BAYESIAN STATISTIC COURSE

Edinburgh, September 4-7th 2007

LECTURE NOTES

Agustín Blasco

AN INTRODUCTION TO BAYESIAN ANALYSIS AND MCMC

Agustín Blasco

Departamento de Ciencia Animal.
Universidad Politécnica de Valencia
P.O. Box 22012. Valencia 46071. Spain
ablasco@dca.upv.es
www.dcam.upv.es/dcia/ablasco

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

Ronald Fisher, 1936.

AN INTRODUCTION TO BAYESIAN ANALYSIS AND MCMC

PROGRAM

1. Do we understand classical statistics?

1.1. Historical introduction

1.2. Test of hypothesis

1.2.1. The procedure

1.2.2. Common misinterpretations

1.3. Standard errors and Confidence intervals

1.3.1. The procedure

1.3.2. Common misinterpretations

1.4. Bias and Risk of an estimator

1.4.1. Unbiased estimators

1.4.2. Common misinterpretations

1.5. Fixed and random effects: a permanent source of confusion

1.5.1. Definition of “fixed “ and “random” effects

1.5.2. Bias, variance and Risk of an estimator when the effect is fixed or random

1.5.3. Common misinterpretations

1.6. Likelihood

1.6.1. Definition

1.6.2. The method of maximum likelihood

1.6.3. Common misinterpretations

Appendix 1.1

Appendix 1.2

Appendix 1.3

2. The Bayesian choice

2.1. Bayesian inference

2.1.1. Base of Bayesian inference

2.1.2. Bayes theorem
2.1.3. Prior information
2.2. Features of Bayesian inference
2.2.1. Point estimates: Mean, median, mode
2.2.2. Credibility intervals
2.2.3. Marginalisation
2.3. Test of hypotheses
2.3.1. Model choice
2.3.2. Bayes factors
2.3.3. Model averaging
2.4. Common misinterpretations
2.5. Bayesian Inference in practice
2.6. Advantages of Bayesian inference

3. Posterior distributions
3.1. Notation
3.2. Cumulative distribution
3.3. Density distribution
3.3.1. Definition
3.3.2. Transformed densities
3.4. Features of a density distribution
3.4.1. Mean
3.4.2. Median
3.4.3. Mode
3.4.4. Credibility intervals
3.5. Conditional distribution
3.5.1. Definition
3.5.2. Bayes Theorem
3.5.3. Conditional distribution of the sample of a Normal distribution
3.5.4. Conditional distribution of the variance of a Normal distribution
3.5.5. Conditional distribution of the mean of a Normal distribution
3.6. Marginal distribution
3.6.1. Definition
3.6.2. Marginal distribution of the variance of a normal distribution

3.6.3. Marginal distribution of the mean of a normal distribution
Appendix 3.1
Appendix 3.2
Appendix 3.3
Appendix 3.4

4. MCMC

4.1. Samples of Marginal Posterior distributions
4.1.1. Taking samples of Marginal Posterior distributions
4.1.2. Making inferences from samples of Marginal Posterior distributions
4.2. Gibbs sampling
4.2.1. How it works
4.2.2. Why it works
4.2.3. When it works
4.2.4. Gibbs sampling features
4.2.5. Example
4.3. Other MCMC methods
4.3.1. Acceptance-Rejection
4.3.2. Metropolis
Appendix 4.1

5. The baby model

5.1. The model
5.2. Analytical solutions
5.2.1. Marginal posterior distribution of the mean and variance
5.2.2. Joint posterior distribution of the mean and variance
5.2.3. Inferences
5.3. Working with MCMC
5.3.1. Using Flat priors
5.3.2. Using vague informative priors
5.3.3. Common misinterpretations
Appendix 5.1
Appendix 5.2
Appendix 5.3

6. The linear model

6.1. The “fixed” effects model

6.1.1. The model

6.1.2. Marginal posterior distributions via MCMC using Flat priors

6.1.3. Marginal posterior distributions via MCMC using vague informative priors

6.1.4. Least Squares as a Bayesian Estimator

6.2. The “mixed” model

6.2.1. The model

6.2.2. Marginal posterior distributions via MCMC

6.2.3. BLUP as a Bayesian estimator

6.2.4. REML as a Bayesian estimator

6.3. The multivariate model

6.3.1. The model

6.3.2. Data augmentation

Appendix 6.1

7. Prior information

7.1. Exact prior information

7.1.1. Prior information

7.1.2. Posterior probabilities with exact prior information

7.1.3. Influence of prior information in posterior probabilities

7.2. Vague prior information

7.2.1. A vague definition of vague prior information

7.2.2. Examples of the use of vague prior information

7.3. No prior information

7.3.1. Flat priors

7.3.2. Jeffrey’s priors

7.3.3. Bernardo’s “Reference” priors

7.4. Improper priors

7.5. The Achilles heel of Bayesian inference

Appendix 7.1

Appendix 7.2

CHAPTER 1

DO WE UNDERSTAND CLASSICAL STATISTICS?

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

1.1. Historical introduction
1.2. Test of hypothesis
    1.2.1. The procedure
    1.2.2. Common misinterpretations
1.3. Standard errors and Confidence intervals
    1.3.1. The procedure
    1.3.2. Common misinterpretations
1.4. Bias and Risk of an estimator
    1.4.1. Unbiased estimators
    1.4.2. Common misinterpretations
1.5. Fixed and random effects: a permanent source of confusion
    1.5.1. Definition of "fixed" and "random" effects
    1.5.2. Bias, variance and Risk of an estimator when the effect is fixed or random
    1.5.3. Common misinterpretations
1.6. Likelihood
    1.6.1. Definition
    1.6.2. The method of maximum likelihood
    1.6.3. Common misinterpretations

Appendix 1.1
Appendix 1.2
Appendix 1.3

1. Historical introduction

The Bayesian School was, in practice, founded by the French aristocrat and politician Count Pierre Simon Laplace through several works published from 1774 to 1812, and it had a preponderant role in scientific inference during the nineteen century (Stigler, 1986). Some years before Laplace's first paper on the matter, the same principle was formalized in a posthumous paper presented at the Royal Society of London and attributed to a rather obscure priest, rev. Thomas Bayes (who never published a mathematical paper during his life). Apparently, the principle upon which Bayesian inference is based was formulated before. Stigler (1983) attributes it to Saunderson (1683-1739), a blind professor of optics who published a large number of papers on several fields of mathematics. Due to the work of Laplace, the greatest statistician of the short history of this rather new subject, Bayesian techniques were commonly used along the 19th and the first decades of the 20th century. For example, the first deduction of the least square method made by Gauss in 1795 (although published in 1809) was made using Bayesian theory. At this time these techniques were known as "inverse probability", because their objective was to find the probability of the causes from their effects (i.e.: to estimate the parameters from the data). The word "Bayesian" is rather recent (Fienberg, 2006), and it was introduced by Fisher (1950) to stress the precedence in time (not in importance) of the work of rev. Bayes. As Laplace ignored the work of Bayes, I think the correct name for this school should be "Laplacian", or perhaps we should conserve the old and explicit name of "inverse probability". Nevertheless, as it is common today, we will use the name "Bayesian" along this book.

Bayesian statistics uses probability to express uncertainty about the unknowns that are being estimated. The use of probability is more efficient than any other method of expressing uncertainty. Unfortunately, to make it possible, inverse probability needs the knowledge of some prior information. For example, we know by published

literature that in many countries the pig breed Landrace has a litter size of 10 piglets around. We perform an experiment to know Spanish Landrace litter size, and a sample of 5 sows is evaluated for litter size in their first parity. Say an average of 5 piglets born is observed. This seems a very unlikely outcome *a priori*, and we should not trust very much our sample. However, it is not clear how to integrate properly the prior information about Landrace litter size in our analysis. We can pretend we do not have any prior information and say that all the possible results have the same *prior* probability, but this leads to some inconsistencies as we will see later in chapter 7. Laplace was aware about this problem, and in his later works he examined the possibility of making inferences based in the distribution of the samples rather than on the probability of the unknowns, founding in fact the frequentist school (Hald, 1998).

Fisher's work on the likelihood in the 20's and the frequentist work of Neyman and Pearson in the 30's and Wald in the 40's eclipsed Bayesian inference. The reason was that they offered inferences and measured uncertainty about these inferences without needing prior information. Fisher developed the properties of the method of maximum likelihood, a method which is attributed to him although Daniel Bernouilli proposed it as early as in 1778 (see Kendall, 1961). When Fisher discussed this method (Fisher, 1935) one of the discussant of his paper noticed that Edgeworth had proposed it in 1908. Probably Fisher did not know this paper when he proposed to use the likelihood in an obscure paper published in 1912 ([1]), when he was 22 years old, and he never cited any precedent. His main contribution was to determine the statistical properties of the likelihood and to develop the concept of information based on it. Neyman and Pearson (1933) used the likelihood ratio as a useful way to perform hypothesis tests. Their theory was based in considering hypothesis tests as a decision problem, choosing between a hypothesis or an alternative, a procedure that Fisher disagreed. Fisher preferred to consider that when a null hypothesis is not rejected we cannot say that this hypothesis is true, but just to take this hypothesis as provisional, very much in the same sense of Popper theory of refutation (although expressed before).

[1] The historian of statistics A. Hald (1998) asks himself why such an obscure paper passed the referee's report and was finally published. At that time Fisher did not use the name of "likelihood".

Although some researchers still used Bayesian methods in the 30's, like the geologist and statistician Harold Jeffreys, the classical Fisher-Neyman-Pearson school dominated the statistical world until the 60's, when a 'revival' started and has been increasing hitherto. Bayesian statistics had three problems to be accepted, two main theoretical problems and a practical one. The first theoretical problem was the difficulty of integrating prior information. To overcome this difficulty, Ramsey (1926) and De Finetti (1937) proposed separately to consider probability as a "belief". Prior information was then evaluated by experts and probabilities were assigned to different events according to their expert opinion; this can work for few traits or effects, but has serious problems in the multivariate case. The other theoretical problem is how to represent "ignorance" because there is no prior information, because we do not like the way in which Ramsey and De Finetti integrates this prior information, or because we would like to assess the information provided by the data without prior considerations. This second theoretical problem is still the main difficulty for many statisticians to accept Bayesian theory, and it is nowadays an area of intense research. The third problem comes from the use of probability to express uncertainty, a characteristic of the Bayesian School. As we will see later, this leads to multiple integrals that cannot be solved even by using approximate methods. This was a main problem to apply Bayesian techniques until the 90's, when a numerical method was applied to find practical solutions to all these integrals. The method, called Monte Carlo Markov Chains (MCMC) allowed to solve these integrals contributing to the current rapid development and application of Bayesian techniques in all fields of science.

Bayesian methods express the uncertainty using probability density functions that we will see in chapter 3. The idea of MCMC is to provide a random sample of a probability function instead of using its mathematical equation. It is considered that the first use of random samples of a density function is due to the Guiness brewer William Searly Gosset (1908), called "Student", in his famous paper in which he presented the t-distributions. The use of Markov chains to find these random samples has its origins in the Los Alamos project for constructing the first atomic bomb. The name "Monte Carlo" was used as a secret code for the project, referring to the Monte Carlo roulette as a kind of random sampling. Metropolis and Ulam (1949) published

the first paper describing MCMC, Geman and Geman (1986) applied this method to image analysis using a particularly efficient type of MCMC that they called “Gibbs sampling” because they were using Gibbs distributions (if they were using normal distributions the method would have been called “normal sampling”), Gelfand and Smith (1990) introduced this technique in the statistical world to obtain probability distributions and Daniel Gianola (Wang et al., 1994) and Daniel Sorensen (Sorensen et al. 1994) brought these techniques into the field of animal breeding. These techniques present several numerical problems and there is a very active area of research in this field to solve them. Today, the association of Bayesian inference and MCMC techniques has produced a dramatic development of the application of Bayesian techniques to practically every field of science.

1.2. Test of Hypothesis

1.2.1. *The procedure*

Let us start with a classical problem: We have an experiment in which we want to test whether there is an effect of some treatment; for example, we are testing whether a selected population for growth rate has a higher growth rate than a control population. In classical statistics, the hypothesis to be tested is that there is no difference between the two treatments; i.e., the difference in growth between the selected and the control group is null. The classical procedure is to establish, *before* making the experiment, the error of rejecting this hypothesis when it is actually true, i.e., the error of saying that there is a difference between selected and control groups when actually there is not. Traditionally, this error, called error Type I, is fixed at a level of 5%, which means that *if the null hypothesis is true* (there is no difference between treatments), repeating an experiment an infinite number of times we can get an infinite number of samples of the selected and control groups, and the difference between the averages of these samples ($\bar{x}_1 - \bar{x}_2$) will be grouped around zero, which is the true value of the difference between selected and control populations (m_1-m_2) (see Figure 1.1). However we do not have money and time to take an infinite number of samples, thus we will only take *one* sample. If our sample lies in the shadow area

of figure 1, we can say that:

1) There is no difference between treatments, and our sample was a very rare sample that only will occur a 5% of times as a maximum if we repeat the experiment an infinite number of times, or

2) The treatments are different, and repeating an infinite number of times the experiment, the difference between the averages of the samples ($\bar{x}_1 - \bar{x}_2$) will not be distributed around zero but around an unknown value different from zero.

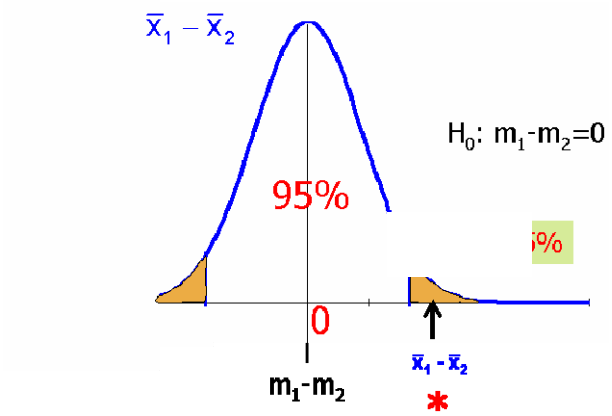

Figure 1.1. Distribution of repeated samples if H_0 is true. When our actual difference between sample averages lies in the shadow area we reject H_0 and say that the difference is "significant". This is often represented by a star.

Neyman and Pearson (1933) suggested that the "scientific behaviour" should be to take option 2 acting as if the null hypothesis was wrong. A result of our *behaviour* will be that 'in the long run' we will be right almost in a 95% of the cases.

There is some discussion in the classical statistical world about what to do when we do not reject the null hypothesis. In this case we can say that we do not know whether the two treatments are different ([2]), or we can accept that both treatments

[2] This attitude to scientific progress was later exposed in a less technical way by Karl Popper (1936). Scientists and philosophers attribute to Popper the theory about scientific progress based in the refutation of pre-existing theories, whereas accepting the current theory is always provisional. However Fisher (1935) based his testing hypothesis theory exactly in the same principle. I do not

have the same effect, i.e. that the difference between treatments is null. Fisher (1925) defended the first choice whereas Neyman and Pearson (1933) defended the second one stressing that we also have a possible error of being wrong in this case (they called it Type II error to distinguish it from the error we managed before).

1.2.2. *Common misinterpretations*

The error level is the probability of being wrong: It is not. We choose the error level *before* making the experiment, thus a small size or a big size experiment may have the same error level. After the experiment is performed, we *behave* accepting or rejecting the null hypothesis as if we had Probability = 100% of being right, hoping to be wrong a small number of times along our career.

The error level is a measure of the percentage of times we will be right: This is not true. You may accept an error level of a 5% and find along your career that your data were always distributed far away from the limit of the rejection (figure 1.2).

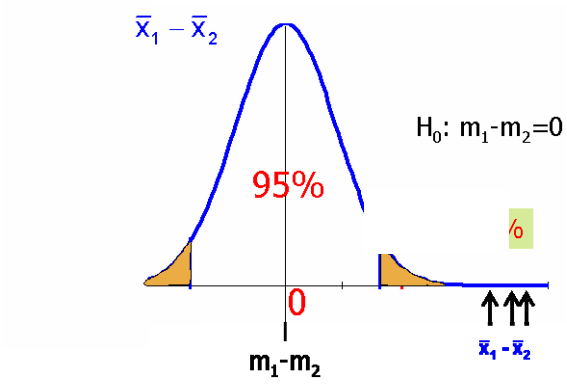

Figure 1.2. An error level of 5% of being wrong when rejecting the null hypothesis was accepted, but along his career, a researcher discovered that his data showed much higher evidence about the null hypothesis being wrong

The *P-value* is a measure of the "significance": This is not true. Modern computer programs give the tail of probability calculated from the sample that is analyzed. The

---

know how far backwards can be traced the original idea, but it is contained at least in the famous essay "On liberty" of James Stuart Mill (1848).

area of probability from the sample to infinite (shadow area in Figure 1.3) gives the probability of finding the current sample or a higher value when the null hypothesis is true. However, a *P-value* of 2% does not mean that the difference between treatments is "significant at a 2%", because if we repeat the experiment we will find another *P-value*. We cannot fix the error level of our experiment depending on our current result because we drive conclusions *not only from our sample but also from all possible repetitions of the experiment* that we have not performed (and we do not have the slightest intention to perform) ([3]).

