CHAPTER 2

THE BAYESIAN CHOICE

2.1. Bayesian inference

2.1.2. *Bayes theorem*

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

If we express the probability of an event B as the number of times N_B that the event B occurs in N outcomes, and the probability of the joint occurrence of two events A and B as the number of times N_{AB} that they occur in these N outcomes, we have

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

where the bar ‘|’ means “given”, i.e. the probability of the other event A is conditioned to that this event B takes place. The probability of taking a train at 12:00 to Valencia is the probability of arriving on time to the train station, given that there is a train to Valencia at this time, multiplied by the probability of having 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 the later.

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

This directly leads to the Bayes theorem:

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

Going back to our problem of estimation of chapter one, we are interested in assessing the effect of selection for growth rate of a rabbit population and we have a selected group of rabbits and a control group in which growth rate has been measured. If we call S the effect of the selected group and C the effect of the control group, we are interested in assessing (S – C). Traditionally, we will find the standard error, or confidence interval, of the difference between the averages of samples of both groups $\left(\bar{x}_S - \bar{x}_C\right)$. Bayesian inference will give a more attractive solution: we will find the distributions probabilities of all possible values of (S-C) according to the information provided by our data. This is expressed as P(S-C|**y**), where **y** is the set of data we use in the experiment. Applying Bayes theorem, we have

$$P(S\text{-}C|\mathbf{y}) = \frac{P(\mathbf{y}|S\text{-}C) \cdot P(S\text{-}C)}{P(\mathbf{y})}$$

thus, to make inferences based in probability we need to know

P(**y**|S-C): This is the distribution of the data for a given value of the unknowns. It 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 they should be normally distributed.

P(**y**) is a constant, the probability of the sample. Our sample is an event that obviously has a probability. After the appearance of MCMC techniques we do not need to calculate it.

P(S-C) is the probability of the difference between selected and control group independently of any set of data. It is interpreted as the information about this difference that we have before making the experiment. This prior information is needed to complete Bayes theorem and to let us make probability statements through P(S-C|**y**).

This gives a pathway for estimation. In classical statistics we do not have, with the exception of likelihood theory, any clear pathway to find good estimators. In Bayesian theory 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 in a topic; other colleagues should have performed related experiments that give some prior information about our experiment. If so, it would be very interesting to blend this information with the information provided by our experiment. This is common in classic papers in the section “Discussion”, in which our current results are compared with the results of other authors. Our conclusions are not only based in our work but also in this previous work, thus a formal integration of all sources of information looks attractive. Unfortunately it is almost impossible to do this formally, 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 in chapter 7. For example, the colour of the skin is determined by a single gene with two alleles (A,a). If a mouse receives the ‘a’ allele from both parents (then it is homozygous aa), its colour is brown, but if it receives an allele ‘A’ from one of the parents (in this case it can be either homozygous AA or heterozygous Aa), his colour is black. We try to know whether a black mouse, son of heterozygous mates (Aa x Aa), is homozygous (AA) or heterozygous (Aa) (figure 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 sure it is heterozygous, 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 get three offspring black. What is the probability for the black mouse is heterozygous, given this data?

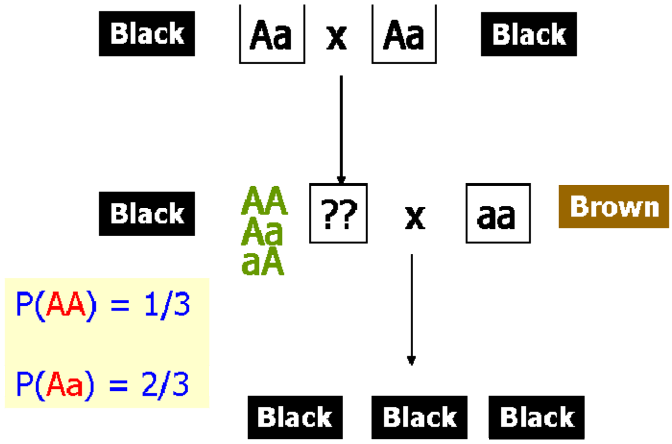


Figure 2.1. Two heterozygous mice have an offspring that may be homozygous or heterozygous. To test this it is crossed to a brown mouse and the offspring is examined. Before performing the experiment we have some prior information due to our knowledge of Mendel's law.

