median. Here the marginal distribution is approximately Normal; thus, the three parameters are approximately the same, and the answer is coincident with the classical analysis: 0.6 kg.

2. *What is the precision of this estimation?*

The most common Bayesian answer is the highest posterior density interval containing a probability of 95%. However, here the marginal posterior distribution is approximately Normal, and we know that the mean $\pm$ twice the standard deviation of the marginal posterior distribution will contain approximately this probability; therefore, we can give either an interval [0.1 kg, 1.1 kg] or just the standard deviation of the difference, 0.25 kg.

3. *What is the probability of the Belgian Landrace cross being higher than the Duroc cross?*

We do not need a test of hypothesis to answer this question. We can just calculate how much probability area of the marginal posterior distribution is positive. We find a 99% probability. Please notice that, with a smaller sample, we could have found a high posterior density interval containing a 95% of probability of [−0.1 kg, 1.3 kg], if, for example, the standard deviation would have been 0.30 kg and still say that the probability of the Belgian Landrace cross being higher than the Duroc is, say, 96%. This is because one question is *the accuracy of the difference*, shown in Fig. 2.7a, and another question is *whether there is a difference*, shown in Fig. 2.8a. In this last figure, we do not need the tail of probability of the right side of Fig. 2.7a. Using Bayesian inference, we should choose the best way for answering each question.

4. *How large can we say is this difference with a probability of 95%?*

We calculate the interval $[k, +\infty)$ (see Fig. 2.9a), and we find out that the value of k is 0.2 kg; thus, we can say that the difference between crosses is at least of 0.2 kg with a probability of 95%; we have a *guaranteed value* of 0.2 kg at 95% probability. We could have also estimated a guaranteed value at 80% or at other probability value; see, for example, Martínez-Álvaro et al. (2016).

5. *Considering that an economical relevant difference between crosses is 0.5 kg, what is the probability of the difference between crosses being relevant?*

We calculate the probability of being higher than 0.5 (Fig. 2.10a), and we find the probability of relevance to be 66%. Thus, we can say that although we consider that both crosses are different, the probability of this difference being relevant is of only 66%.

Example 2 We now take a sensory analysis from Hernández et al. (2005). Here a rabbit population selected for growth rate is compared with a control population, and sensory properties of meat from the *l. dorsi* are assessed by a panel test. The panels test scored from 0 to 5, and the data was divided by the standard deviation of each panelist in order to avoid a scale effect. In this example, it is difficult to determine what a relevant difference is, because it is difficult to understand whether a difference in 0.3 points of liver flavour, for example, is high or not. Thus, instead of assessing the differences between the selected (S) and control (C) population, the ratio of the selection and control effects S/C is analysed (see Fig. 2.12). This allows expressing the superiority of the selected over the control population (or conversely the superiority of the control over the selected population) in percentage. We will take the trait liver flavour. The result of the classical analysis is that the least square means of the selected and control populations are 1.38 ± 0.08 and 1.13 ± 0.08 points, respectively, and they were found to be significantly different. These means and their standard error are rather inexpressive about the effect of selection on meat quality. Now, the Bayesian analysis answers the following questions:

1. *What is the probability of the selected population being higher than the control population?*

We calculate the probability of the ratio being higher than 1. We find a 99% probability; thus, we conclude they are different.

2. *How much higher is the liver flavour of the selected population with respect to the control population?*

As in Example 1, we can give the mean, the mode or the median, and all of them are approximately coincident; we find that the liver flavour of the selected population is 23% higher than the liver flavour of the control population.

3. *Which is the precision of this estimation?*

The 95% high posterior density interval goes from 1.03 to 1.44, which means that the liver flavour of the selected population is between 3% and 44% higher than this flavour in the control population with a probability of 95%.

4. *How large can we say is this difference with a probability of 95%?*

We calculate the *guaranteed value*, the k of the interval $[k, +\infty)$, and we find that the value of k for the ratio S/C is 1.06; thus, we can say that selected population is at least 6% higher than control population with a probability of 95%. We can say alternatively that the probability of selected population being lower than 6% of the control population has a probability of only 5%. In practice, lower probabilities are used for guaranteed values; often 80% is enough (see, for example, Zomeño et al. 2013).