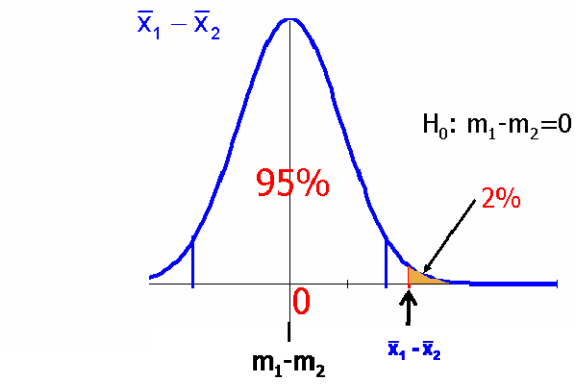

Figure 1.3. A *P-value* of 2% gives the probability of finding the current sample or a higher value if the null hypothesis holds. However this does not mean that the difference between treatments is "significant at a 2%".

There is a frequent tendency to think that if a P-value is small, when we will repeat the experiment it will still be small. This is not necessarily true. For example, if we obtain a P-value of 5% and the TRUE value is the same as the value obtained in our sample, when repeating the experiment half of the samples will give a significant value (P<0.05) and half of them will not (P>0.05) (figure 1.4). Of course, if the true value is much higher, only few samples will give a significant difference when

[3] Fisher insisted many times in that P-values can be used to reject the null hypothesis but not to accept any alternative (about which no probability statements are made), and clearly established the difference between his way of performing hypothesis tests and the way of Neyman and Pearson (see, for example, Fisher, 1956). On their side, Neyman and Pearson never used P-values because P-values do not mean anything in their way or accepting or rejecting hypothesis; rejection areas are established *before* making the experiment, and a P-value is an area obtained after the experiment is performed. However, it is noticeable how both Neyman and Pearson significance and Fisher P-values are now blended in modern papers just because now P-values are easy to compute. In general, I think that they create more confusion than help in understanding results.

repeating the experiment, but in this case it is unlikely to find a P-value near 5% in a previous essay. We do not know where the true value is, thus we do not know whether we are in the situation of figure 1.4 or in other situation.

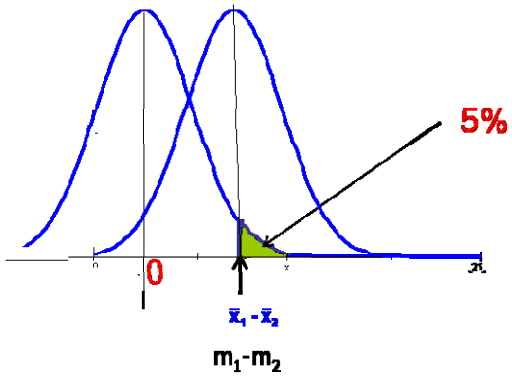

Figure 1.4. Distribution of the samples when the true value is the same as the actual sample. The current sample gives us a P- value of 5%. If the true value is the same as our sample, notice that when repeating the experiment, half of the times we will obtain "non significant " values (P>0.05).

Obviously a low P-value shows a higher evidence for rejecting the null hypothesis than a high P-value, but it is not clear *how much evidence* it provides. A P-value gives the probability of finding the current sample *or a higher value*, but we are not interested in how probable is to find our sample, but in how probable our hypothesis is, and to answer to this question we need to use prior probabilities, as we will see in chapter 2. Fisher said that a P-value of 0.05 shows either that the null hypothesis is not true or a rare event happened. How rare is the event is not clear. For example, Berger and Sellke (1987) show an example in which under rather general conditions, a P-value of 0.05 corresponds to a probability of the null hypothesis being true of a 23%, far more than the 5% suggested by the P-value. A conscious statistician knows what a P-value means, but the problem is that P-values suggest to the average researcher that they have found more evidence that they actually have, and they tend to believe that this 5% given by a P-value is the probability of the null hypothesis being right, which is not.

According to the procedure of classical statistics, we cannot use P-values as a *measure of significance* because the error level is defined *before* the experiment is performed (at least before it is analyzed). Thus, performing the classic procedure *we do not have any measure of how much evidence we have for rejecting the null hypothesis*, and this is one of the major flaws of classical statistics.

Modern statisticians are trying to use *P-values* to express the amount of evidence the sample gives, but there is still a considerable discussion and no standard methods are hitherto implemented (see Sellke et al., 2001 and Bayarri and Berger, 2004, for a discussion).

Significant difference means that a difference exists. This is not always true. We may be wrong one each twenty times as an average, if the error level is a 5%. The problem is that when measuring many traits, we may detect a false significant difference once each twenty traits (^{4}). The same problem arises when we are estimating many effects. It is not infrequent to see pathetic efforts of some authors for justifying some second or third order interaction that appears in an analysis when all the other interactions are not significant, without realising that this interaction can be significant just by chance.

N.S. (non significant difference) means that there is no difference between treatments. This is usually false. First, treatments are always different because they are not going to be *exactly equal.* A pig selected population can differ from the control in less than a gram of weight at some age, but this is obviously irrelevant. Second, in well designed experiments, N.S. appears when the difference between treatments is irrelevant, but this only happens for the trait for which the experiment was designed, thus all other measured traits can have relevant differences between treatments whereas we still obtain N.S. from our tests. The safest interpretation of N.S. is "we do not know whether treatments differ or not"; this is Fisher's interpretation for N.S.

Our objective is to find whether two treatments are different. We are not interested in finding whether or not there are differences between treatments because they are not going to be *exactly* equal. Our objective in an experiment is to find **relevant** differences. How big should be a difference in order to consider it as *relevant* should be defined before making the experiment. A relevant value is a quantity

[4] Once each twenty traits as a maximum if the traits are uncorrelated. If they are correlated the frequency of detecting false significances is different.

under which differences between treatments have no biological or economical meaning. In classical statistics, the size of the experiment is usually established for finding a significant difference between two treatments when this difference is considered to be relevant.

Significant difference means Relevant difference: This is often false. What is true is that if we have a good experimental design, a *significant* difference will appear just when this difference is *relevant.* Thus, if we consider that 100 g/d is a relevant difference between a selected and a control population, we will calculate the size of our experiment in order to find a significant difference when the difference from the averages of our samples $|\bar{x}_1 - \bar{x}_2| \geq 100$ g/d, and we will not find a significant difference if it is lower than this. The problem arises in field data, where no experimental design has been made, in poorly designed experiments and in well designed experiments when we analyze other trait than the trait used to find the size of the experiment. In these cases there is no link between the *relevance* of the difference and its *significance*, and we can find:

1) **Significant differences that are completely irrelevant**: This first case is innocuous, although if *significance* is confused with *relevance*, the author of the paper will stress this result whit no reason. *We will always get significant differences if the sample is big enough.* Thus 'significance' itself is of little value.

2) **Non significant differences that are relevant**: This means that the size of the experiment is not high enough. Sometimes experimental facilities are limited because of the nature of the experiment, but a conscious referee should reject for publication "N.S." differences that are relevant.

3) **Non significant differences that are irrelevant, but have high errors**: Sometimes the estimation we have can be, by chance, near zero, but if the standard error of the estimation is high this means that when repeating the experiment, the difference may be much higher and relevant. For example, if a relevant difference for growth rate is 100g/d in pigs and the difference between the selected and control populations is 10 g/d with a s.e. of 150 g/d, when repeating the experiment we may find a difference higher than 100g/d; i.e., we

can get a relevant difference. Thus, a “N.S.” difference should not be interpreted as “there is no relevant difference” unless the precision of this difference is good enough.

4) **Significant differences that are relevant, but have high errors:** This may lead to a dangerous misinterpretation. Imagine that we are comparing two breeds of rabbits for litter size. We decide that one kit will be enough to consider the difference between breeds to be relevant. We obtain a significant difference of 2 kits with a risk of a 5% (we got one ‘star’). However, the confidence interval at a 95% probability of this estimation goes from 0.1 to 3.9 kits. Thus, we are not sure about whether the difference between breeds is 2 kits, 0.1 kits, 0.5 kits, 2.7 kits or whatever other value between 0.1 and 3.9. It may happen that the true difference is 0.5 kits, which is irrelevant. However, typically, all the discussion of the results is organised around the 2 and the ‘star’. We will typically say that ‘we found significant and important differences between breeds’, although we do not have this evidence. The same applies when comparing our results with other published results; typically the standard errors of both results are ignored when discussing similarities or dissimilarities.

We always know what a relevant difference is. Actually, for some problems we do not know: a panel of expertises analyse the aniseed flavour of some meat and they find significant differences of three points in a scale of ten points, is this relevant? Which is the relevant value for enzyme activities? Sometimes it is difficult to precise which the relevant value is, and in this case we are completely disoriented when we are interpreting the tables of results, because in this case we cannot distinguish between the four cases we have listed before. In appendix 1.1 I propose some practical solutions to this problem.

Tests of hypothesis are always needed in experimental research. I think that for most biological problems we do not need any hypothesis test: The answer provided by a test is rather elementary: *Is there a difference between treatments?* YES or NOT. However this is not actually the question for most biological problems. In fact, we know that the answer to this question is always YES, because two treatments are not going to be *exactly equal*. Thus, usually our question is whether these treatments

differ in more than a relevant quantity. To answer to this question we should estimate the difference between treatments accompanied by a measurement of our uncertainty. I think that the common practice of presenting results as LS-means and levels of significance or *P-values* should be substituted by presenting differences between treatments accompanied by their uncertainty expressed as confidence intervals when possible.

1.3. Standard errors and Confidence intervals

1.3.1. *The procedure*

If we take an infinite number of samples, the sample averages (or the difference between two sample averages) will be distributed around the true value we want to estimate, as in Figure 1. The standard deviation of this distribution is called “standard error” (s.e.), to avoid confusion with the standard deviation of the population. A large standard error means that the sample averages will take very different values, many of them far away from the true value. As we do not take infinite samples, but just one, a large standard error means that we do not know whether we are close or not to the true value, but a small standard error means that we are close to the true value because most of the possible sample averages when repeating the experiment (conceptually, which means imaginary repetitions) will be close to the true value.

When the sampling distribution is Normal ([5]), about twice the standard error around the true value will contain a 95% of the sample averages. This permits the construction of the so-called Confidence Intervals at 95% by establishing the limits within the true value is expected to be found. Unfortunately, we do not know the true value, thus it is not possible to establish confidence intervals as in Figure 1, and we have to use our estimate instead of the true value to define the limits of the confidence interval. Our confidence interval is (sample average ± 2 s.e.). A consequence of this way of working is that each time we repeat the experiment we

[5] Computer programs (like SAS) ask about whether you have checked the normality of your data, but normality of the data is not needed if the sample is large enough. Independently of the distribution of the original data, the average of a sample is distributed normally, if the simple size is big enough. This is often forgotten, as Fisher complained (Fisher 1925).

have a new sample average (a new “estimate of the true value”) and thus a new confidence interval.

For example, assume we want to estimate the litter size of a pig breed and we obtain a value of 10 with a confidence interval with a 95% of probability C.I.(95%)=[9, 11]. This means that if we repeat the experiment, we will get many confidence intervals: [8, 10], [9.5, 11.5] … etc. and a 95% of these intervals will contain the true value. However we are not really going to repeat the experiment an infinite number of times, and thus we only have got one interval! What shall we do? In classical statistics we *behave as* if our interval would be one of the intervals containing the true value. We hope, *as a consequence of our behaviour*, to be wrong a maximum of a 5% of times along our career.

1.3.2. *Common misinterpretations*

The true value is between ± s.e. of the estimate: We do not know whether this happens or not. First, the distribution of the samples when repeating the experiment might be not normal as it is in Figure 1. This is common when estimating correlation coefficients and they are close to 1 or to -1. Part of the problem is the foolish notation that scientific journals admit for s.e. It is nonsense to write a correlation coefficient as 0.95 ± 0.10. Modern techniques (for example, bootstrap) taking advantage of easy computation with modern computers can show the actual distribution of a sample. A correlation coefficient sampling distribution may be asymmetric, like in figure 4. If we take the most frequent value as our estimate (-0.9), the s.e. has little meaning.

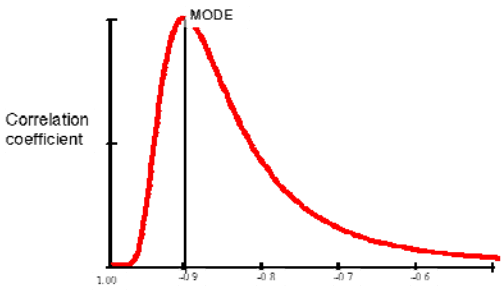

Figure 4. Sampling distribution of a correlation coefficient. Repeating the experiment, the samples are not distributed symmetrically around the true value**.**

A C.I. (95%) means that the probability of the true value to be contained in the

interval is a 95%: This is not true. We say that the true value is contained in the interval with probability P=1, i.e., with total certainty. We utter that our interval is one of the "good ones" (figure 5). We may be wrong, but we *behave* like this and we hope to be wrong only a 5% of times as a maximum along our career. As in the case of the test of hypothesis, we make inferences not only from our sample but from the distribution of samples in ideal repetitions of the experiment.

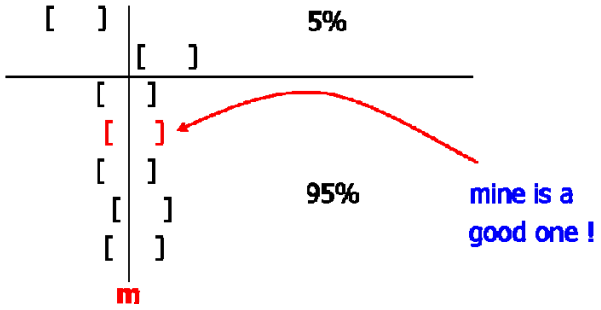

Figure 5. Repeating the experiment many times, a 95% of the intervals will contain the true value m. We do not know whether our interval is one of these, but we assume that it is. We hope not to be wrong many times along our career

Conceptual repetition leads to paradoxes: Several paradoxes produced by drawing conclusions not only from our sample but from conceptual repetitions of it have been noticed. The following one can be found in Berger and Wolpert (1982).

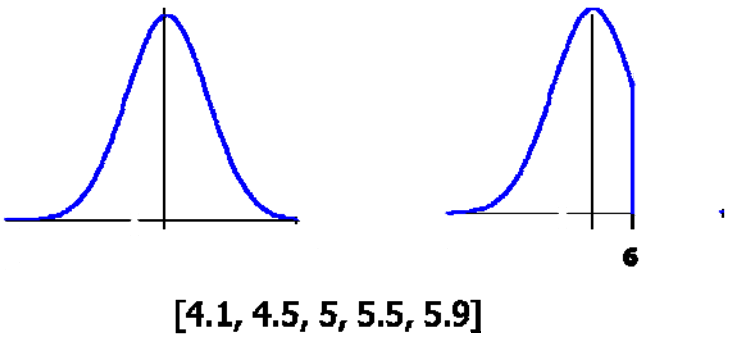

Figure 6. Repeating an experiment an infinite number of times we arrive to different conclusions if our pH-meter is broken or not, although all our measurements were correctly taken.

Imagine we are measuring a pH and we know that the estimates will be normally distributed around the true value when repeating the experiment an infinite number of times. We obtain a sample with five measurements: 4.1, 4.5, 5, 5.5 and 5.9. We then calculate our CI 95%. Suddenly, a colleague tells us that the pH-meter was broken

and it could not measure a pH higher than six. Although we did not find any measure higher than six and then all the measurements we took were correct, if we repeat the experiment an infinite number of times we will obtain a truncated distribution of our samples (figure 6). This means that we should change our confidence interval, since all possible samples higher than 6 would be recorded as 6. Then another colleague tells us that the pH-meter was repaired before we started our experiment, and we write a paper changing the CI 95% to the former values. But our former colleague insists in that the pH-meter was still broken, thus we change again our CI.

Notice that we are changing our CI *although none of our measurements led in the area in which the pH-meter was broken.* We change our CI not because we had wrong measures of the pH, but because *if we would repeat the experiment an infinite number of times* this will produce a different distribution of our samples. As we make inferences not only from our samples, but from imaginary repetitions of the experiment (that we will never perform), our conclusions are different if the ph-meter is broken although all our measurements were correct.

1.4. Bias and Risk of an estimator

1.4.1. *Unbiased estimators*

In classical statistics we call *error of estimation* to the difference between the true value u and the estimated value û

$$e = u - \hat{u}$$

We call *loss function* to the square of the error

$$l(\hat{u},u) = e^2$$

and we call Risk to the mean of the losses([6])

[6] All of this is rather arbitrary and other solutions can be used. For example, we may express the error as a percentage of the true value, the loss function may be the absolute value of the error instead of

$$R(\hat{u},u) = E[l(\hat{u},u)] = E(e^2)$$

A good estimator will have a low risk. We can express the risk as

$$R(\hat{u},u) = E(e^2) = E(\bar{e}^2 + e^2 - \bar{e}^2) = E(\bar{e}^2) + E(e^2 - \bar{e}^2) = \bar{e}^2 + var(e) = Bias^2 + var(e)$$

where we define Bias as the mean of the errors $\bar{e}$. An *unbiased estimator* has a null bias. This property is considered particularly attractive in classical statistics, because it means that when repeating the experiment an infinite number of times, the estimates are distributed around the true value like in Figure 1. In this case the errors are sometimes positive and sometimes negative and their mean is null (and so is its square).

1.4.2. *Common misinterpretations*

A transformation of an unbiased estimator leads to another unbiased estimator: This is often not true. It is frequent to find researchers that carefully obtain unbiased estimators for the variance and then use them to estimate the standard deviation by computing their square root. For example, people working with NIR (near infrared spectroscopy, an analytical method) estimate the variance of the error of estimation by using unbiased estimators, and then they calculate the standard error by computing the square root of these estimates. However, *the square root of an unbiased estimator of the variance is not an unbiased estimator of the standard deviation*. It is possible to find unbiased estimations of the standard deviation, but they are not the square root of the unbiased estimator of the variance (see for example Kendall et al., 1992). Fisher considered, from his earliest paper (Fisher, 1912) that the property of unbiasedness was irrelevant due to this lack of invariance to transformations.