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' at the same time from both, because in this case it will be brown. This means that we have only three possibilities: or it received two alleles 'A', or it received one 'A' from the father and an '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 2/3, because there are two favourable possibilities in a total of three. We should blend this prior information with the information provided by the experiment, in our case having three offspring black when crossing this mouse with a brown mouse (aa).

When we have vague prior information: In most cases prior information is not so 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 can consider than their circumstances were only partially applicable to our case. However they provide information useful for us, and independently of whether they provide useful information or not, we need prior information in order to apply Bayes theorem. It can be proposed that probability describes beliefs. 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.

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, we will examine the problem of representing ignorance in chapter 7. 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.

2.2. Features of Bayesian inference

2.2.1. *Point estimates*

All information is contained in the probability distribution P(S-C|**y**), thus we do not really need point estimates to make inferences about the success of the selection experiment. In both, classical and Bayesian statistics, it looks somewhat strange to say that our estimate of something is 10, just to immediately state that we do not know whether its true value is between 9 and 11. However if we need a point estimate for some reason, we have in a Bayesian context several choices (figure 2.3).

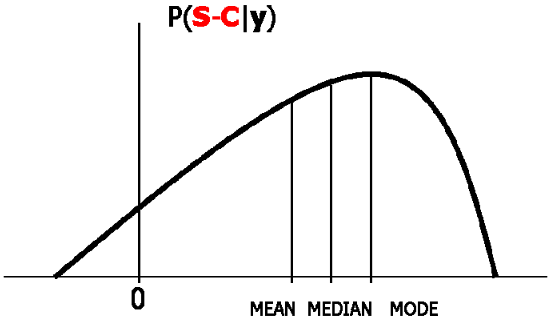


Figure 2.3. Mean, median and mode of the probability distribution of the difference between selected and control groups, given the information provided by our data.

It can be shown (see, for example, Bernardo and Smith, 1994) that each one minimizes a different Risk. Calling 'u' the unknown to be estimated and û its estimator, we have:

MEAN: minimizes *RISK* = $E(\hat{u} - u)^2$
MEDIAN: minimizes *RISK* = $E|\hat{u} - u|$
MODE: minimizes *RISK* = 0 if û=u, *RISK* = 1 otherwise

MEAN: It is quite common to give the mean of the distribution as an estimator because it minimizes the risk that is more familiar to us. However the risk function of the mean has two inconveniences. First, it penalizes high errors, since we work with the square of the error, and it is not clear why we should do this. Second, this risk function is not invariant to transformations; i.e., the risk of u^2 is not the square of the risk of u.

MODE: It is quite popular for two reasons: one is that it is the most probable value, and the second one is that in the era previous to MCMC it was easier to calculate than the other estimates, since no integrals were needed but only to find the maximum of the distribution. Unfortunately, mode has a horrible loss function. To understand what this function means, see the (rather artificial) example shown in figure 2.4. It represents the probability distribution of a correlation coefficient given our data. This probability 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.

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 other transformation). The median has also the interesting property of being invariant to transformations one-to-one (for example, if we have five values and we calculate the square of them, the median value is still the same). A short demonstration is in Appendix 2.1.