5. *Assuming being 10% higher as relevant, what is the probability of the selected population being 10% higher than the control population?*

We calculate the probability of the ratio of being higher than 1.10, and we find this value to be 88%. This means that the probability of the effect of selection on liver flavour being relevant is 88%. This is not related to significance thresholds or rejection areas; we can state that this is the actual probability.

Example 3 Progesterone participates in the release of mature oocytes, facilitation of implantation and maintenance of pregnancy. Most progesterone functions are exerted through its interaction with specific nuclear progesterone receptor. Peiró et al. (2008) analyse a possible association of a *PGR* gene polymorphism (GG, GA, AA) with litter size in a rabbit population. They consider that 0.5 kits per litter is a relevant quantity. The GG genotype had higher litter size than GA genotype; the difference between genotypes D was relevant and $P(\mathrm{D} > 0) = 99\%$. The GA genotype had similar litter size as the AA genotype (probability of similarity = 96%), which indicates that the genetic determination of this trait is dominant. Here, the probability of similitude (Fig. 2.11a) means that the area of the posterior distribution of P(GA-AA | $\boldsymbol{y}$) included between −0.5 and +0.5 kits was 96%.

Example 4 Hernandez et al. (1998) estimated the correlation between moisture and fat percentage in hind leg meat of rabbit, obtaining a coefficient of -0.95 ± 0.07. This standard error is not very useful, since the sampling distribution of the correlation coefficient is not symmetrical (the correlation cannot be lower than −1; thus, the '±' is misleading).[12] A Bayesian analysis obtained the marginal

[12]Using bootstrap techniques, a drawing of the sampling distribution of the correlation coefficient can be obtained, showing that when repeating many times the experiment, the samples will not be symmetrical. However, this sampling distribution does not allow making probability statements about the unknowns but only about the sampling distribution, and it is more complex than the MCMC procedures used for estimating marginal posterior distributions that we will see in Chap. 4.

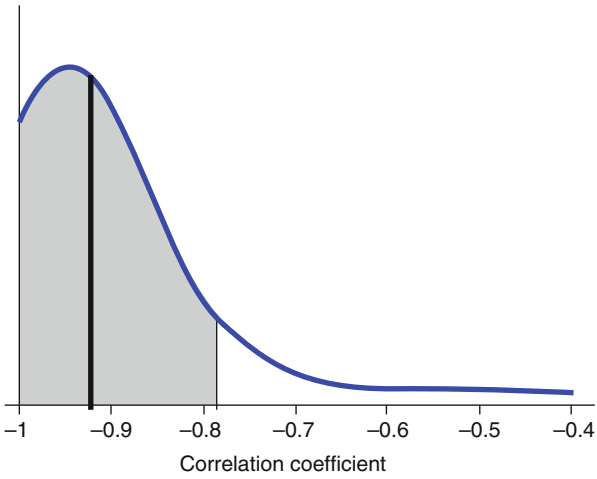

Fig. 2.14 Probability distribution of a correlation coefficient. Notice that it is highly asymmetric

posterior distribution shown in Fig. 2.14. Here the distribution is asymmetrical; thus, mode, mean and median are not coincident. A usual choice for a point estimation is to take the mean (−0.93) because it minimises the quadratic risk, which is conventional, although there are reasons for taking the median, as we said before. Here the HPD interval at 95% is [−1.00, −0.79], not symmetrical around the mean, and shows better the uncertainty about the correlation than the s.e. of the classical analysis. The probability of this correlation being negative is almost one.

2.6 Advantages of Bayesian Inference

We will summarise here the advantages of Bayesian inference:

- We have a *measure of uncertainty* for both hypothesis tests and credibility intervals, since we work with probabilities. We do not have 'prior' risks.
- We are not worried about bias (there is no such thing as 'bias' in a Bayesian context).
- We should not decide whether an effect is fixed or random (all of them are random); we will consider different prior distributions instead, as we will see in Chap. 7, Sect. 7.3.2.
- We work with marginal probabilities, i.e. all multivariate problems are converted to univariate, and we take into account errors of estimating other parameters.
- We have a method for inferences, a path to follow. We know that we should calculate marginal posterior distributions, and we can express uncertainty in several ways using them.

Appendix 2.1

1. *The mean minimises the risk:* $R = E_u(\widehat{u} - u)^2$.