Unbiased estimators should be always preferred: Not always. As the Risk is the sum of the bias plus the variance of the estimator, it may happen that a biased

---

its square and the risk might be the mode instead of the mean of the loss function, but in this chapter we will use these definitions.

estimator has a lower risk, and thus it is a better estimator than another unbiased estimator (figure 7).

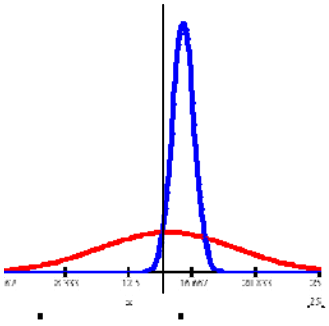

Figure 7. A biased estimator (blue) is not distributed around the true value but has lower risk than an unbiased estimator (red) that is distributed around the true value with a much higher variance.

For example, take the case of the estimation of the variance. We can estimate the variance as

$$\hat{\sigma}^2 = \frac{1}{k}\sum_{1}^{n}(x_i - \bar{x})^2$$

It can be shown that the bias, variance and risk of this estimator are

$$\mathrm{BIAS}(\hat{\sigma}^2) = \sigma^2 - \frac{n-1}{k}\sigma^2 \qquad \mathrm{var}(\hat{\sigma}^2) = \frac{2(n-1)}{k^2}\sigma^4$$

$$\mathrm{RISK}(\hat{\sigma}^2) = \mathrm{BIAS}^2 + \mathrm{var} = \left(\sigma^2 - \frac{n-1}{k}\sigma^2\right)^2 + \frac{2(n-1)}{k^2}\sigma^4$$

depending on the value of k we obtain different estimators. For example, to obtain the estimator of minimum risk, we derive the Risk respect to k, equal to zero and obtain a value of k = n+1. But there are other common estimators for other values of k. When k=n we obtain the maximum likelihood (ML) estimator, and when k=n–1 we obtain the residual (or restricted) maximum likelihood estimator (REML) (see Blasco 2001). Notice that when k=n–1 the estimator is unbiased, which is a favourite reason of REML users to prefer this estimator. However, the Risk of REML is higher than the risk of ML because its variance is higher, thus ML should be preferred… or even

better, the minimum risk estimator (that nobody uses).

1.5. Fixed and random effects

1.5.1. *Definition of "fixed" and "random" effects*

Churchill Eisenhart proposed in 1941 a distinction between two types of effects. The effect of a model was "*fixed*" if we were interested in its particular value and "*random*" if it could be considered just one of the possible values of a random variable. Consider, for example, an experiment in which we have 40 sows in four groups of 10 sows each, and we feed each group with a different food. We are interested in knowing the effect of each food in the litter size of the sows, and then each sow has five parities. The effect of the food can be considered as a "fixed" effect, because we are interested in finding the food that leads to higher litter sizes. We also know that there some sows are more prolific than other sows, but we are not interested in the prolificacy of a particular sow, we consider that each sow effect is a "random" effect. When repeating an experiment an infinite number of times, the fixed effect always has the same values, whereas the random effect changes in each repetition of the experiment. When repeating our experiment, we will always give the same four foods, but the sows will be different; the effect of the food will be always the same but the effect of the sow will randomly change in each repetition.

In Figure 8 we can see how the true value of the effects and their estimates are distributed. When repeating the experiment, the true value of the fixed effect remains constant and all its estimates are distributed around this unique true value. In the case of the random effect, each repetition of the experiment leads to a new true value, thus the true value is not constant and it is distributed around its mean.

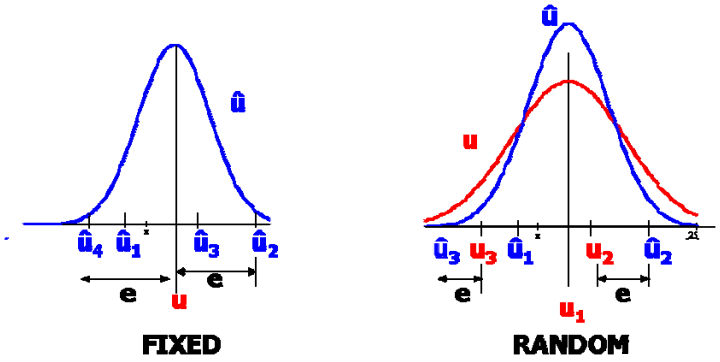

Figure 8. Distribution of the effects and their estimates when repeating the experiment an infinite number of times. When the effects are fixed the true value is constant, but when the effect is random it changes its value in each repetition. In red, the distribution of the true values; in blue, the distribution of the estimates.

1.5.2. *Bias, variance and Risk of an estimator when the effect is fixed or random*

By definition, bias is the mean of the errors,

FIXED BIAS = E(e) = E(u–û) = E(u) – E(û) = u – E(û)

RANDOM BIAS = E(e) = E(u–û) = E(u) – E(û)

In the case of *fixed* effects, as the true value is constant, u = E(u) and when the estimator is *unbiased*, the estimates are distributed around the true value. In the case of *random* effects the true value is not constant and when the estimator is *unbiased* the average of the estimates will be around the average of the true values, a property which is much less attractive ([7]).

The variances of the errors are also different

FIXED: var(e) = var(u – û) = var(û)

[7] Henderson (1973) has been critiqued for calling the property E(u) = E(û) "unbiasedness", in order to defend that his popular estimator "BLUP" was unbiased. This property should always mean that the estimates are distributed around the true value. In the case of random effects this means u= E(û|u), a property that BLUP does not have (see Robinson, 1981).

RANDOM: var(e) = var(u – û) = var(u) – var(û)

In the case of fixed effects, as the true value is a constant, var(u) = 0, then the best estimators are the ones with smallest variance var(û) because this variance is the same as the variance of the error, which is the one we want to minimize. In the case of random effects the true values have a distribution and the variance of the error is the difference between the variance of the true values and the variance of their estimator (see Appendix 1.2 for a demonstration). Thus, the best estimator is the one with a variance as big as the variances of the true values. An estimator with small variance is not good because its estimates will be around its mean E(û) and the errors will be high because the true value changes in each repetition of the experiment (see figure 8). Moreover, its variance cannot be higher than the variance of the true value and the covariance between u and û is positive (see Appendix 1.2).

The source of the confusion is that a good estimator is not the one with small variance, but the one with *small error variance*. A good estimator will give values close to the true value in each repetition, the error will be small, and the variance of the error also small. In the case of fixed effects this variance of the error is the same as the variance of the estimator and in the case of random effects the variance of the error is small when the variance of the estimator is close to the variance of the true value ([8]).

1.5.3. *Common misinterpretations*

An effect is fixed or random due to its nature: This is not true. In the example before, we might have considered the four types of foods as random samples of all different types of food. Thus, when repeating the experiment, we would change the food (we should not be worried about this because we are not going to repeat the experiment; all are “conceptual” repetitions). Conversely, we might have considered

[8] In quantitative genetics, this is a well known property of selection indexes. In appendix 1.3 there is an example of its use in selection.

the sow as a “fixed” effect and we could have estimated it, since we had five litters per sow ([9]). Thus the effects can be fixed or random depending on what is better for us when estimating them.

We are not interested in the particular value of a random effect: Sometimes we can be interested in it. A particular case in which it is interesting to consider the effects as random is the case of genetic estimation. We know the covariances of the effects of different relatives, thus we can use this prior information if the individual genetic effects are considered as random effects. We have smaller errors of estimation than considering the genetic effects as fixed.

Even for random effects to be unbiased is an important property: The property of unbiasedness is not attractive for random effects, since repeating the experiment the true values also change and the estimates are not distributed around the true value.

Random effects have always lower errors than fixed effects: We need good prior information. We still need to have a good estimation of the variance of the random effect. This can come from the literature or from our data, but in this last case we need data enough and the errors of estimation are high when having few data.

BLUP is the best possible estimator: As before, we can have biased estimators with higher risk as unbiased estimators. The reason for searching estimators with minimum variance (“best”) among the unbiased ones is because there are an infinite number of possible biased estimators with the same risk, depending on their bias and their variance. By adding the condition of unbiasedness, it can be found a single estimator, called “BLUP”.

1.6. Likelihood

[9] Fisher, considered that the classification of the effects in fixed and random was worse than considering all the effects as random, as they were considered before (Fisher, 1956).

1.6.1. *Definition*

The concept of likelihood and the method of maximum likelihood (ML) were developed by Fisher between 1912 and 1922, although there are historical precedents attributed to Bernouilli (1778, translated by C.G. Allen, see Kendall, 1961). By 1912 the theory of estimation was in an early state and the method was practically ignored. However, Fisher (1922) published a paper in which the properties of the estimators were defined and he found that this method produced estimators with good properties, at least asymptotically. The method was then accepted by the scientific community and it is now frequently used.

To arrive to the concept of likelihood, I will put an example of Blasco (2001). Consider finding the average weight of rabbits of a breed at 8 wk of age. We take a sample of one rabbit, and its weight is y_0 = 1.6 kg. The rabbit can come from a population normally distributed which mean is 1.5 kg, or from other population with a mean of 1.8 kg or from other possible populations. Figure 9 shows the density functions of several possible populations from which this rabbit can come, with population means m_1=1.50 kg, m_2= 1.60 kg, m_3= 1.80 kg. Notice that, at the point y_0, the probability density of the first and third population $f(y_0|m_1)$ and $f(y_0|m_3)$ are lower than the second one $f(y_0|m_2)$. It looks very *unlikely* that a rabbit of 1.6 kg comes from a population which mean is 1.8 kg. Therefore, it seems more *likely* that the rabbit comes from the second population.

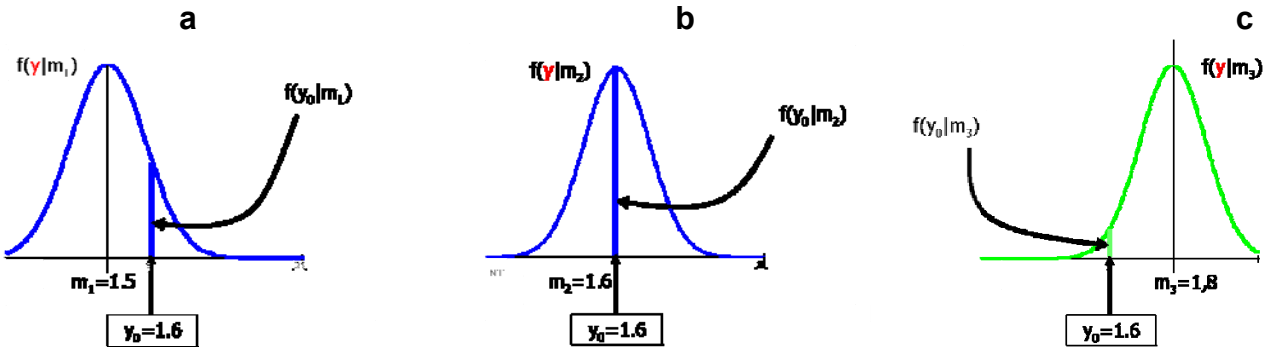

Figure 9. Three likelihoods for the sample y_0= 1.6. a: likelihood if the true mean of the population would be 1.5, b: likelihood if the true mean of the population would be 1.6. c: likelihood if the true mean of the population would be 1.8

All the values $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$, … define a curve with a maximum in $f(y_0|m_2)$ (Figure 10).

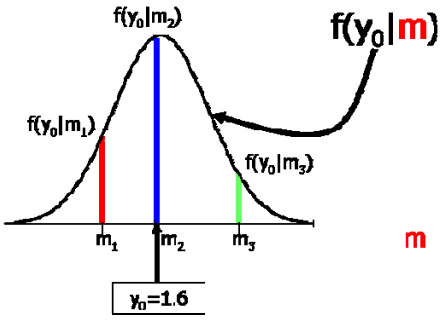

Figure 10. Likelihood curve. It is not a probability because its values come from different probability distributions, bur it is a rational degree of belief. The notation stress that the variable (in red) is m and not y_0 that is a given fixed sample.

This curve varies with m, and the sample y_0 is a fixed value for all those density functions. It is obvious that the new function defined by these values is *not* a density function, since each value belongs to a different probability density function.

We have here a problem of notation, because here the variable is 'm' instead of 'y', because we have fixed the value of $y=y_0=1.6$. Speaking about a set of density functions $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$… for a given y_0 is the same as speaking about a function $L(m|y_0)$ that is not a density function ([10]). However this notation hides the fact that $L(m|y_0)$ is a family of density functions indexed at a fixed value $y=y_0$. We will use a new notation, representing the variable in red colour and the constants in black colour. Then $f(y_0|m)$ means a family of density functions in which the variable is m that are indexed at a fixed value y_0. For example, if these normal functions of our example are standardized (s.d. = 1), then the likelihood will be represented as

$$\mathbf{f}(y_0 \mid m) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_0 - m)^2}{2}\right] = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(1.6 - m)^2}{2}\right]$$

where the variable is in red colour. We will use 'f' exclusively for density functions in a generic way; i.e., f(x) and f(y) may be different functions (Normal or Poisson, for example), but they will be always density functions.

[10] Classic texts of statistics like the Kendall's one (Kendall et al., 1998) contribute to the confusion by using the notation L(y|m) for the likelihood. Moreover, some authors distinguish between "given a parameter" (always fixed) and "giving the data" (which are random variables). They use (y|m) for the first case and (m;y) for the second. Thus, likelihood can be found in textbooks as L(m|y), L(y|m), f(y|m) and L(m;y).

1.6.2. *The method of maximum likelihood*

Fisher (1912) proposed to take the value of m that maximized $f(y_0|m)$ because from all the populations defined by $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$, … this is the one that *if this were the true value* the sample would be most probable. Here the word *probability* can lead to some confusion, since these values belong to different density functions and the likelihood function defined taking all of these values is not a probability function. Thus, Fisher preferred to use the word *likelihood* for all these values considered together ([11]).

Fisher (1912, 1922) not only proposed a method of estimation, but also proposed the likelihood as a *degree of belief* different from the probability but allowing to express uncertainty in a similar manner. What Fisher proposed is to use the whole likelihood curve and not only its maximum, a practice rather unusual. Today, frequentist statisticians typically use only the maximum of the curve because it has good properties in repeated sampling (figure 11). Repeating the experiment an infinite number of times, the estimator will be distributed near the true value, with a variance that can also be estimated. But all those properties are asymptotic and thus there is no guarantee about the goodness of the estimator when samples are small. Besides, the ML estimator is not necessarily the estimator that minimizes the risk. Nevertheless, the method has an interesting property apart from its frequentist properties: any reparametrization leads to the same type of estimator. For example, the ML estimator of the variance is the square of the ML estimator of the standard deviation, and in general a function of a ML estimator is also a ML estimator.

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

[11] Speaking strictly, these quantities are *densities* of probability. As we will see in 3.3.1, probabilities are areas defined by f(y), like $f(y)\Delta y$.

likelihood function is sharp, its maximum gives a more *likely* value of the parameter than other values near it. Conversely, if the likelihood function is rather flat, other values of the parameter will be almost as *likely* as the one that gives the maximum to the function. The frequentist school also discovered that the likelihood was useful for construction of hypothesis tests, since the likelihood ratio between the null and the alternative hypothesis has good asymptotical frequentist properties, and it is currently used for testing hypotheses. We will come back to this in chapter 10.

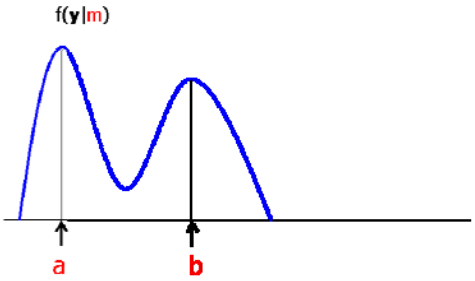

Figure 11. Likelihood curve. Here m can take "likely" the values 'a' or 'b', however the frequentist school will only take the maximum at 'a'

1.6.3. *Common misinterpretations*

The method of maximum likelihood finds the estimate that makes the sample most probable: This is strictly nonsense, since each sample has its probability depending on the true value of the distribution from which it comes. For example, if the true value of the population is the case **c** in figure 9 ($m_{TRUE} = m_3 = 1.8$), our sample $y_0 = 1.6$ is rather improbable, but its probability is not modified just because we use a maximum likelihood method to estimate the true value of m. Our maximum likelihood estimate will be $\hat{m} = m_2 = 1.6$, but the true probability of our sample still will be very low because it really comes from population **c** of figure 9. Therefore, the method of ML is *not* the one that makes the sample most probable. This method provides a value of the parameter that *if this were the true value* the sample would be most probable ([12]). As Fisher says, for the case of estimating the true coefficient of correlation ρ from the value r obtained in a sample:

[12] Some authors say that likelihood maximizes the probability of the sample *before performing the experiment*. They mean that f(y|m) can be considered a function of both, y and m, before taking

> "We define likelihood as a quantity proportional to the probability that, *from a population having that particular value of ρ*, a sample having the observed value *r*, should be obtained".
>
> Fisher, 1921

A likelihood four times bigger than other likelihood gives four times more evidence in favour of the first estimate: This is not true. Unfortunately, likelihoods are not quantities that can be treated as probabilities because each value of the likelihood comes from a different probability distribution. Then they do not follow the laws of the probability (e.g., they do not sum up to one, the likelihood of excluding events is not the sum of their likelihoods, etc.). Therefore a likelihood four times higher than other one does not lead to a "degree of rational belief" four times higher, as we will see clearly in chapter 7. There is an obvious risk of confusing likelihood and probability, as people working in QTL should know ([13]).