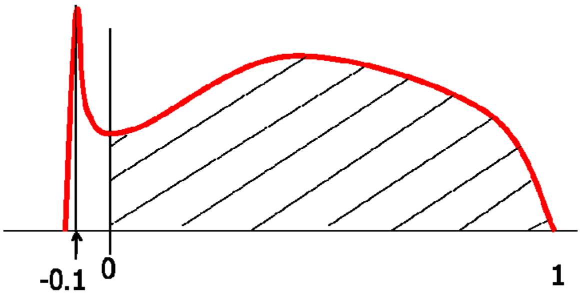


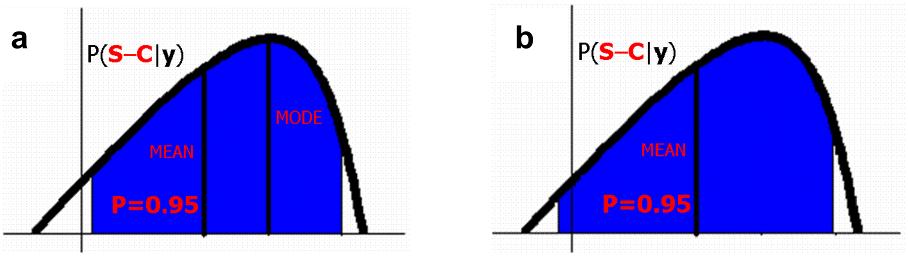
Figure 2.4. Probability distribution of a correlation coefficient given the data. The mode is negative, but the coefficient of correlation is probably positive.

When the number of data increases, the distributions tend to be normal (see Appendix 2.2), and then mean, median and mode tend to be coincident. Nevertheless, some parameters like the correlation coefficient show asymmetric distributions near the limits of the parametric space (near -1 or +1 in the case of the correlation coefficient) even with samples that are not small.

2.2.2. *Credibility intervals*

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 through MCMC procedures is the possibility of easy construction of all kind of intervals. This allows us to ask questions that we could not ask within the classical inference approach. For example, if we give the median and the mode and we ask for the precision of our estimation, we can find the shortest interval with a 95% probability of containing the true value (what is called the *Highest posterior density interval* at 95%). 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 here this interval is independent on the estimate we give, and it can be asymmetric around the mean or the mode (figure 2.5.a). Of course, in the Bayesian case we can also obtain the symmetric interval about the mean or the mode containing 95% of the probability (figure 2.5.b).

We can also calculate the probability of the difference between S and C being higher than 0 (Figure 2.6.a), 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 (Figure 2.6.b). This can be more practical than a test of hypothesis, since we will know the exact probability of S being higher than C. As we will argue later, we do not need hypothesis tests for most biological problems.



Shortest interval with P=0.95

Symmetric interval with P=0.95

Figure 2.5. Credibility intervals containing the true value with a probability of 95%. **a.** Shortest interval (not symmetric around the mean or the mode). **b.** Symmetric interval around the mean.

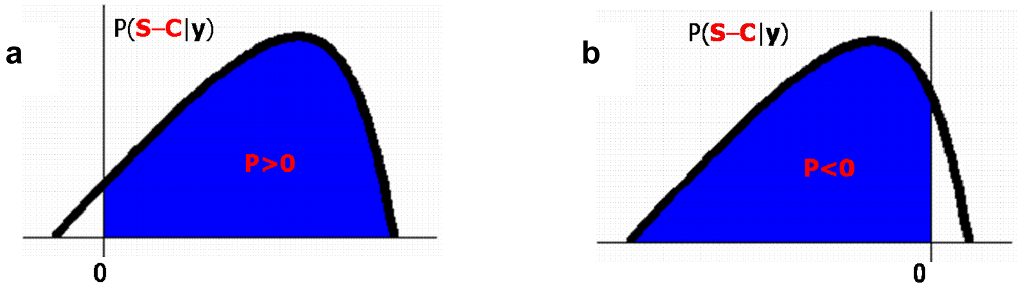


Figure 2.6. Credibility intervals. **a.** Interval [0, +∞) showing the probability of S-C of being equal or higher than zero. **b.** Interval (-∞, 0]) showing the probability of S-C of being equal or lower than zero

In some cases it may be important to know how big we can state that this difference is with a probability of a 95%. By calculating the interval [*k*, + ∞) containing 95% of the probability (Figure 2.7.a) 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 (-∞, *k*] and state that the probability of S-C being higher than *k* is only a 5% (Figure 2.7.b).