We should find the value of $\widehat{u}$ that minimises R

$$\frac{\partial R}{\partial \widehat{u}} = 2 \cdot E_u(\widehat{u} - u) = 0 \quad \rightarrow \quad E_u(u) = E_u(\widehat{u}) = \widehat{\mathrm{u}}$$

where $E_u(\widehat{u}) = \widehat{u}$, because u is not included in $\widehat{u}$. The estimate $\widehat{u}$ is only function of the data, and we are taking the mean E_u with respect to the values of u.

2. *The median minimises the risk:* $R = E_u|\widehat{u} - u|$

The loss function $|\widehat{u} - u|$ can be expressed as

$$L(\widehat{u}, u) = \left\{ \begin{array}{ll} \widehat{u} - u & \text{when } \widehat{u} > u \\ u - \widehat{u} & \text{when } u > \widehat{u} \end{array} \right\}$$

The median m is defined (see Chap. 3, Sect. 3.4.2) as

$$\int_{-\infty}^{m} f(x)dx = 0.50$$

We should prove now that the median m has a lower risk of the type $R = E_u|\widehat{u} - u|$ than any other estimator $\widehat{u}$.

Assume that $\widehat{u}$ is the estimator of minimum risk of this type, and consider without loss of generality that $\widehat{u} > m$ (an analogous demonstration can be made for $\widehat{u} < m$). Consider now the difference between the loss functions $D = L(m, u) - L(\widehat{u}, u)$.

$$\begin{array}{ll} m > u, \widehat{u} > u, & D = m - u - (\widehat{u} - u) = m - \widehat{u} \\ m < u, \widehat{u} < u, & D = u - m - (u - \widehat{u}) = \widehat{u} - m \\ m < u, \widehat{u} > u, & D = u - m - (\widehat{u} - u) = 2u - m - \widehat{u} < 2\widehat{u} - m - \widehat{u} = \widehat{u} - m \end{array}$$

We do not consider the case $m > u, \widehat{u} < u$ because we have assumed that $\widehat{u} > m$.

Thus, we can write

$$D \leq \left\{ \begin{array}{ll} m - \widehat{u} & \text{when } m > u\,(50\% \text{ of times, since } m \text{ is the median}) \\ \widehat{u} - m & \text{when } m < u\,(50\% \text{ of times, since } m \text{ is the median}) \end{array} \right\}$$

$$E_u(D) = E_u[L(m,u) - L(\widehat{u},u)]$$
$$\leq \frac{1}{2}(m-\widehat{u}) + \frac{1}{2}(\widehat{u}-m) = 0 \rightarrow E_u[L(m,u)] \leq E_u[L(\widehat{u},u)]$$

But as $\widehat{u}$ is the estimator of minimum risk, m cannot have a lower risk and should be $m = \widehat{u}$; thus, the median is the estimator of minimum risk.

3. *The mode minimises the risk: RISK = 0 if $\widehat{u} = u$, RISK = 1 otherwise*

The demonstration of this requires complex operations that are out of the scope of this book. The reader interested in it can consult, for example, Leonard and Hsu (1999) or http://web.uvic.ca/~dgiles/blog/zero_one.pdf.

Appendix 2.2

We know (see Chap. 3, Sect. 3.3.2) that, if y is a function of x,

$$f(y) = f(x) \cdot \left|\frac{dx}{dy}\right|$$

then

$$\text{median}\,(y) \rightarrow \int_{-\infty}^{m_y} f(y)dy = \frac{1}{2} = \int_{-\infty}^{m_y} f(x) \cdot \left|\frac{dx}{dy}\right| dy = \int_{-\infty}^{m_x} f(x)dx \rightarrow \text{median}(x)$$

Appendix 2.3

If we take a flat prior $f(\theta) = \text{constant}$, as we know that $x = \exp(\log x)$,

$$f(\theta|y) \propto f(y|\theta)f(\theta) \propto f(y|\theta) \propto \exp[\log f(y|\theta)]$$

We can develop a Taylor series of $\log f(\theta|\,y)$ around the mode m, up till the second term.

$$f(\theta|y) \propto \exp\left\{(\theta - m)\left[\frac{\partial \log f(\boldsymbol{y}|\theta)}{\partial \theta}\right]_{\theta=m} + \frac{1}{2}(\theta-m)^2\left[\frac{\partial^2 \log f(\boldsymbol{y}|\theta)}{\partial \theta^2}\right]_{\theta=m}\right\}$$