Appendix 1.1. Definition of relevant difference

In both classical and Bayesian statistics it is important to know which difference between treatments should be considered "relevant". It is usually obtained under economical considerations; for example, which difference between treatments justifies to do an investment or to prefer one treatment. However there are traits like the results of a sensory panel test or the enzymatic activities for which it is difficult to determine what a relevant difference between treatments is. To find significant differences is not a solution to this problem because we know that if the sample is big

---

samples and it can be found an m that maximizes f(y|m) for each given y. Again, it maximizes the density of probability only if m is the true value, and in my opinion the sentence *before performing the experiment* is a clear abuse of the common language.

[13] Figures of marginal likelihoods, common in papers searching for QTL, are often interpreted as probabilities. Incidentally, one of the founders of the frequentist theory, Von Mises (1957, pp. 157-158), accuses Fisher of exposing with great care the differences between likelihood and probability, just to forget it later and use the word 'likelihood' as we use 'probability' in common language.

enough, we will always find significant differences. I propose considering that a relevant difference depends on the variability of the trait. To have one finger more in a hand is relevant because the variability of this trait is very small, but to have one hair more in the head is not so relevant (although for some of us it is becoming relevant with the age). Take an example of rabbits: carcass yield has a very small variability; usually the 95% of rabbits have a carcass yield (Spanish carcass) between 55% ± 2%, thus a difference between treatments of a 2.75%, which is a 5% of the mean, is a great difference between treatments. Conversely, 95% of commercial rabbits have litter size between 10 ± 6 rabbits, thus a 5% of the mean as before, 0.5 rabbits, is irrelevant. If we take a list of the important traits in animal production, we will see that for most of them the economical relevance appears at a quantity placed between ½ or ⅓ of the standard deviation of the trait. Therefore, I propose to consider that a relevant difference between treatments is, for all traits in which it is not possible to argue economical or biological reasons, a quantity placed between ½ or ⅓ of the standard deviation of the trait. This sounds arbitrary, but it is even more arbitrary to compare treatments without any indication of the importance of the differences found in the samples.

Another solution that we will see in chapter 2 would be to compare ratios of treatments instead of differences between treatments. It can be said that a treatment has an effect a 10% bigger than the other, or its effect is a 92% of the other one. This can be complex in classical statistical, mainly because the s.e. of a ratio is not the ratio of the s.e., and it should be calculated making approximations that do not always work well ([14]), but is trivial for Bayesian statistics when combined with MCMC.

Appendix 1.2

If u, û are normally distributed, the relationship between u and û is linear

[14] For example, the delta method, commonly used in quantitative genetics to estimate s.e. of ratios, does not work well for correlations (Visscher, 1998).

u = b P + e = û +e

cov(u,û) = cov(u, û+e) = cov(u,û) + cov(û,e) = var(û) + 0 = var(û)

var(e) = var(u–û) = var(u)+var(û)–2cov(u,û) = var(u) + var(û) – 2 var(û) =
= var(u) – var(û)

if the relationship is not linear, then cov(û,e) may be different from zero and then this new term should be considered.

Appendix 1.3

Let us estimate the genetic value 'u' of a bull using the average of their daughters '$\bar{y}$' by regression.

$$u = b \cdot \bar{y} + e = \hat{u} + e$$

$$\text{var}(\hat{u}) = b^2\,\text{var}(\bar{y}) = \frac{\text{cov}^2(u,\bar{y})}{\left[\text{var}(\bar{y})\right]^2} \cdot \text{var}(\bar{y}) = \frac{\left(\sigma_u^2\right)^2}{\text{var}(\bar{y})} = \frac{\left(\sigma_u^2\right)^2}{\frac{1}{n^2}\left(n\sigma_y^2 + n(n-1)\sigma_u^2\right)} = \frac{n \cdot \left(\sigma_u^2\right)^2}{\sigma_y^2 + (n-1)\cdot\sigma_u^2}$$

When n increases, var(û) increases.

When $n \to \infty$, $\text{var}(\hat{u}) \to \sigma_u^2$ $\text{var}(e) = \text{var}(u - \hat{u}) \to 0$

CHAPTER 2

THE BAYESIAN CHOICE

> "Si un événement peut être produit par un nombre n de causes différentes, les probabilités de l'existence de ces causes prises de l'événement sont entre elles comme les probabilités de l'événement prises de ces causes, et la probabilité de l'existence de chacune d'elles est égale à la probabilité de l'événement prise de cette cause, divisé par la somme de toutes les probabilités de l'événement prises de chacune de ces causes".
>
> **Pierre Simon, Marquis de Laplace**, 1774.

2.1. Bayesian inference
- 2.1.1. Bases of Bayesian inference
- 2.1.2. Bayes theorem
- 2.1.3. Prior information

2.2. Features of Bayesian inference
- 2.2.1. Point estimates: Mean, median, mode
- 2.2.2. Credibility intervals
- 2.2.3. Marginalisation

2.3. Test of hypotheses
- 2.3.1. Model choice
- 2.3.2. Bayes factors
- 2.3.3. Model averaging

2.4. Common misinterpretations

2.5. Bayesian Inference in practice

2.6. Advantages of Bayesian inference

Appendix 2.1

Appendix 2.2

2.1. Bayesian inference

2.1.1. *Bases of Bayesian inference*

Bayesian inference is based on the use of probability for expressing uncertainty. It looks more natural to express uncertainty saying that the probability of two treatments being different is 98%, than saying that our behaviour should be to admit they are not equal 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 to find which value of this parameter, if it would be the true value, would produce our data with a highest probability. To examine the distribution of the data or the distribution of an estimator based on a combination of the data is less attractive than to examine the distribution of the probability of the parameter we want to estimate. This was recognised by all founders of what we call now “classical statistics”, as the citations heading this lecture notes and chapter 7 show. All of them also preferred probability statements to express uncertainty, but they thought it was not possible to construct these statements. The reason is that in order to make probability statements based in our data we need some prior information and it is not clear how to introduce this prior information in our analysis or how to express lack of information using probability statements. However, Bayesian statisticians say they have found solutions for this problem 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 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 chapter 7 the discussion about the Bayesian solutions for 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.

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 $(\bar{x}_S - \bar{x}_C)$. 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

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

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?

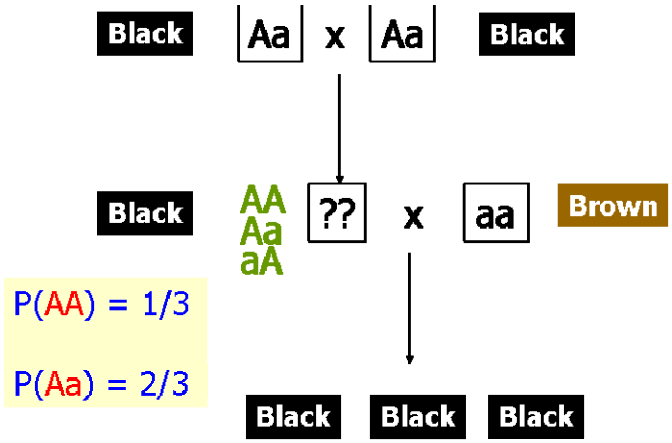

Figure 2.1. Two heterozygous mice have an offspring that may be homozygous or heterozygous. To test this it is crossed to a brown mouse and the offspring is examined. Before performing the experiment we have some prior information due to our knowledge of Mendel's law.

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). We will do this in chapter 7.

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. This was not correctly perceived along the 19th century, and it was always assumed that we cannot integrate prior information properly. The first solution to this problem was provided independently by the philosopher of the mathematics Ramsay (1926) and the Italian actuary Bruno de Finetti (1934) ([15]). 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. They 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 a 6 when throwing a dice. If we have data

[15] The famous economist J.M. Keynes proposed probability as a state of belief before (Keynes, 1921), but his probabilities could not be used for computations.

enough, our data will overcome the prior opinion we have and our experiment will 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 in previous data, not in unfounded guessing or in arbitrary statements.

When this solution was proposed, some statisticians were scandalised by thinking that science was becoming something subjective. Kempthorne express 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.

However, in fields in which the expert opinion is used to take decisions (like in economy) this did not represent any problem. In biological sciences it is preferred that the data dominates and the 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 rank of variation of this parameter, being the average about 0.4. However, in a former experiment performed 11 years ago with the same population, the heritability was 0.11. Then, three vague states of beliefs were tested (figure 2.2)

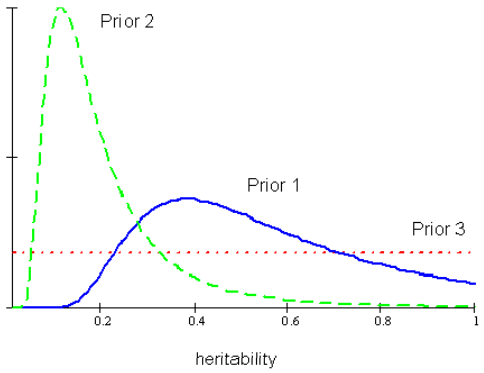

Figure 2.2. Three different priors showing three different states of belief about the heritability of ovulation rate in French Landrace pigs (from Blasco et al., 1998).

the first considered that it was more probable to find heritability around 0.4 and the second that it was more probable to find it around 0.11. Both curves are asymmetric because heritability is rarely high in most traits. A third state of opinion was considered trying to express indifference about the value of heritability: 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 final conclusions and decisions to be taken.

It can be argued that when the prior belief is very sharp, it will dominate and the results will reflect our prior belief instead of what our experiment brings. But, why should we perform an experiment if we have sharp prior beliefs? If we are pretty 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 around 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 if we analyze the data in a Bayesian way will be similar to our prior opinion independently of the result of our experiment. But if we analyze the experiment using classical statistics and we find a heritability of 0.9, 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. The problem arises in the multivariate case, in which we cannot state a prior opinion because human minds cannot think in many dimensions. We will deal with this problem in chapter 7. ([16])

[16] There is a Bayesian statistician that quotes Kant (Robert, 1992, pp. 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 one of the first and most important questions to know whether there is some knowledge independent of experience. Call this knowledge ‘a priori’ and make distinction from the empiric knowledge in that the sources of this last one are ‘a posteriori’, based in the experience” (Kant, 1781, pp. 98). Of course, no Bayesian scientist would use priori knowledge as something not based in previous experiences. Therefore, I do not find any relationship between Kant’s philosophy and the Bayesian paradigm.

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 the results we will obtain ignoring it and basing our inferences only in our data. Unfortunately it does not seem easy to describe ignorance using probability statements. Along the 19th century and the first three decades of the 20th 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 aspect; prior 3 of figure 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'); for example, it is not the same to say that we do not know which is the true value of the heritability in the example before than to say that all possible values have the same probability. Moreover, this principle leads to paradoxes, as we will see in chapter 7. Other alternatives have been proposed: Jeffreys (1961) proposes priors that are invariant to transformations and Bernardo (1979) proposes priors that have minimum information. All these priors are called "objective" or "non-informative" by Bayesian statisticians, however it should be noted that *all of them are informative* (although usually the information they provide is rather vague) and *the information they provide is not objective*, at least using the commons meaning of this word ([17]).

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*

[17] 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.

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).

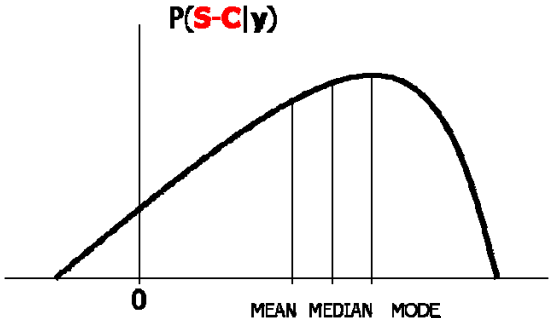

Figure 2.3. Mean, median and mode of the probability distribution of the difference between selected and control groups, given the information provided by our data.

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 $\hat{u}=u$, *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.

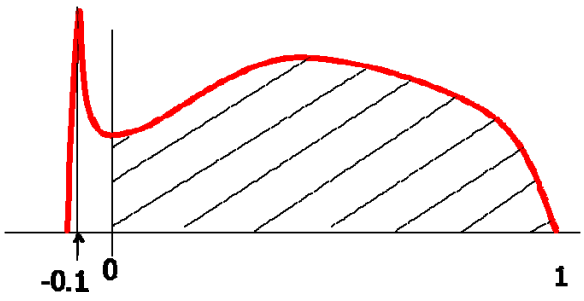

Figure 2.4. Probability distribution of a correlation coefficient given the data. The mode is negative, but the coefficient of correlation is probably positive.

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 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 ([18]).

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

[18] Please do not confuse the median of the distribution with the median of a sample when we want to estimate the population mean. In this last case the median has less information than the arithmetic mean, but we are talking now about probability distributions, in the continuous case with an infinite number of points, we are not using a sampling estimator.

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.

Notice that although *Risk* has the same definition as in the frequentist case, i.e. the mean of the loss function, here the variable is not û, which is a combination of the data, and in a Bayesian context the data are fixed, we do not repeat the experiment conceptually 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*

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).

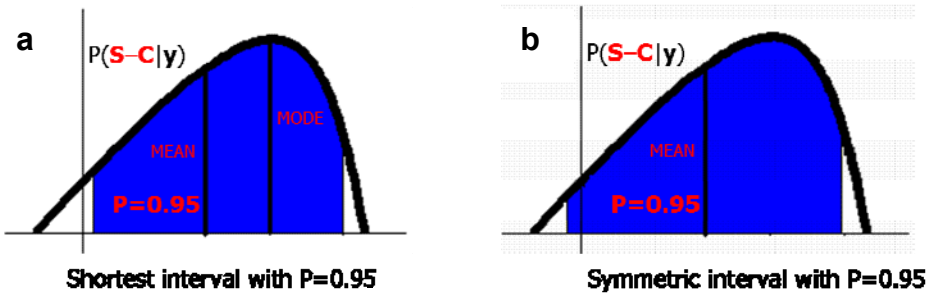

Figure 2.5. Credibility intervals containing the true value with a probability of 95%. a. Shortest interval (not symmetric around the mean or the mode). b. Symmetric interval around the mean.

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.

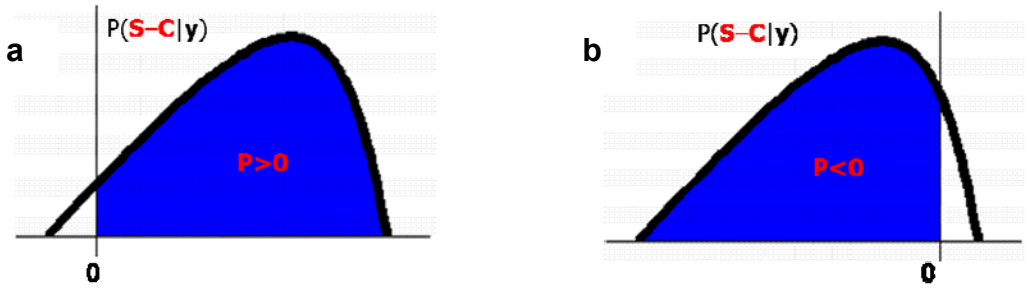

Figure 2.6. Credibility intervals. a. Interval [0, +∞) showing the probability of S-C of being equal or higher than zero. b. Interval (-∞, 0]) showing the probability of S-C of being equal or lower than zero

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) ([19]).

[19] These intervals have the advantage of being invariant to transformations (see Appendix 2.1). 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.

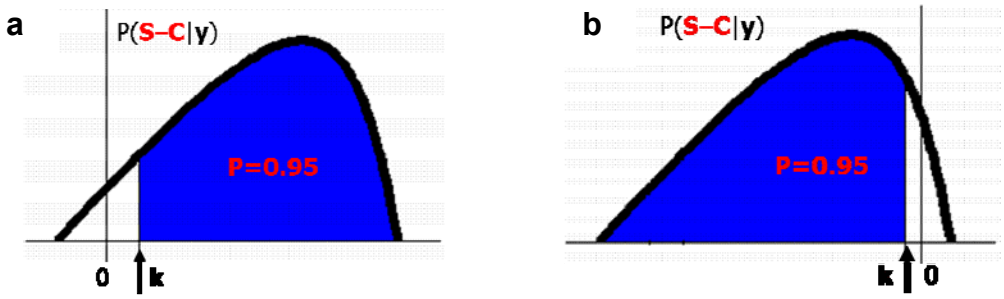

Figure 2.7. Credibility intervals. a. Interval [k, +∞) showing the lowest value of an interval containing the true value with a probability of 95%. b. Interval (-∞, k] showing the highest value of an interval containing the true value with a probability of 95%.

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).

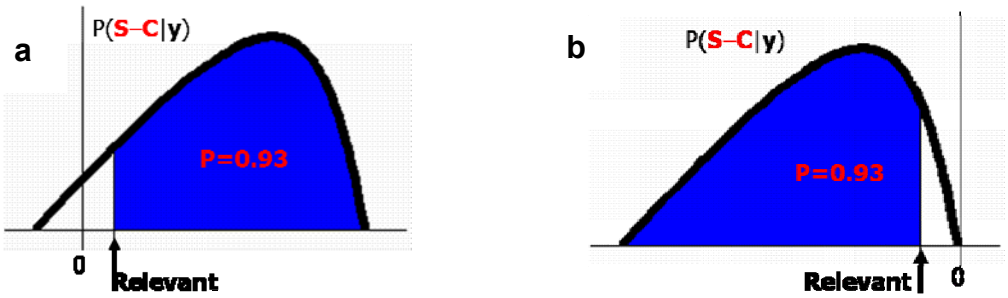

Figure 2.8. Credibility intervals. a. Interval from a relevant quantity to +∞, showing the probability of the difference S-C of being relevant. b. Intervals from –∞ to a relevant quantity, showing the probability of the difference S-C of being relevant

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".