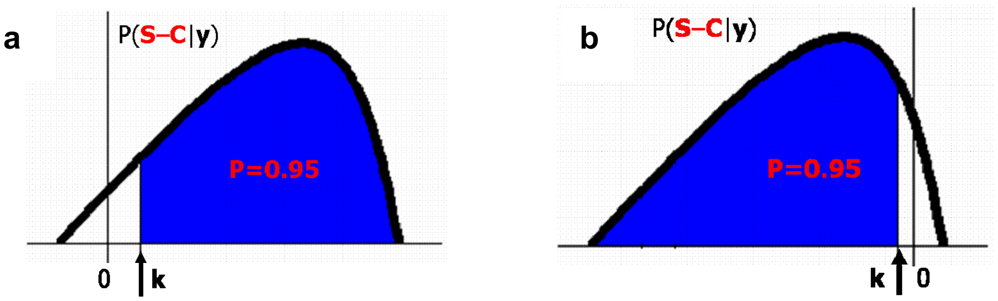


Figure 2.7. Credibility intervals. **a.** Interval [k, +∞) showing the lowest value of an interval containing the true value with a probability of 95%. **b.** Interval (-∞, k] showing the highest value of an interval containing the true value with a probability of 95%.

In practice, we are interested not only in finding whether S is higher than C or not, but in whether this difference is *relevant* (see 1.2.2 and Appendix 1.1 for a discussion). S may be higher than C, but this difference may be irrelevant. We can calculate the probability of this difference being relevant. For example, if we are measuring lean content in pigs, we can consider that 1 mm of backfat is a relevant difference between S and C groups, and calculate the probability of S-C being more than 1 mm (Figure 2.8.a).

We can be interested in finding whether S is different from C. When we mean “different” we mean higher or lower than a *relevant* quantity, since we are sure that S is different from C because they are not going to be *exactly equal*. For example, if we are comparing ovulation rate of two lines of mice and we consider that one ovum is the relevant quantity, we can find the probability of the difference between strains of being higher or lower than one ovum (Figure 2.8.a).

We can establish the probability of similarity (Figure 2.9.a) and say for example that S=C with a probability of 96% (here ‘=’ means that the differences between S and C are irrelevant). This is different from saying that their difference is “non significant” in the frequentist case, because N.S. means that *we do not know* whether they are different. If we have few data, a low probability of similarity does not necessarily means that S and C are different; we simply do not know it (Figure 2.9.b). The important matter is that we can make the difference between “S is equal to C” and “we do not know whether they are different or not”.

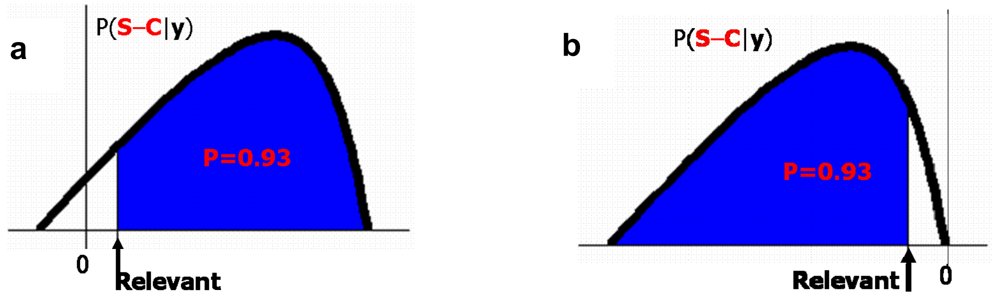


Figure 2.8. Credibility intervals. **a.** Interval from a relevant quantity to +∞, showing the probability of the

difference S-C of being relevant. **b.** Intervals from –∞ to a relevant quantity, showing the probability of the difference S-C of being relevant