Since the mode is a maximum, the first derivative is null, thus

$$f(\theta|y) \propto \exp\left(\frac{1}{2} \cdot \frac{(\theta - m)^2}{\left[\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2}\right]^{-1}_{\theta=m}}\right)$$

which is the kernel of a Normal distribution with mean m and variance the inverse of the second derivative of the log of the density function of the data applied in the mode of the parameter (we will find this expression again when we will introduce the concept of information in Chap. 10). As the Normal distribution is symmetric, mode, mean and median is the same.

References

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

Blasco A (2005) The use of Bayesian statistics in meat quality analyses. Meat Sci 69:115–122

Blasco A, Gou P, Gispert M, Estany J, Soler Q, Diestre A, Tibau J (1994) Comparison of five types of pig crosses. I. Growth and Carcass traits. Livest Prod Sci 40:171–178

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

de Finetti B (1937) La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincairé 7:1–68. Translated in Kyburg HE, Smokler HE (1964) Studies in subjective probability. Wiley, New York

Donkin WF (1851) On certain questions relating to the theory of probabilities. Philos Mag 1:353–368. 2:55–60

Fisher R (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh

Fisher R (1936) Uncertain inference. Proc Am Acad Arts Sci 71:245–258

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. Chapman and Hall, Boca Raton, FL

Hernández P, Guerrero L, Ramírez J, Mekkawy W, Pla M, Ariño B, Ibáñez N, Blasco A (2005) A Bayesian approach of the effect of selection for growth rate on sensory meat quality of rabbit. Meat Sci 69:123–127

Hernandez P, Pla M, Blasco A (1998) Carcass characteristics and meat quality of rabbit lines selected for different objectives. II. Relationships between meat characteristics. Livest Prod Sci 54:125–131

Howie D (2002) Interpreting probability: controversies and developments in the early twentieth century. Cambridge University Press, New York

Howson C, Urbach P (1996) Scientific reasoning. The Bayesian approach. Open Court, Chicago, IL

Jeffreys H (1931) Scientific inference. Oxford University Press, Oxford

Jeffreys H (1961) Theory of probabilities, 3rd edn. Clarendon Press, Oxford

Kant E (1781/1998) Critique of pure reason. Reprinted in translation. Cambridge University Press, Cambridge

Kempthorne O (1984) Revisiting the past and anticipating the future. In: Statistics: an appraisal. Proc. 50th Anniversary Iowa State Statistical Laboratory. The Iowa State University Press, Ames, pp 31–52

Keynes JM (1921) A treatise on probability. Macmillan Publ. Co, London

Laplace PS (1774/1986) Memoir on the probabilities of the causes of events (Trad. by Stigler SM). Stat Sci 1:364–378

Leonard T, Hsu JSJ (1999) Bayesian methods. Cambridge University Press, New York

Martínez-Álvaro M, Hernández P, Blasco A (2016) Divergent selection on intramuscular fat in rabbits: responses to selection and genetic parameters. J Anim Sci 94:4993–5003
Pearson K (1920) The fundamental problems of practical statistics. Biometrika 13:1–16
Pearson E (1962) Some thoughts on statistical inference. Ann Math Stat 33:394–403
Peiró R, Merchán M, Santacreu MA, Argente MJ, García ML, Folch JM, Blasco A (2008) Progesterone receptor gene as candidate gene for reproductive traits in rabbits. Genetics 180:1699–1705
Ramsey FP (1931) Truth and probability. In: Braithwaite RB (ed) The foundation of mathematics and other logical essays. Routledge & Kegan, London
Robert CP (1992) L'Analyse statistique bayesienne. Economica, Paris, France
Zomeño C, Hernandez P, Blasco A (2013) Divergent selection for intramuscular fat content in rabbits. I Direct response to selection. J Anim Sci 91:4526–4531

http://www.springer.com/978-3-319-54273-7

Bayesian Data Analysis for Animal Scientists
The Basics
BLASCO, A.
2017, XVIII, 275 p. 160 illus., 151 illus. in color.,
Hardcover
ISBN: 978-3-319-54273-7