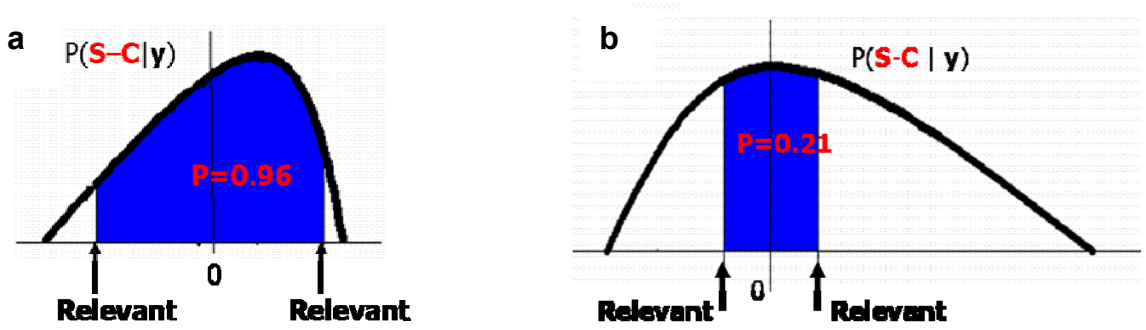

Figure 2.9. Probability of similarity between S and C. a.S and C are similar. b. We do not have data enough to determine whether S is higher, lower or similar to C.

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.

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).

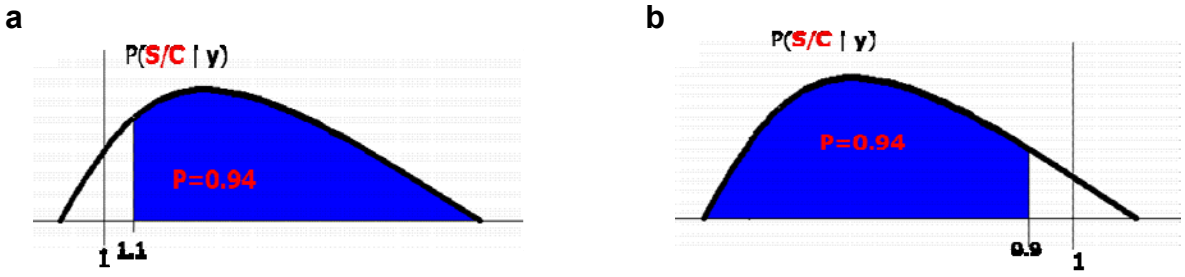

Figure 2.10. Credibility interval for the ratio of levels of a treatment. a. The probability of S being a 10% higher than C is a 94%. b. The probability of S being a 90% of C is a 94%

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, $\sigma^2 = 0.5$ and $\sigma^2 = 1$. The marginal posterior distribution of the difference between treatments will be the sum of P(S-C given the data and given that $\sigma^2 = 0.5$) and P(S-C given the data and given that $\sigma^2 = 1$) 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 $P(S\text{-}C \mid y, \sigma^2)$ for all values of σ^2 from 0 to ∞.

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

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*

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 in that there is no need of hypothesis tests for most biological experiments. This has been emphasised by Gelman et al. (2003), Howson and Urbach (1996) and Robert (1992) for both, frequentist and Bayesian statistics. Nevertheless there are circumstances in we may need to select one among several statistical models, for example when we are trying to be read off some noise effects of the model that have too many levels to test them two by two (e.g. herd-year-season effects with many levels), or when we have many noise effects and we would like to know whether we can be read off all of them. 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 from others. Test of hypothesis is only one of the possible ways of model choice. We can use criteria based in concepts like amount of information, criteria based on probability or heuristic criteria that we have seen by simulation or by our experience that work well in determined circumstances. We can also choice a model based in its predictive properties by using cross-validation techniques (frequentist or Bayesian). Bayesian tests are based in comparing the probabilities of several hypothesis given our data, but there are heuristic criteria that, without being properly Bayesian, use some Bayesian elements like the mean of a posterior distribution (for example, the now popular Deviance Information Criterion, DIC). Cross validation can be also performed using Bayesian criteria like the probability of predicting part of the sample given other elements of the sample. We will expose here the way in which Bayesian model choice is performed. In last chapter we will discuss the difficulties of the Bayesian approach, and other alternative approaches.

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. For example, there are two different models that are used in the study of allometric growth; one of them (we call it H_1) says that the relative growth of two tissues ‘x’ and ‘y’ of an organism is represented by what we nowadays know as “Huxley equation”

$$y = b \cdot x^k$$

where ‘b’ and ‘k’ are parameters to be estimated. The other model (H_2) is the Bertalanfly (1983) equation

$$\frac{y}{A_y} = q\frac{x}{A_x} + (1-q)\left(\frac{x}{A_x}\right)^2$$

where A_x and A_y are the adult weight of these tissues and q the only parameter to be estimated.

We can calculate the probability of each hypothesis $P(H_1|\mathbf{y})$, $P(H_2|\mathbf{y})$ 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 $P(\mathbf{y})$ is the sum of the probabilities of two excluding events: H_1 is the true hypothesis or it is H_2.

We calculate $P(H_2|\mathbf{y})$ in the same way. After calculating $P(H_1|\mathbf{y})$, $P(H_2|\mathbf{y})$ we can choose the most probable hypothesis. This can be extended to several hypotheses, we can calculate $P(H_1|\mathbf{y})$, $P(H_2|\mathbf{y})$, $P(H_3|\mathbf{y})$, … 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

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

where

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

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

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

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 $f(\theta)$. 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 ([20]).

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

[20] 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.

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 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.

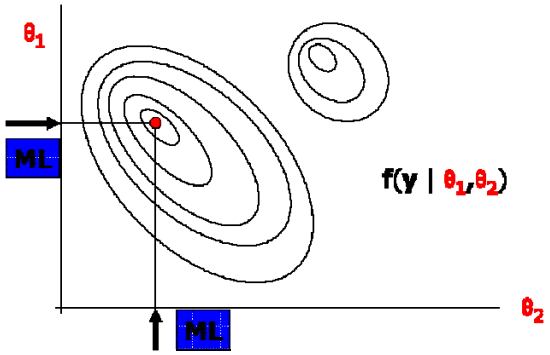

Figure 2.10. Probability function of the data for two parameters. Lines should be interpreted as level curves of a map.

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 the actual probability, and it is a matter of opinion whether this probability is high or low.

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, unpublished results) analyze a possible association of a *PGR* gene polymorphism (GG, GA, AA) with litter size in a rabbit population. They consider that 0.5 rabbits was a relevant quantity. The GG genotype had higher litter size than GA genotype, the difference between genotypes D was relevant and P(D>0)=99%. The GA genotype had similar litter size than the AA genotype (Probability of similarity = 96%), which indicates that the genetic determination of this trait is dominant. Here the

probability of similarity (figure 2.9.a) means that the area of the posterior distribution of P(GA-AA | **y**) included between -0,5 and +0.5 kits was 96%.

Example 4. Hernández 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). A Bayesian analysis obtained the marginal posterior distribution shown in figure 2.11. Here the distribution is asymmetrical, thus mode, mean and median are not coincident. A usual choice is to take the mean (-0.93) because it minimizes the quadratic risk, which is conventional, 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 one, as expected.

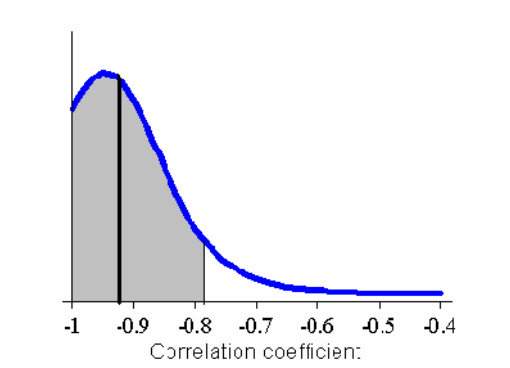

Figure 2.11. Probability distribution of a correlation coefficient. Notice that it is highly asymmetric.

2.6. Advantages of Bayesian inference

We will resume here the advantages of Bayesian inference:

- We are not worried about bias (there is nothing like bias in a Bayesian context).
- We should not decide whether an effect is fixed or random (all of them are random).

- We often do not need often Hypothesis tests because posterior distribution of differences between effects substitutes them.
- We have a *measure of uncertainty* for both hypothesis tests and credibility intervals, we work with Probabilities. We do not have “prior” risks.
- We work with marginal probabilities: i.e., all multivariate problems are converted in univariate and we take into account errors of estimating other parameters.
- We have a method for inferences, a path to be followed.

Appendix 2.1

We know (see chapter 3) 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 2.2

If we take a flat prior f(θ) = constant and we realize that x= exp (logx),

f(θ|y) $\propto$ f(y|θ) f(θ) $\propto$ f(y|θ) = exp [logf(y|θ)]

We can develop a Taylor series up till the second term around the mode M of logf(θ|y).

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

but the first derivative around the mode is null because it is a maximum, thus

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

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. As the normal distribution is symmetric, mode mean and median is the same.

CHAPTER 3

POSTERIOR DISTRIBUTIONS

"If science cannot measure the degree of probability involved, so much the worse for science. The practical man will stick to his appreciative methods until it does, or will accept the results of inverse probability of the Bayes/Laplace brand till better are forthcoming".

Karl Pearson,1920.

3.1. Notation
3.2. Cumulative distribution
3.3. Density distribution
    3.3.1. Definition
    3.3.2. Transformed densities
3.4. Features of a density distribution
    3.4.1. Mean
    3.4.2. Median
    3.4.3. Mode
    3.4.4. Credibility intervals
3.5. Conditional distribution
    3.5.1. Definition
    3.5.2. Bayes Theorem
    3.5.3. Conditional distribution of the sample of a Normal distribution
    3.5.4. Conditional distribution of the variance of a Normal distribution
    3.5.5. Conditional distribution of the mean of a Normal distribution
3.6. Marginal distribution
    3.6.1. Definition

3.6.2. Marginal distribution of the variance of a normal distribution
3.6.3. Marginal distribution of the mean of a normal distribution
Appendix 3.1
Appendix 3.2
Appendix 3.3
Appendix 3.4

3.1. Notation

We have used an informal notation in chapter 2 because it was convenient for a conceptual introduction to the Bayesian choice. From now, we will use a more formal notation. Bold type will be used for matrixes and vectors, and capital letters for matrixes. Thus '**y'** represents a vector, '**A**' is a matrix and 'y' is a scalar. Unknown parameters will be represented by Greek letters, like μ and σ^2.

The letter 'f' will always be used for probability density functions. Thus f(x) and f(y) are different functions (for example, a normal function and a gamma function) although both are probability density functions.

We use sometimes "density" or "distribution" as short names of "probability density distributions".

The letter 'P' will be reserved for probability. For example, P(**y**) is a number between 0 and 1 representing the probability of the sample **y**.

The variables will be in red colour, independently on whether they are after a statement of '|' (given) or not. For example,

$$f(y \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} \right]$$

is a function of σ^2, but it is not a function of y or a function of μ, that are considered as constants. Thus, in this example we represent a family of density functions for different values of σ^2. An example of this type of function is the likelihood (see chapter 1).

The sign $\propto$ means "proportional to". We will often work with proportional functions, since it is easier and as we will see later, the results from proportional functions can be reduced to (almost) exact results easily with MCMC. For example, if c and k are constants, the distribution N(0,1) is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \propto \exp\left(x^2\right) \propto k \cdot \exp(c) \cdot \exp\left(x^2\right) = k \cdot \exp\left(c + x^2\right)$$

Thus we can add or subtract constants from an exponent, or multiply or divide if the constants are not in the exponent. What we cannot do is

$$f(x) \propto \exp\left(x^2\right) \neq k + \exp\left(x^2\right)$$
$$f(x) \propto \exp\left(x^2\right) \neq \exp\left(c \cdot x^2\right) = \left[\exp\left(x^2\right)\right]^c$$

3.2. Cumulative distribution

Take the example of a single gene with two alleles (A,a) producing three genotypes aa, Aa, AA. We define a variable x with three values 0,1,2 corresponding to the three genotypes. If we cross two individuals Aa, according to Mendel's law, the probability of each offspring genotype is

	aa	Aa	AA
x_0 =	0	1	2
$P(x = x_0)$ =	¼	½	¼

We define the cumulative distribution function as

$F(x_0) = P(x \leq x_0)$

in this example, this cumulative distribution function is (figure 3.1)

$F(0) = P(x \leq 0) = ¼$
$F(1) = P(x \leq 1) = ¼ + ½ = ¾$
$F(2) = P(x \leq 2) = ¼ + ½ + ¼ = 1$
For all values $x<0$ we have $F(x) = 0$
For all values $0 \leq x < 1$ we have $F(x) = ¼$
For all values $1 \leq x < 2$ we have $F(x) = ¾$
For all values $x \geq 2$ we have $F(x) = 1$

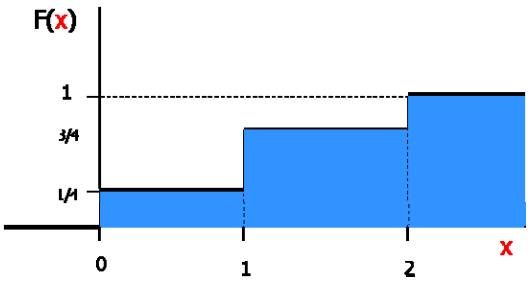

Figure 3.1. Cumulative distribution function

The probability of x to be between 1 and 2 (equal or lower than 2 but higher than 1)

$P(1 < x \leq 2) = P(x \leq 2) – P(x \leq 1) = F(2) – F(1) = 1 – ¾ = ¼$

Usually the cumulative distribution functions are continuous, for example

$$F(x) = x = P(x_0 \leq x)$$

is a straight line in which probability augments with x at a constant rate (figure 3.2)

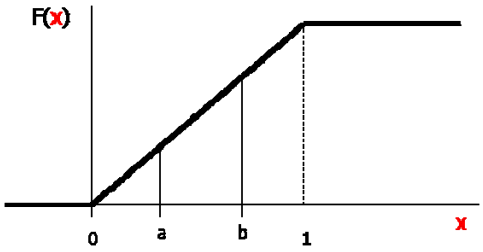

Figure 3.2. Cumulative distribution function

The probability of x to be between a and b is

$P(a < x \leq b) = P(x \leq b) – P(x \leq a) = F(a) – F(b)$

3.3. Density distribution

3.3.1. *Definition*

We define the probability density function (figure 3.3.) as

$$f(x) = \frac{\Delta F(x)}{\Delta x}$$

this function is always positive because $\Delta F(x)$ and Δx are both positive. In the example of figure 3.1., the values of $\Delta F(x)$ are ¼ until 0, ½ from 0 to 1 and ½ from 1 upwards. Then,

$\sum f(x) \cdot \Delta x = \sum \Delta F(x) = ¼ + ½ + ¼ = 1$

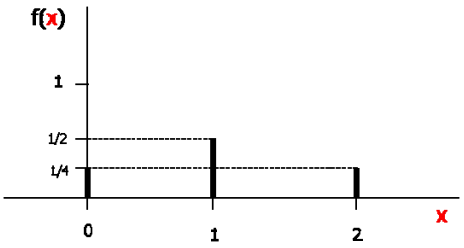

Figure 3.2. Probability density function

in the continuous case

$$f(x) = \frac{dF(x)}{dx}$$

this means (figure 3.3a) that

$$F(x_0) = \int_{-\infty}^{x_0} f(x)\,dx = P(x \le x_0)$$

and, analogously as in the discrete case

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

The probability of x to be between two values ‘a’ and ‘b’ is

$$P(x \le b) - P(x \le a) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx$$

which is the area defined by f(x) between ‘a’ and ‘b’ (figure 2.3**b**)

Figure 3.3. Probability density functions. a. In blue, $P(x \le x_0)$, b. In blue, $P(a \le x \le b)$

Notice that $f(x_0)$ is not a probability. $F(x_0)$ is a probability, by definition. Areas defined by f(x) are probabilities, and the area $f(x_0)\cdot\Delta x$ is approximately a probability when Δx is small (figure 3.4). These small probabilities are usually expressed as f(x)dx.

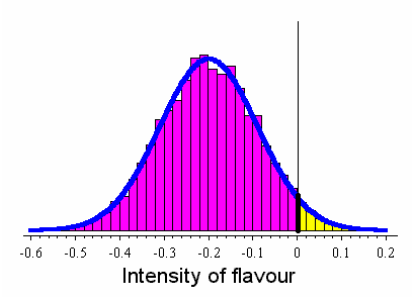

Figure 3.4. Probabilities are areas of f(x), for example the shadowed area of f(x) for all x>0. The small rectangles f(x)·Δx are approximate probabilities.