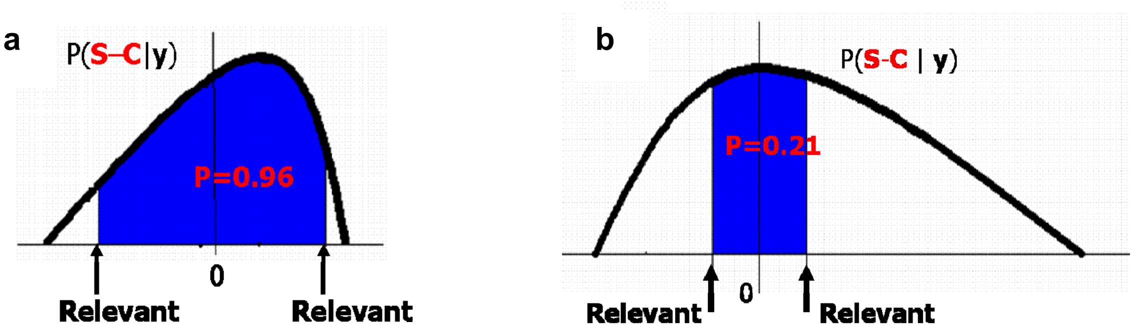


Figure 2.9. Probability of similarity between S and C. **a.**S and C are similar. **b.** We do not have data enough to determine whether S is higher, lower or similar to C.

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 2.9 in which the relevant quantity is very small, we can have problems related to the prior distribution. Figure 2.9 shows posterior distributions, and they have been derived using a prior. 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 chapter 7, but if the area in blue of figure 2.9 is extremely small, even in these cases the prior can have an influence in the conclusions.

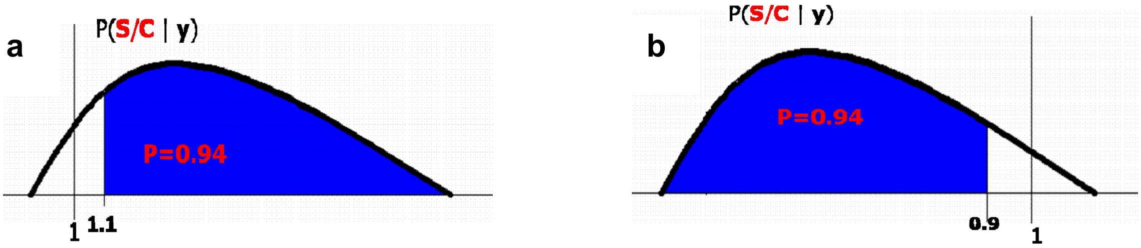


Figure 2.10. Credibility interval for the ratio of levels of a treatment. **a.** The probability of S being a 10% higher than C is a 94%. **b.** The probability of S being a 90% of C is a 94%

However, for many traits, it is difficult to state which is a relevant difference. For example, if we measure the effect of selection on enzymes activities it is not clear what we can consider to be 'relevant'. I have proposed a procedure for these cases in Appendix 1.1, but we have 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 distribution of S-C. For example, we can calculate the probability of the selected population being a 10% higher than the control population for a particular trait (figure 2.10). 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 (figure 2.10.b).

2.2.3. *Marginalisation*

One of the main advantages of Bayesian inference is the possibility of marginalisation. In classical statistics we estimate, for example, first the variance of the error of the model, we take it as being estimated without any error of estimation and we use it later to estimate other parameters or other errors. In animal breeding, when applying BLUP or selection indexes, we should estimate before the genetic parameters and take these estimates as the true ones. In Bayesian statistics we can take into account the inaccuracy of estimating these variance components. This is possible because Bayesian inference is based in probabilities. Inferences are made from the *marginal* posterior distribution, having integrated out, weighing by their probabilities, all the possible values of al the other unknown parameters. In our example about treatments S and C, we do not know the residual variance, that has to be estimated from the data. Suppose that this residual variance only can take two values, σ^2 = 0.5 and σ^2 = 1. The marginal posterior distribution of the difference between treatments will be the sum of P(S-C given the data and given that σ^2 = 0.5) and P(S-C given the data and given that σ^2 = 1) multiplied by the respective probabilities of σ^2 taking these values.

$$P(S\text{-}C \mid y) = P(S\text{-}C \mid y, \sigma^2 = 0.5)\, P(\sigma^2 = 0.5) + P(S\text{-}C \mid y, \sigma^2 = 1)\, P(\sigma^2 = 1)$$

When σ^2 can take all possible values from 0 to ∞, instead of summing up we calculate the integral of P(S-C | y, σ^2) for all values of σ^2 from 0 to ∞.

$$P(S-C\mid y) = \int_0^\infty P\left(S-C,\sigma^2 \mid y\right)d\sigma^2 = \int_0^\infty P\left(S-C\mid y,\sigma^2\right)P\left(\sigma^2\right)d\sigma^2$$

Thus, we take all possible values of the unknowns, we multiply by their probability and we sum up. This has two 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*

Suppose we have two models to be compared, or two hypotheses to be tested. One hypothesis can be that S-C = 0 (i.e., S=C) and an alternative hypothesis can be S-C≠0 (i.e., S≠C in our former model). This is how the classical hypothesis tests are presented; we test what is called “nested” hypothesis: one model has an effect; the other model has not this effect. In the Bayesian case we have a wider scope. We can compare models that are not nested. We can calculate the probability of each hypothesis P(H_1|**y**), P(H_2|**y**) using Bayes theorem

$$P(H_1 \mid \mathbf{y}) = \frac{P(\mathbf{y}\mid H_1)\cdot P(H_1)}{P(\mathbf{y})} = \frac{P(\mathbf{y}\mid H_1)\cdot P(H_1)}{P(\mathbf{y}\mid H_1) + P(\mathbf{y}\mid H_2)}$$

where the probability of the sample P(**y**) is the sum of the probabilities of two excluding events: H_1 is the true hypothesis or it is H_2.

We calculate P(H_2|**y**) in the same way. After calculating P(H_1|**y**), P(H_2|**y**) we can choose the most probable hypothesis. This can be extended to several hypotheses, we can calculate P(H_1|**y**), P(H_2|**y**), P(H_3|**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 exact probabilities of these hypotheses, thus if we say that 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. To calculate the probability of each hypothesis, we have to act like in marginalisation. We give all possible values to θ, given our data, we multiply by their probability and we sum up. In the continuous case we integrate instead of summing. For each hypothesis H, we have

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

.

these integrals are highly dependent on the prior information f(θ), which makes Bayesian model choice extremely difficult. Heuristic solutions not based in the Bayesian paradigm but using Bayesian elements have been proposed (intrinsic Bayes factors, posterior Bayes factors, etc.) and will be discussed in last chapter.

2.3.2. *Bayes factors*

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

$$\frac{P(H_1\,|\,\mathbf{y})}{P(H_2\,|\,\mathbf{y})} = \frac{\dfrac{P(\mathbf{y}\,|\,H_1)\cdot P(H_1)}{P(\mathbf{y})}}{\dfrac{P(\mathbf{y}\,|\,H_2)\cdot P(H_2)}{P(\mathbf{y})}} = \frac{P(\mathbf{y}\,|\,H_1)\cdot P(H_1)}{P(\mathbf{y}\,|\,H_2)\cdot P(H_2)} = BF\cdot\frac{P(H_1)}{P(H_2)}$$

where

$$BF = \frac{P(\mathbf{y}\,|\,H_1)}{P(\mathbf{y}\,|\,H_2)}$$

is called “Bayes Factor” (although Bayes never used it, this was 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

$$BF = \frac{P(\mathbf{y}\,|\,H_1)}{P(\mathbf{y}\,|\,H_2)} = \frac{P(H_1\,|\,\mathbf{y})}{P(H_2\,|\,\mathbf{y})}$$

and we can use Bayes factors to compare the posterior probabilities of two hypotheses, which is the most common use of the Bayes factor. The main problem with Bayes factors is that they are sensitive to prior distributions of the unknowns f(θ). Moreover, if we have complex models Bayes factors are difficult to calculate.