3.3.2. *Transformed densities*

We know a density f(x) and we want to find the density f(y) of a function y = g(x). In Appendix 3.1 we show that

$$f(y) = f(x)\left|\frac{dx}{dy}\right| = f(x)\left|\frac{dy}{dx}\right|^{-1}$$

For example, we have a Normal distribution f(x) and we want to know the distribution of y = exp(x). We have

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$f(y) = f(x)\cdot\left|\frac{dy}{dx}\right|^{-1} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\cdot\frac{1}{\exp(x)}$$

As x = ln(y), we finally have

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2} \right] \cdot \frac{1}{\exp(\ln y)} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2} \right] \cdot \frac{1}{y}$$

In the multivariate case, we know f(x, y) and we want the density f(u, w), where u and w are functions of (x,y)

u = g(x,y) ; w = h(x,y)

The corresponding formula to the univariate case is

$$f(u,w) = f(x,y) \cdot \begin{vmatrix} \frac{\partial^2 f(x,y)}{\partial u^2} & \frac{\partial^2 f(x,y)}{\partial u \partial w w} \\ \frac{\partial^2 f(x,y)}{\partial u \partial w} & \frac{\partial^2 f(x,y)}{\partial w^2} \end{vmatrix}$$

For example, we know the densities of x and y and we would like to know the density of $\frac{x}{x+y}$. This happens, for example, when in genetics we have the densities of the additive and environmental variance of a trait and we want to have the density of the heritability, which is the ratio of the additive and the sum of both variance components. In this case f(x, y) is a bivariate Normal distribution and

u = x

w = $\frac{x}{x+y}$

$$f(u,w) = f(x,y) \cdot \begin{vmatrix} \frac{\partial^2 f(x,y)}{\partial u^2} & \frac{\partial^2 f(x,y)}{\partial u \partial w w} \\ \frac{\partial^2 f(x,y)}{\partial u \partial w} & \frac{\partial^2 f(x,y)}{\partial w^2} \end{vmatrix}$$

3.4. Features of a density distribution

3.4.1. *Mean*

The *expectation* or *mean* of a density function is

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

If we have a function of the random variable

$y=g(x)$

we know (Appendix 3.1) that

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

thus, its expectation is

$$E(y) = \int_{-\infty}^{\infty} y \cdot f(y)\,dy = \int_{-\infty}^{\infty} g(x) \cdot f(x)\left|\frac{dx}{dy}\right| dy = \int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx$$

For example, if

$y = x^2$

$$E(y) = \int_{-\infty}^{\infty} y \cdot f(y)\,dy = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx$$

3.4.2. *Median*

The median is the value dividing the density distribution in two parts ach one with a 50% of probability, i.e.: the median is the value m_x that

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

3.4.3. *Mode*

The mode is the maximum of the density function

Mode = arg max f(x)

The mode it is the most probable value in the discrete case, and in the continuous case is the value around which the probability is maximum (i.e.: the value of x for which $f(x) \cdot \Delta x$ is maximum).

3.4.4. *Credibility intervals*

A credibility interval of a given probability, for example a 90%, is the interval [a, b] containing a probability of 90%. Any values 'a' and 'b' for which

$$\int_{a}^{b} f(x)dx = 0.90$$

constitute a credibility interval [a, b] at 90%. Observe that there are infinite credibility intervals at 90% of probability, but they have different length (see figure 2.5). One of them is the shortest one, and in Bayesian inference, when the density function used is the posterior density, it is called the Highest Posterior Density interval at 90% ($HPD_{90\%}$).

3.5. Conditional distribution

3.5.1. *Definition*

We say that the conditional distribution of x given $y=y_0$ is

$$f(x|\,y=y_0)=\frac{f(x,y=y_0)}{f(y=y_0)}$$

Following our notation, in which the variables are in red and the constants are in black, we can express the same formula as

$$f(x|\,y)=\frac{f(x,y)}{f(y)}$$

If we consider that y can take several values, the formula can be expressed as

$$f(x|\,y)=\frac{f(x,y)}{f(y)}$$

and in this case it represents a family of density functions, with as different density function for each value of y = y_0, y_1, y_2, …

For two given values ‘x’ and ‘y’, the formula is

$$f(x\,|\,y)=\frac{f(x,y)}{f(y)}$$

3.5.2. *Bayes theorem*

Although f(x) is not a probability, we found that $f(x)\cdot\Delta x$ is indeed a probability (see fig. 3.4), thus applying Bayes theorem, we have

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

$$f(x\,|\,y)\Delta x=\frac{f(y\,|\,x)\Delta y\cdot f(x)\Delta x}{f(y)\Delta y}\qquad\longrightarrow\qquad f(x\,|\,y)=\frac{f(y\,|\,x)\cdot f(x)}{f(y)}$$

thus we have now a version of Bayes theorem for density functions. Considering x as the variable and 'y' as a given (constant) value,

$$f(x\,|\,y)=\frac{f(y\,|\,x)\cdot f(x)}{f(y)}$$

which can be expressed proportionally, since f(y) is a constant,

$$f(x\,|\,y)\propto f(y\,|\,x)\cdot f(x)$$

For example, if we have a normal distribution $y \sim N(\mu,\sigma^2)$ in which we do not know the parameters μ, σ^2, the uncertainty about both parameters can be expressed as

$$f(\mu\,|\,\sigma^2,y)\propto f(y\,|\,\mu,\sigma^2)f(\mu)$$

$$f(\sigma^2\,|\,\mu,y)\propto f(y\,|\,\sigma^2,\mu,)f(\sigma^2)$$

3.5.3. *Conditional distribution of the sample of a Normal distribution*

Let us consider now a random sample **y** from a normal distribution

$$f(\mathbf{y}\,|\,\mu,\sigma^2)=f(y_1,y_2\mathrm{K}\;y_n\,|\,\mu,\sigma^2)=f(y_1\,|\,\mu,\sigma^2)\cdot f(y_2\,|\,\mu,\sigma^2)\cdot\mathrm{K}\cdot f(y_n\,|\,\mu,\sigma^2)=$$

$$= \prod_1^n f\left(y_i \mid \mu, \sigma^2\right) = \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right] = \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2}{2\sigma^2}\right]$$

this is the conditional distribution of the data because it is conditioned to the given values of 'μ' and 'σ^2'; i.e., for each given value of the mean and the variance we have a different distribution. For example; for a given mean μ=5 and variance σ^2=2, and for a sample of three elements **y'** = [y_1, y_2, y_3] we have

$$f\left(\mathbf{y} \mid \mu, \sigma^2\right) = \frac{1}{\left(\sqrt{2\pi}\right)^3 (2)^{\frac{3}{2}}} \exp\left[-\frac{\left(y_1-5\right)^2 + \left(y_2-5\right)^2 + \left(y_3-5\right)^2}{2 \cdot 2}\right]$$

This is a trivariate multinormal distribution, with three variables y_1, y_2 and y_3.

3.5.4. *Conditional distribution of the variance of a Normal distribution*

We can also write the conditional distribution of the variance for a given sample '**y**' and a given mean 'μ'. We do not know f(σ^2 | μ, **y**), but we can apply Bayes theorem and we obtain

$$f\left(\sigma^2 \mid \mu, \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \sigma^2, \mu\right) f\left(\sigma^2\right)$$

applying the principle of indifference (see 2.1.3), we will consider in this example that *a priori* all values of the variance have the same probability density; i.e., f(σ^2)=constant. This leads to

$$f\left(\sigma^2 \mid \mu, \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \sigma^2, \mu\right)$$

but we know the distribution of the data, that we have assumed to be Normal, thus we can write the conditional distribution of the variance,

$$f\left(\sigma^2 \mid \mu, \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \sigma^2, \mu\right) = \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right]$$

Notice that the variable is coloured in red, thus here the variance is the variable and the sample and the mean are given constants. For example, if the mean and the sample are

$\mu = 1$

$\mathbf{y}' = [2, 3, 4]$

for this given mean and this given sample, the conditional distribution of the variance is

$$f\left(\sigma^2 \mid \mu, \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{3}{2}}} \exp\left[-\frac{(2-1)^2 + (3-1)^2 + (4-1)^2}{2\sigma^2}\right] = \frac{1}{\left(\sigma^2\right)^{\frac{3}{2}}} \exp\left[-\frac{7}{\sigma^2}\right]$$

Notice that this is not a Normal distribution. A Normal distribution does not have the variable (in red) there; a Normal distribution looks like

$$f\left(x\right) = \frac{1}{\left(\sqrt{2\pi}\right)(10)^{\frac{1}{2}}} \exp\left[-\frac{(x-5)^2}{2 \cdot 10}\right]$$

where 5 and 10 are the mean and the variance of the variable x.

Which is, then, the conditional distribution of the variance? There is a type of distribution called “inverted gamma” that looks like

$$f\left(x \mid \alpha, \beta\right) \propto \frac{1}{x^{\alpha+1}} \exp\left[-\frac{\beta}{x}\right]$$

where ‘α’ and ‘β’ are parameters that determine the shape of the function. In figure 2.2 we find three different shapes we can obtain by given different values to ‘α’ and ‘β’, a flat line and two curves, one of them sharper than the other one. We can see that the conditional distribution of the variance is of the type “inverted gamma”. If we take

$$\beta = 7$$
$$\alpha + 1 = \frac{3}{2}$$

we obtain an “inverted “gamma”, and in general the variance of a Normal distribution is an inverted gamma ([21]) with parameters

$$\beta = \frac{1}{2}\sum_{1}^{n}(y_i - \mu)^2$$

$$\alpha = \frac{n}{2} - 1$$

3.5.5. *Conditional distribution of the mean of a Normal distribution*

We will write now the conditional distribution of the mean for a given sample ‘**y**’ and a given variance ‘σ^2’. We do not know $f(\mu \mid \sigma^2, \mathbf{y})$, but we can apply Bayes theorem and we obtain

$$f(\mu \mid \sigma^2, \mathbf{y}) \propto f(\mathbf{y} \mid \mu, \sigma^2) f(\mu)$$

Applying the principle of indifference (see 2.1.3), we will consider in this example that *a priori* all values of the variance have the same probability density; i.e., $f(\mu)$=constant. This leads to

$$f(\mu \mid \sigma^2, \mathbf{y}) \propto f(\mathbf{y} \mid \mu, \sigma^2)$$

[21] This type of inverted gamma is usually named “inverted chi-square”

but we know the distribution of the data, that we have assumed to be Normal, thus we can write the conditional distribution of the mean,

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right]$$

Notice that the variable is coloured in red, thus here the mean is the variable and the sample and the variance are given constants. For example, if the variance and the sample are

$\sigma^2 = 9$

$\mathbf{y}' = [2, 3, 4]$

for this given variance and this given sample, the conditional distribution of the mean is

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto \frac{1}{(9)^{\frac{3}{2}}} \exp\left[-\frac{(2-\mu)^2 + (3-\mu)^2 + (4-\mu)^2}{2 \cdot 9}\right] = \frac{1}{27} \exp\left[-\frac{3\mu^2 - 18\mu + 32}{18}\right]$$

This can be transformed in a Normal distribution easily

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto \frac{1}{27} \exp\left[-\frac{\mu^2 - 6\mu + \frac{32}{3}}{\frac{18}{3}}\right] = \frac{1}{27} \exp\left[-\frac{\mu^2 - 2 \cdot 3\mu + 9 - 9 + \frac{32}{3}}{\frac{18}{3}}\right] =$$

$$= \frac{1}{27} \exp\left[-\frac{-9 + \frac{32}{3}}{\frac{18}{3}}\right] \exp\left[-\frac{1}{2}\frac{(\mu - 3)^2}{\frac{18}{2 \cdot 3}}\right] \propto \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}} \exp\left[-\frac{1}{2}\frac{(\mu - 3)^2}{2}\right]$$

which is a Normal distribution with mean 3 and variance 2. In a general form, we have (Appendix 3.2)

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right] \propto \frac{1}{\sqrt{2\pi}\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}} \exp\left[-\frac{\left(\mu - \bar{y}\right)^2}{2\frac{\sigma^2}{n}}\right]$$

which is a Normal distribution with mean the sample mean and variance the given variance divided by the sample size.

3.6. Marginal distribution

3.6.1. *Definition*

We saw in 2.2.3 the advantages of marginalisation. When we have a bivariate density f(x, y), a marginal density f(x) is

$$f\left(x\right) = \int_{-\infty}^{\infty} f\left(x, y\right) dy = \int_{-\infty}^{\infty} f\left(x \mid y\right) f\left(y\right) dy$$

The marginal density f(x) takes the *average* of all values of 'y' for each x; i.e., takes all values of 'y', multiplies them by their probability and sums.

A density can be marginal for a variable and conditional for another one. For example, a marginal density conditioned in 'z' is

$$f\left(x \mid z\right) = \int_{-\infty}^{\infty} f\left(x, y \mid z\right) dy = \int_{-\infty}^{\infty} f\left(x \mid y, z\right) f\left(y \mid z\right) dy$$

where 'y' has been marginalised and 'z' is conditioning the values of 'x'. Notice that the marginalised variable 'y' does not appear in the formula because for each value of x all its possible values have been considered, multiplied by their respective probability and summed up, thus we do not need to give a value to 'y' in order to obtain f(x|z). However, the conditional variable appears in the formula because for

each given value of 'z' we obtain a different value of f(x|z). We will see an example in next paragraph.

3.6.2. *Marginal distribution of the variance of a normal distribution*

The marginal density of the variance *conditioned to the data* is

$$f\left(\sigma^2 \mid \mathbf{y}\right) = \int_{-\infty}^{\infty} f\left(\mu, \sigma^2 \mid \mathbf{y}\right) d\mu$$

Here the mean is *marginalised* and the data are *conditioning* the values of the variance, which means that we will obtain a different distribution for each sample. We will call this distribution "the marginal distribution of the variance" as a short name, because in Bayesian inference it is implicit that we are always conditioning in the data. Bayesian inference is always based in the sample, not in conceptual repetitions of the experiment; the sample is always 'given'.

We do not know $f\left(\mu, \sigma^2 \mid \mathbf{y}\right)$, but we can find it applying Bayes theorem because we know the distribution of the data $f\left(\mathbf{y} \mid \mu, \sigma^2\right)$. If the prior information $f\left(\mu, \sigma^2\right)$ is constant because we apply the principle of indifference as before, we have

$$f\left(\mu, \sigma^2 \mid \mathbf{y}\right) = \frac{f\left(\mathbf{y} \mid \mu, \sigma^2\right) f\left(\mu, \sigma^2\right)}{f\left(\mathbf{y}\right)} \propto f\left(\mathbf{y} \mid \mu, \sigma^2\right) f\left(\mu, \sigma^2\right) \propto f\left(\mathbf{y} \mid \mu, \sigma^2\right)$$

and the marginal density is

$$f\left(\sigma^2 \mid \mathbf{y}\right) = \int_{-\infty}^{\infty} f\left(\mu, \sigma^2 \mid \mathbf{y}\right) d\mu \propto \int_{-\infty}^{\infty} f\left(\mathbf{y} \mid \mu, \sigma^2\right) d\mu = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{n}{2}} \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_{1}^{n} \left(y_i - \mu\right)^2}{2\sigma^2}\right] d\mu$$

We solved this integral in Appendix 3.3, and we know that the solution is

$$f\left(\sigma^2 \mid \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \bar{y}\right)^2}{2\,\sigma^2}\right]$$

which is an Inverted Gamma distribution, as in 3.4.4, with parameters α, β

$$\beta = \frac{1}{2}\sum_1^n \left(y_i - \bar{y}\right)^2$$

$$\alpha = \frac{n-1}{2} - 1 = \frac{n-3}{2}$$

For example, if we take the same sample as in 3.4.4

y' = [2, 3, 4]

for this given sample, the marginal distribution of the variance is

$$f\left(\sigma^2 \mid \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{3-1}{2}}} \exp\left[-\frac{\left(2 - \frac{2+3+4}{3}\right)^2 + \left(3 - \frac{2+3+4}{3}\right)^2 + \left(4 - \frac{2+3+4}{3}\right)^2}{2\sigma^2}\right] = \frac{1}{\sigma^2} \exp\left[-\frac{1}{\sigma^2}\right]$$

Notice that the mean does not appear in the formula. In 3.5.4, when we calculated the density of the variance conditioned to the mean and to the data we had to give a value for the mean and we had to give the data. Here we only should give the data because the mean has been *marginalised.*

3.6.3. *Marginal distribution of the mean of a Normal distribution*

The marginal density of the mean *conditioned to the data* is

$$f\left(\mu \mid \mathbf{y}\right) = \int_0^\infty f\left(\mu, \sigma^2 \mid \mathbf{y}\right) d\sigma^2$$

Here the variance is *marginalised* and the data are *conditioning* the values of the variance, which means that we will obtain a different distribution for each sample. We will call this distribution "the marginal distribution of the mean" as a short name, because, as we said before, in Bayesian inference it is implicit that we are always conditioning in the data.