2.3.3. *Model averaging*

Another possibility is to do model averaging. This is an interesting procedure for inferences that has no counterpart in frequentist statistics. It consists in using simultaneously both 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 H_1 has a probability of 70% and H_2 has a 30%. This is unsatisfactory, because although we can choose H_1 as the true model and estimate θ with it, there is a considerable amount of evidence in favour of H_2. Here we face the problem we saw in chapter 1 when having insufficient data to choose one model; our data do 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 of being true, as we explained in chapter 1. In a Bayesian context we can multiply each hypothesis by its probability because they are the true probabilities of each model, and we can make inferences from *both hypotheses*, weighting each one according to the evidence we have of each one.

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

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 meaning than the same parameters in the Gompertz growth curve. Another example: we cannot compare a linear regression coefficient with the linear coefficient of a quadratic regression, because the parameters are not the same.

2.4. Common misinterpretations

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

Bayesian statistics is subjective, thus researchers find what they want: A part of Bayesian statistics (when there is vague prior information) can be subjective, but this does not mean *arbitrary*, as we have discussed before. It is true that no mind how many data we have, we always can define a prior probability that will dominate the results, but if we really believe in this highly informative prior, why should we perform any experiment? Subjective priors should be always vague and data are almost always going to dominate. Moreover, in multivariate cases subjective priors are almost impossible to be defined properly.

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

The posterior of today is the prior of tomorrow: This is Bayesian propaganda. When we are analyzing 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 normally will try to avoid the effect of any prior, having data enough as said before. We almost never will 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 u_0 is a constant that we do not know, we use the random variable 'u' (which is not the true value) to make probability statements about this unknown true value u_0. Unfortunately, frequentist statisticians use 'u' as the true value, thus this is a source of confusion; which is worse, some Bayesian statisticians use 'u' 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_0^2' and we will use this nomenclature in this book.

Bayesian statistics ignores what would happen if the experiment is repeated: We can be interested in which would be the distribution of a Bayesian estimator if the experiment would be 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 happens 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 ([4]).

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

Credibility intervals should always contain a 95% of probability: The choice of the 95% was made by Fisher (1925) because approximately two standard deviations of the normal function included the 95% of its values. Here we are not working with significance levels, thus we obtain actual probabilities. If we obtain 89% of probability we should ask ourselves whether this is enough for the trait we are examining. For example, I never play lottery, but if a magician says to me that if I play tomorrow I have an 80% of probabilities to win, I will play. However if the magician says to me that if I take the car tomorrow to travel I have a 95% of probabilities to survive, I will not take it.

When 0 is included in the credibility interval 95%, there are no significant differences: First, there is nothing like "significant differences" in a Bayesian context. We do not have significance levels; we measure exactly the probability of a difference being equal or greater than zero. Second, in a frequentist context "significant differences" is the result of a hypothesis tests, and we are not performing any test by using a credibility interval; the result of a frequentist test is "yes" or "not", but we are here evaluating the precision of an unknown. Finally, the Bayesian answer to assess whether S is equal or greater than C is not a HPD 95% but a [k,+∞] interval with a 95% of probability or 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, 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… , but in whether it is lower than a value that would be small enough to consider that this difference is irrelevant. Then we can find probabilities of similitude as in figure 9.

We need hypothesis tests to check whether an effect exists or not: Some people say that if S=C there is no effect of selection, thus a model considering this effect is wrong and cannot be used to draw inferences like "P(S-C)>0 is very low". This worries particularly geneticists that think they cannot say that a trait has no dominance effects (for example) unless we have a test, because if we say that the dominant variance is irrelevant we are assuming its existence. As we said before, the question is not properly formulated. We are not interested in knowing whether the dominance variance or the selection effect is 0.0000000 … but whether it is lower than a quantity that will be irrelevant for us. We are not interested in knowing whether the heritability of a trait is 0.000000… but in whether it is small enough to prevent a selection program for this trait. Even if we perform a test, a negative answer is not that we are sure about the absence of this effect, but only that *our data are compatible with the absence of this effect*, which is not the same. In practice it is much easier to work with posterior probabilities than to perform model choice tests.

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

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

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 we cannot apply Bayes theorem. We can try to find more or less esoteric interpretations of the Bayes factor to intuitively understand what it means, but we cannot make inferences. When we make inferences from Bayes factors, it is 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 chapter 1, Bayes factors show which hypothesis, *if it would be the true hypothesis and not otherwise* would make the sample more probable. This is not enough to make inferences because it does not leads to which hypothesis is the most probable one, which is the question, *and it is the only question*, that permits to draw inferences.

Bayes factors are equivalent to the maximum likelihood ratio: The maximum likelihood ratio is a technique to construct hypothesis tests in the frequentist world, since it leads to chi-square distributions that can be used to draw 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 only can be used for inferences when prior probabilities of both hypotheses are the same.

Bayesian statistics gives 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 that are not of interest 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 for one, and examine the likelihood curve substituting all unknowns but this one by their maximum likelihood estimates. In figure 2.11 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. The objective in classical analysis is to find the maximum of this figure, which will be the top of the high "hill" forgetting the other 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.

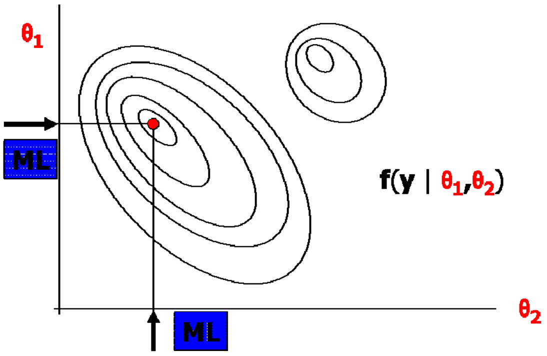


Figure 2.10. Probability function of the data for two parameters. Lines should be interpreted as level curves of a map.

2.5. Bayesian inference in practice

In this section we will follow the examples given by Blasco (2005) with small modifications (the references for works quoted in this paragraph can be found in Blasco, 2005). 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 ask first 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 of 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 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%. But here the marginal posterior distribution is approximately normal, thus we know that the mean ± twice the standard deviation of the marginal posterior distribution will contain approximately this probability, thus we can either give an interval [0.1 kg, 1.1 kg] or just the standard deviation of the difference, 0.25 kg.

3) *Which 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% of probability. Please notice that we could have found a high posterior density interval containing a 95% of probability of [0.0 kg, 1.2 kg], if for example the standard deviation would have been 0.30kg, and still say that the probability of the Belgian Landrace cross being higher than the Duroc is, say, a 97%. This is because one question is *the accuracy of the difference*, which is measured in Figure 2.5.a, and another question is *whether there is a difference*, which is answered in Figure 2.6.a, in which we do not need the tail of probability of the right side of Figure 2.5.a.

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

We calculate the interval [k, +∞) (see figure 2.7.a) and we find that the value of k is 0.2 kg, thus we can say that the difference between crosses is at least a 0.2kg with a probability of 95%.

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

We calculate the probability of being higher than 0.5 (Figure 2.8.a) and we find the value to be 66%. Thus, we can say that although we consider that both crosses are different, the probability of this difference being relevant is only a 66%.

Example 2. We take now 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 score from 0 to 10, and data were 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, 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 analyzed (see figure 2.7). 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 flavor. 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 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) *Which is the probability of the selected population being higher than the control population?*

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

2) *How much higher is the liver flavor 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 as the marginal distribution is also approximately normal all of them are coincident, we find that the liver flavor of the selected population is a 23% higher than the liver flavor 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 flavor of the selected population is between a 3% to a 44% higher than this flavor in the control population with a probability of a 95%.

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

We calculate the interval [k, +∞) 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 a 6% higher than control population with a probability of 95%, or that the probability of selected population being lower than a 6% of the control population has a probability of only a 5%.

5) *Considering being a 10% higher as relevant, which is the probability of the selected population of being a 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 flavor being relevant is 88%. This is not related to significance thresholds or rejection areas, we can state that this is