We do not know the function $f\left(\mu,\sigma^2 \mid \mathbf{y}\right)$, but applying Bayes theorem we can find it, because we now the distribution of the data, $f\left(\mathbf{y} \mid \mu,\sigma^2\right)$.

$$f\left(\mu,\sigma^2 \mid \mathbf{y}\right) = \frac{f\left(\mathbf{y} \mid \mu,\sigma^2\right) f\left(\mu,\sigma^2\right)}{f\left(\mathbf{y}\right)} \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right) f\left(\mu,\sigma^2\right)$$

if we admit the indifference principle to show vague prior information, then $f(\mu,\sigma^2)$ is a constant, thus

$$f\left(\mu,\sigma^2 \mid \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right)$$

and the marginal density of the mean is

$$f\left(\mu \mid \mathbf{y}\right) \propto \int f\left(\mathbf{y} \mid \mu,\sigma^2\right) d\sigma^2 = \int_0^{\infty} \frac{1}{(2\pi)^{\frac{n}{2}} \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right] d\sigma^2$$

This integral is solved in Appendix 3.4, and the result is

$$f\left(\mu \mid \mathbf{y}\right) \propto \left[1 + \frac{n}{n-1} \cdot \frac{\left(\mu - \bar{y}\right)^2}{s^2}\right]^{-\frac{n-2}{2}}$$

where

$$\bar{y} = \frac{1}{n}\sum_{1}^{n} y_i \quad ; \quad s^2 = \frac{1}{n-1}\sum_{1}^{n}(y_i - \bar{y})^2$$

This is a Student t-distribution with n–1 degrees of freedom, having a mean which is the sample mean and a variance that is the sample quasi-variance, thus

$$f(\mu \mid \mathbf{y}) \propto t_{n-1}(\bar{y}, s^2)$$

For example, if **y'** = [2, 3, 4]

$$\bar{y} = \frac{1}{3}(2+3+4) = 3$$

$$s^2 = \frac{1}{3-1}\left[(2-3)^2 + (3-3)^2 + (4-3)^2\right] = 1$$

$$f(\mu \mid \mathbf{y}) \propto \left[1 + \frac{3}{2-1}\cdot\frac{(\mu-3)^2}{1}\right]^{-\frac{3}{2}}$$

Notice that the variance does not appear in the formula. In 3.5.5, when we calculated the density of the mean conditioned to the variance and to the data, we had to give a value for the variance and we had to give the data. Here we only should give the data because the variance has been *marginalised.*

Appendix 3.1

We know a density f(x) and what we want is to find the density f(y) of a function

y = g(x)

We will first suppose that g(x) is a strictly increasing function, as in figure 3.5.a. By the definition of distribution function, we have

$F(y_0) = P(y \leq y_0) = P[g(x) \leq y_0]$

but as g(x) is an increasing function, we know that

$g(x) \leq y_0 \rightarrow x \leq x_0$

then, we have

$F(y_0) = P(x \leq x_0) = F(x_0) \rightarrow F(y) = F(x)$

because this applies for every x_0.

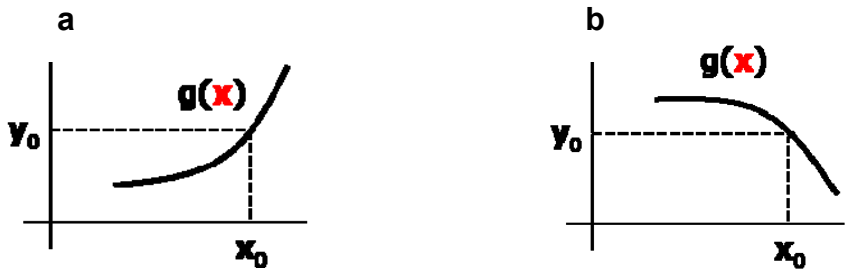

Figure 3.5. a. When g(x) is a monotonous increasing function. b when g(x) is a monotonous decreasing function.

Now, by definition of density function, we have

$$f(y) = \frac{dF(y)}{dy} = \frac{dF(x)}{dy} = \frac{dF(x)}{dx} \cdot \frac{dx}{dy} = f(x) \cdot \frac{dx}{dy}$$

Now suppose that g(x) is a strictly decreasing function, as in figure 3.5.b. By the definition of distribution function, we have

$F(y_0) = P(y \leq y_0) = P[g(x) \leq y_0]$

but as g(x) is an decreasing function, we know that

$g(x) \leq y_0 \rightarrow x \ x_0$

then, we have

$F(y_0) = P(x \geq x_0) = 1 - P(x \leq x_0) = 1 - F(x_0) \rightarrow F(y) = 1 - F(x)$

because this applies for every x_0. Now, by definition of density function, we have

$$f(y) = \frac{dF(y)}{dy} = \frac{d[1-F(x)]}{dx} \cdot \frac{dx}{dy} = -f(x) \cdot \frac{dx}{dy}$$

Finally, putting together the two cases, we have that

$$f(y) = f(x) \cdot \left|\frac{dx}{dy}\right| = f(x) \cdot \left|\frac{dy}{dx}\right|^{-1}$$

Appendix 3.2

We consider first that the numerator in the exp is

$$\sum_1^n (y_i - \mu)^2 = \sum_1^n (y_i - \bar{y} + \bar{y} - \mu)^2 = \sum_1^n \left[(y_i - \bar{y}) - (\mu - \bar{y})\right]^2 =$$

$$= \sum_1^n (y_i - \bar{y})^2 + \sum_1^n (\mu - \bar{y})^2 - 2\sum_1^n (y_i - \bar{y})(\mu - \bar{y}) = \sum_1^n (y_i - \bar{y})^2 + n(\mu - \bar{y})^2$$

because the double product is null

$$\sum_1^n (y_i - \bar{y})(\mu - \bar{y}) = (\mu - \bar{y})\sum_1^n (y_i - \bar{y}) = (\mu - \bar{y})\left[\sum_1^n (y_i) - n\bar{y}\right] = (\mu - \bar{y})\left[\sum_1^n (y_i) - n\frac{1}{n}\sum_1^n (y_i)\right] = 0$$

Then, substituting in the formula, we have

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2}{2\sigma^2}\right] \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\bar{y}\right)^2}{2\sigma^2}\right] \exp\left[-\frac{n\left(\mu-\bar{y}\right)^2}{2\sigma^2}\right] \propto$$

$$\propto \exp\left[-\frac{n\left(\mu-\bar{y}\right)^2}{2\sigma^2}\right] \propto \frac{1}{\sqrt{2\pi}\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}} \exp\left[-\frac{\left(\mu-\bar{y}\right)^2}{2\frac{\sigma^2}{n}}\right]$$

Appendix 3.3

$$f\left(\sigma^2 \mid \mathbf{y}\right) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2}{2\sigma^2}\right] d\mu$$

We can place out of the integral everything but μ, thus considering the factorization we have made in Appendix 3.2 of the numerator within the exp, we can write

$$f\left(\sigma^2 \mid \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\bar{y}\right)^2}{2\sigma^2}\right] \int \exp\left[-\frac{n\left(y_i-\mu\right)^2}{2\sigma^2}\right] d\mu$$

We can include within the integral any constant or variable with the exception of μ, thus we multiply out of the integral and divide inside the integral by the same expression, and write

$$f\left(\sigma^2 \mid \mathbf{y}\right) \propto \frac{\sqrt{2\pi}\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\bar{y}\right)^2}{2\,\sigma^2}\right] \int \frac{1}{\sqrt{2\pi}\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}} \exp\left[-\frac{\left(\mu-\bar{y}\right)^2}{2\,\frac{\sigma^2}{n}}\right] d\mu$$

The expression inside the integral is a density function (a normal function) of μ, and the integral is 1, thus

$$f\left(\sigma^2 \mid \mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \bar{y}\right)^2}{2\,\sigma^2}\right]$$

Appendix 3.4

$$f\left(\mu \mid \mathbf{y}\right) \propto \int_0^\infty \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right] d\sigma^2$$

We put the 2π in the constant of proportionality. We can include within the integral any constant or variable with the exception of σ^2, thus we divide out of the integral and multiply inside the integral by the same expression, and write

$$f\left(\mu \mid \mathbf{y}\right) \propto \frac{\Gamma(\alpha)}{\left[\sum_1^n \left(y_i - \mu\right)^2\right]^{\frac{n-2}{2}}} \int_0^\infty \frac{1}{\Gamma(\alpha)} \left[\frac{\sum_1^n \left(y_i - \mu\right)^2}{2}\right]^{\frac{n-2}{2}} \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right] d\sigma^2$$

Now, if we call x = σ^2, the inside part of the integral looks like

$$\mathbf{f}\left(x \mid \alpha, \beta\right) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{-(\alpha+1)} \exp\left[-\frac{\beta}{x}\right]$$

where

$$\beta = \frac{\sum_{1}^{n}(y_i - \mu)^2}{2} \quad ; \quad \alpha = \frac{n}{2} - 1 = \frac{n-2}{2}$$

This is, putting Γ(α) in the integration constant, a inverted gamma distribution, thus the value of the integral is 1. The, we have

$$f(\mu \mid \mathbf{y}) \propto \frac{1}{\left[\sum_{1}^{n}(y_i - \mu)^2\right]^{\frac{n-2}{2}}}$$

This can be transformed as follows; according to the factorization of Apendix 3.2:

$$f(\mu \mid \mathbf{y}) \propto \left[\sum_{1}^{n}(y_i - \mu)^2\right]^{-\frac{n-2}{2}} = \left[\sum_{1}^{n}(y_i - \bar{y})^2 + n(\mu - \bar{y})^2\right]^{-\frac{n-2}{2}} = \left[(n-1)s^2 + n(\mu - \bar{y})^2\right]^{-\frac{n-2}{2}} =$$

$$= (n-1)s^2\left[1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}\right]^{-\frac{n-2}{2}} \propto \left[1 + \frac{n}{n-1}\cdot\frac{(\mu - \bar{y})^2}{s^2}\right]^{-\frac{n-2}{2}}$$

where

$$\bar{y} = \frac{1}{n}\sum_{1}^{n} y_i \quad ; \quad s^2 = \frac{1}{n-1}\sum_{1}^{n}(y_i - \bar{y})^2$$

CHAPTER 4

MCMC

> Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell *(Biometrika,* Vol.I.p.219). The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains the measurements of 3000 criminals in a random order. Finally each consecutive set of 4 was taken as a sample—750 in all—and the mean, standard deviation, and correlation *r* of each sample determined.
>
> **William Searly Gosset** ("**Student**"), 1908.

4.1. Samples of Marginal Posterior distributions
- 4.1.1. Taking samples of Marginal Posterior distributions
- 4.1.2. Making inferences from samples of Marginal Posterior distributions

4.2. Gibbs sampling
- 4.2.1. How it works
- 4.2.2. Why it works
- 4.2.3. When it works
- 4.2.4. Gibbs sampling features
- 4.2.5. Example

4.3. Other MCMC methods
- 4.3.1. Acceptance-Rejection
- 4.3.2. Metropolis

Appendix 4.1

Appendix 4.2

4.1. Samples of marginal posterior distributions

4.1.1. *Taking samples of marginal posterior distributions*

We have seen in chapter 2 that two great advantages of Bayesian inference are *marginalisation* and the possibility of calculating actual *probability intervals* (called credibility intervals by Bayesians). Both to marginalize and to obtain these intervals, integrals should be performed. For very simple models this is not a difficulty, but the difficulty increases when models have several effects and different variance components. These difficulties stopped the progress of Bayesian inference for many years. Often the only practical solution was to find a multivariate mode, renouncing to the possibility of marginalisation. But this mode was given without any error or measure of uncertainty, because it was also necessary to calculate integrals to find credibility intervals. Most of these problems disappeared when it was made available a system of integration based in random sampling of Markov chains. Using these methods we do not obtain the posterior marginal distributions, but just random samples from them. This may look disappointing, but has many advantages as we will see soon.

Let us put an example. We need to find the marginal posterior distribution of the difference between the selected and the control group for the meat quality trait “intensity of flavour”, given the data, measured by a panel test in a scale from 1 to 5. But instead of this, we are going to obtain samples of the marginal posterior distribution of the selection and control effects given the data.

f(S | **y**): [3.1, 3.3, 4.1, 4.8, 4.9,…]
f(C | **y**): [2.4, 2.6, 2.6, 2.6, 2.8,...]

as both are random samples of the marginal posterior distributions of the effects, the difference sample by sample (i.e.; 3.1–2.4= 0.7, 3.3–2.6= 0.7, 4.1–2.6=1.5, 4.8–2.6= 2.2, etc.) gives a list of numbers that is a random sample of the difference between treatments

f(S–C | **y**) : [0.7, 0.7, 1.5, 2.2, 2.1,…]

These lists are called *Markov chains*. As they are formed by random samples, they are called "*Monte Carlo*". We can make a histogram with these numbers and obtain an approximation to the posterior distribution of S–C given the data **y** (figure 5.1).

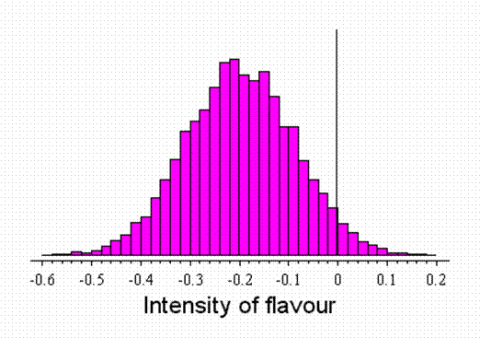

Figure 4.1 An histogram made by random sampling the posterior distribution of the difference between the selected and the control population f(S–C | y) for Intensity of flavour

From this random sample it is easy to make Bayesian inferences as we will see later. For example, if we want to estimate the mean of this posterior distribution we just calculate the average of the *chain* of numbers sampled from f(S–C | **y**).

This chain of sampled numbers from the posterior distribution can be as large as we want, thus we can estimate the posterior distribution as accurately as we need. Notice that it is not the same to estimate the posterior distribution with 500 samples than with 5,000 or 50,000. The samples can also be correlated. There is a sampling error that depends on the size of the sample but also on how correlated are the samples. For example, if we take 500 samples and the correlation between them is 1 we do not have 500 samples but always the same one. It can be calculated the "effective number" that we have; i.e., sample size of uncorrelated numbers that estimates the posterior distribution with the same accuracy that we do with our current chain of samples.

Another important point is that we can directly sample form marginal distributions. If we find a way to obtain random samples (x_i, y_i) of a joint posterior distribution f(x,y), each x_i is a random sample of the marginal distribution f(x) and each y_i is a random sample of the marginal distribution f(y). For example, if x and y can only take discrete

values 1, 2, 3, … , 10, we take a chain of 500 samples of the joint posterior distribution and we order this sample according to the values of x, we have

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1 & & 2\ 2\ 2\ 2\ 2\ 2 & & 3\ 3 & & 10 \\ 1,1,1,2,2,3,4,4,4,5,5,\cdots, & & 1,1,2,3,3,3,\cdots, & & 1,1,\cdots, & & 10 \end{pmatrix}$$

We have seen that in the continuous case a marginal density f(x) is

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

The equivalent in the discrete case is

$$f(x_i) = \frac{1}{n_{ik}} \sum_k f(x_i, y_k)$$

where n_{ik} is the number of samples (x_i, y_k). We can see that to calculate the frequency of x=1 we take all pairs of values (1,1), (1,1), (1,2), (1,2), (1,3),… which is to take all possible values of (1,y_k) and sum. Thus the first row is composed by random samples of the marginal distribution f(x) and the second row is composed by random samples of the marginal distribution f(y).

4.1.2. *Making inferences from samples of marginal posterior distributions*

From a chain of samples we can make inferences. Let us take the former example, in which we have a chain of random samples of the posterior distribution for the difference between the selected and the control population. We have now a chain of 30 samples. Let us order the chain from the lowest to the highest values

f(S–C | **y**) : [-0.2, -0.2, -0.1, -0.1, -0.1, 0.0, 0.0, 0.1, 0.2, 0.2, 0.2, 0.2, 0.5, 0.6, 0.6,
0.7, 0.7, 1.1, 1.1, 1.3, 1.5, 1.8, 1.8, 1.8, 1.8, 2.0, 2.0, 2.1, 2.1, 2.2]

Now we want to make the following inferences:

1) *Which is the probability of S being higher than C?* (Figure 2.6.a)

$P(S>C) = P(S–C>0)$

We estimate this probability counting how many samples higher than zero we have and divide by the total number of samples. We have 23 samples higher than zero from a total of 30 samples, thus our estimate is

$$P(S>C) = \frac{23}{30} = 0.77$$

2) *Which is the probability of the difference between groups being higher than 0.5?* (Figure 2.8.a)

We count how many samples are higher than 0.5. We find 17 samples higher than 0.5. As we have 30 samples, the probability of the difference between groups being higher than 0.5 is

$$P(S–C>0.5) = \frac{17}{30} = 0.57$$

3) *Which is the probability of the difference between groups being different from zero?* (Figure 2.9.a)

Strictly speaking, the probability of being different from zero is 1, since this difference will never take exactly the value 0.000000……., thus we have to define the minimum value of this difference from which lower values will be considered in practice as null. It is the same difference that is used in experimental designs when we decide that higher values will appear as significant and lower values as non significant. We call any higher value a *relevant* difference between groups or treatments. We decide that a *relevant* difference will be any one equal or higher than $\pm$ 0.1. We see that only two samples are lower than 0.1 and higher than -0.1, thus

$$P(|S—C| \geq \text{relevant}) = \frac{2}{30} = 0.97$$

4) *Which is the probability of the difference between groups being between 0.1 and 2.0?* (Figure 3.3.b)

We have 20 samples between both values (including them, thus this probability is

$$P(0.1 \leq S—C \leq 2.0) = \frac{20}{30} = 0.67$$

5) *Which is the minimum value that can take the difference between treatments, with a probability of 70%*? (Figure 2.7.a)

Let us take the *last* 70% of the samples of our ordered chain. A 70% of 30 samples is 21 samples, thus we take the *last* 21 samples of the chain. The first value of this set, which is the lowest one as well, is 0.2, thus we say that the difference between groups is at least 0.2 with a probability of 70%.

6) *Which is the maximum value that the difference between groups can take with a probability of 0.90?*

We take the *first* 90% of the samples of our ordered chain. A 90% of 30 samples are 27 samples, thus we take the first 27 samples, and the highest value of this set (the last sample) is 2.0. Thus we say that the difference between groups will be as a maximum 2.0 with a probability of 90%.

7) *Which is the shortest interval containing a 90% of probability?* (Figure 2.5.a)

The shortest interval (i.e., the most precise one) is calculated by considering all possible intervals containing the same probability. As a 90% of 30 samples is 27 samples, such an interval will contain 27 samples. Let us consider all possible intervals with 27 samples. These intervals are [-0.2, 2.0], [-0.2, 2.1], [-0.1, 2.2]. The

first interval has a length of 2.2, the second one 2.3 and the third one 2.3, thus the shortest interval containing a 90% of probability is [-0.2, 2.0].

8) *Give an estimate of the difference between groups*

Although it is somewhat illogical to say that this difference has a value just to say later that we are not sure about this value and we should give an interval, it is usual to give point estimates of the differences between treatments. We have seen that we can give the mean, median or mode of the posterior distribution. The mean is the average of the chain, and the median the value in the middle, the value between the sample 15 and 16.

Estimate of the mean and median of the posterior distribution P(S-C):

$$\text{Mean} = \frac{1}{30}\sum \text{ (-0.2, -0.2, -0.1, -0.1, -0.1, 0.0, 0.0, 0.1, 0.2, 0.2, 0.2, 0.2, 0.5, 0.6, 0.6, 0.7, 0.7, 1.1, 1.1, 1.3, 1.5, 1.8, 1.8, 1.8, 1.8, 2.0, 2.0, 2.1, 2.1, 2.2)} = 0.86$$

$$\text{Median} = \frac{0.6 + 0.7}{2} = 0.65$$

To estimate the mode, we need to draw the distribution, since we have a finite number samples and it is not possible to estimate with them accurately the mean (it ca happens, for example, that we have few samples of the most probable value).

In this example, mode and median differ, showing that the distribution is asymmetric. Which estimate should be given is a matter of opinion, we just should know the advantages and disadvantages, expressed in 2.2.1 ([22]).

[22] Sometimes the chains can sample outliers (particularly when we are making combinations of chains, for example finding the ratio of chains when the denominator is not far from zero). Another advantage of the median when working with MCMC is that the median is very robust to outliers.

4.2. Gibbs sampling

4.2.1. *How it works*

Now the question is how to obtain these samples. We will start with a simple example: how to obtain random samples from a joint posterior distribution f(x,y) that are also sets of samples from the marginal posterior distributions f(x), f(y). We will use MCMC techniques, and the most common one is called “Gibbs sampling” for reasons we commented in 1.1. What we need for obtaining these samples is:

1. To obtain univariate distributions of each unknown parameter *conditioned* to the other unknown parameters; i.e., to obtain f(x|y) and f(y|x).

2. To find a way to extract random samples from these conditional distributions.

The first step is easy, as we have seen in 3.4. The second step is easy if the conditional distribution have a recognisable form; i.e., they are Normal, Gamma, Poisson or other known distribution. We have algorithms that permit us to extract random samples of known distributions. For example,

a) Take a random sample x between 0 and 1 from a random number generator (all computers have this).

b) Calculate

$$y = \sqrt{-2\log x}\cdot\cos(2\pi x)$$

Then y is a random sample of a N(0,1).

When we do not have this algorithm because the conditional is not a known function or we do not have algorithms to extract random samples from it, other MCMC techniques can be used, but they are much more laborious as we will se later.

Once we have conditional functions from which we can sample, the Gibbs sampling mechanism starts as follows (Figure 4.1):

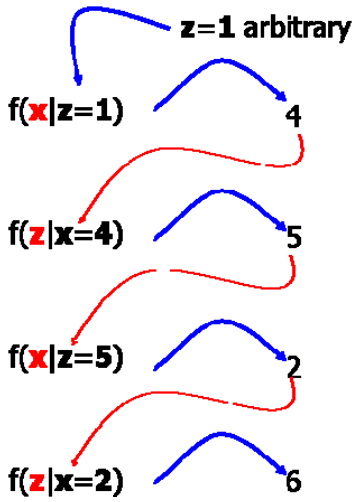

Figure 4.2. Gibbs sampling at work

1) Start with an arbitrary value for z, for example z=1
2) Extract one random sample from the conditional distribution f(x|z=1). Suppose this random sample is x=4
3) Extract one random sample from the conditional distribution f(z|x=4). Suppose this random sample is z=5
4) Extract one random sample from the conditional distribution f(x|z=5). Suppose this random sample is x=2
5) Extract one random sample from the conditional distribution f(z|x=4). Suppose this random sample is z=6
6) Continue with the process until obtaining two long chains

   x: 4, 2, …

   z: 5, 6, …
7) Disregard the first samples. We will see later how many samples should be disregarded.
8) Consider that the rest of the samples not disregarded are samples from the marginal distributions f(x) and f(z).

4.2.2. *Why it works*

Consider a simple example in order to understand this intuitively: to obtain a posterior distribution of f(x, z) sampling from the conditionals f(x | z) and f(z | x). Figure 4.1 shows f(x,z) represented as lines of equal probability (as level curves in a map). We suppose we have

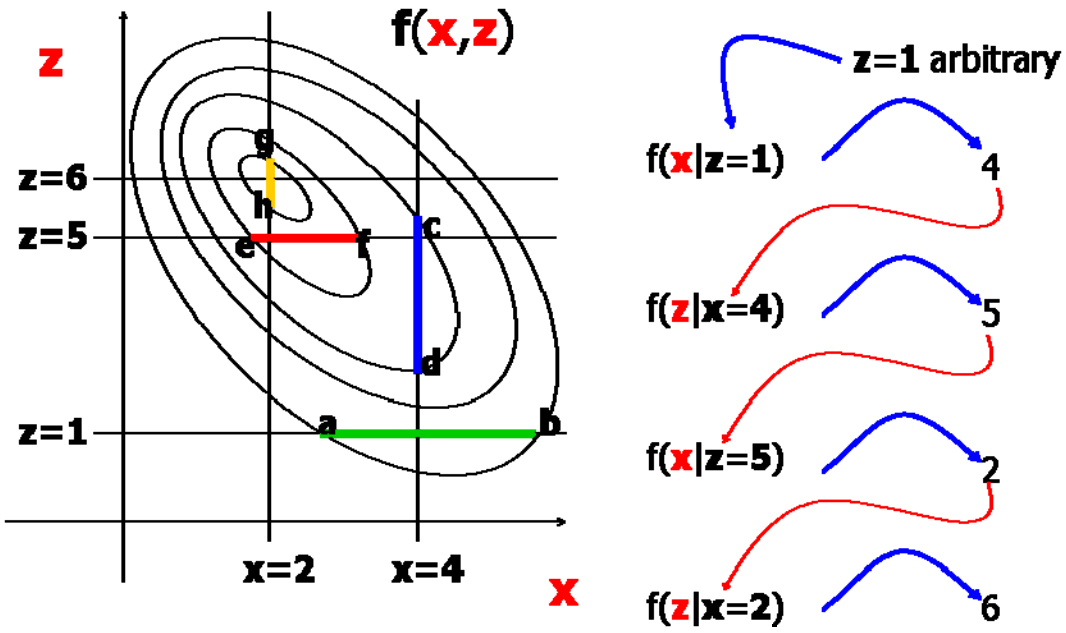

Figure 4.3. Gibbs sampling. The curves represent

Take an arbitrary value for z, say z=1. Sample a random number, which is the univariate density function that has all possible values for x but all values of z are z=1. This function is represented in figure 4.1 as the line z=1, that has more density of probability between a and b than in other parts of this line. Therefore, the number sampled from f(x | z=1) will be found between ‘a’ and ‘b’ more probably than in other parts of the conditional function. Suppose that the number sampled is x = 4. Now sample a random number from the conditional density f(z | x=4). This function is represented in figure 4.1 as the line x=4, that has more density of probability between ‘c’ and ‘d’ than in other parts of this line. Therefore, the number sampled from f(z | x=4) will be found between ‘c’ and ‘d’ more probably than in other parts of the

conditional function. Suppose that the number sampled is z = 5. Now sample a random number from the conditional density f(x | z=5). This function is represented in figure 4.1 as the line z=5, that has more density of probability between ‘e’ and ‘f’ than in other parts of this line. Therefore, the number sampled from f(x | z=5) will be found between ‘e’ and ‘f’ more probably than in other parts of the conditional function. Suppose that the number sampled is x = 2. Now sample a random number from the conditional density f(z | x=2). This function is represented in figure 4.1 as the line x=2, that has more density of probability between ‘g’ and ‘h’ than in other parts of this line. Therefore, the number sampled from f(z | x=2) will be found between ‘g’ and ‘h’ more probably than in other parts of the conditional function. Suppose that the number sampled is z=6, we will carry on with the same procedure until we obtain a chain of samples of the desired length.

Observe the tendency to sample from the highest areas of probability more often than to the lowest areas. At the beginning, z=0 and x=4 were points of the posterior distribution, but they were not random extractions, thus we were not interested on them. However, after many iterations, we find more samples in the highest areas of probability than in the lowest areas, thus we find random samples from the posterior distribution. This explains why the first points sampled should be discarded, and only after some cycles of iteration are samples taken at random.

4.2.3. *When it works*

1) *Strictly speaking, it cannot be demonstrated that we are finally sampling from a posterior distribution.* A Markov chain must be *reducible* to converge to a posterior distribution. Although it can be demonstrated that some chains are not reducible, there is no general procedure to ensure reducibility.

2) Even in the case in which the chain is reducible, *it is not known when sampling from the posterior distribution begins*.

3) Even having a reducible chain and when the tests ensure convergence, the converged distribution may not be stationary. Sometimes there are large

sequences of sampling that give the impression of stability, and after many iterations the chains moves to another area of stability.

The above problems are not trivial, and they occupy a part of the research in MCMC methods. Practically speaking, what people do is to launch several chains with different starting values and to observe their behaviour. No pathologies are expected for a large set of problems (for example, when using multivariate distributions), but some more complicated models (for example, threshold models with environmental effects in which no positives are observed in some level of one of the effects), should be examined with care. By using several chains, we arrive to an iteration from which the variability among chains may be attributed to Monte-Carlo sampling error, and thus support the belief that samples are being drawn from the posterior distribution. There are some tests to check whether this is the situation (Gelman and Rubin, 1992). Another possibility is to use the same seed and different initial values; in this case both chains should converge and we can establish a minimum difference between chains to accept the convergence (Johnson 1996). When having only one chain, a common procedure is to compare the first part and the last part of a chain (Geweke, 1992). Good practical textbooks dealing with MCMC application are Gelman et al. (2003), Gilks et al. (1996) and Robert and Casella (2004).

It should be noted that these difficulties are similar to finding a global maximum in multivariate likelihood with several fixed and random effects. With small databases maximum likelihood should not be used, since most properties of the method are asymptotic. With a large database, second derivative algorithms cannot be used by operative reasons, thus there is a formal incertitude about whether the maximum found is global or local. Here again, people use several starting values and examine the behaviour of their results. Complex models are difficult to handle in one or the other paradigm. However, MCMC techniques transform multivariate problems in univariate approaches, and inferences are made using probabilities, which has as easier interpretation.

4.2.4. *Gibbs sampling features*

To give an accurate description of the Gibbs sampling procedure used to estimate the marginal posterior distributions, in a scientific paper we should offer

In the Material and Methods section

1. *Number of chains:* When using several chains and they converge, we have the psychological persuasion that no convergence problems were found. We have no limit for the number of chains; in very simple problems, like linear models with treatment comparison, one chain is enough, but with more complex problems it is convenient to compute at least two chains. For very complex problems, ten or more chains can be computed. Whereas figure 4.4.a indicates that in a complex problem convergence arrived, figure 4.4.b questions convergence.

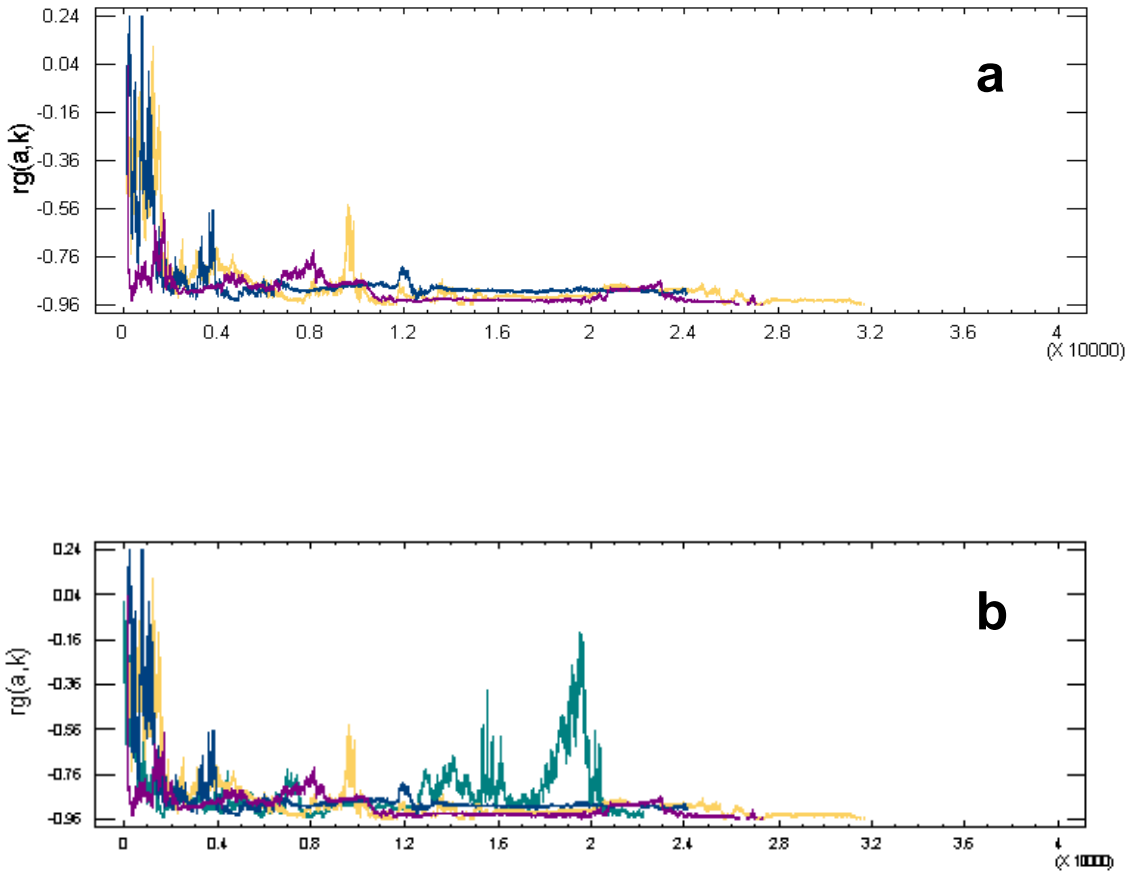

Figure 4.4 Several Gibbs sampling chains showing convergence (a) or not (b)

2. *Length of the chains*: It is customary to give the length of the chains in order to have an idea about the complexity of the problem. Very long chains are performed when there are convergence problems or when all the samples are extremely correlated.

3. *Burn-in*: We have seen in 4.3.2 that the first samples are not taken at random and should be disregarded. When we start considering that the samples are random samples from the joint (and consequently from the marginal) distributions is usually made by visual inspection of the chain. In many problems this is not difficult, and in simple problems convergence is raised after few iterations (figure 4.5). Although there are some methods to determine the burn-in (for example, Raftery and Lewis, 1992), they require the chain to have some properties that we do not know whether the chain actually has, thus visual inspection is a common method to determine the burn-in.

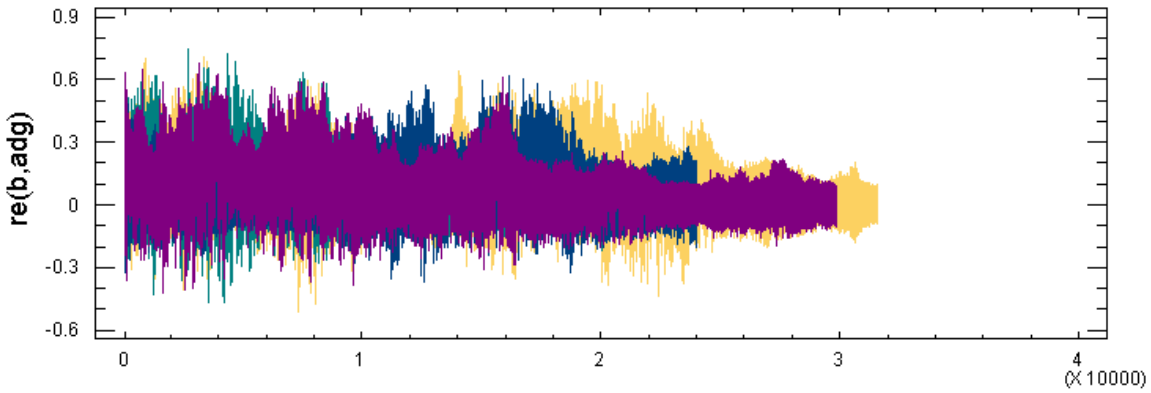

Figure 4.5 Several Gibbs sampling chains showing convergence from the first iterations

4. *Sampling lag*: Samples are correlated. We have seen in 4.2.2 how we start in sampling in a part of the function and that it is not probable to jump to the other side of the posterior distribution for a new sample. If the correlation between two successive samples is very high (say, 0.99) we will need more samples to obtain the same precision. For example, if the samples are independent, the sample mean has a variance which is the variance of the distribution divided by the number of samples (σ^2/n), but when the samples

are correlated, this variance is higher because the covariances should be taken into account. Collecting a number of samples high enough is not a problem, we can arrive to the accuracy we desire, but to avoid collecting a high number of samples, consecutive samples are disregarded and, for example, only one each 20 samples is collected, which decreases the correlation between two consecutive samples. This is called the 'lag' between samples.

5. *Actual sample size*: It is the equivalent number of independent samples that will have the same accuracy as the sample we have. For example, we can have 50.000 samples highly correlated that will lead to the same precision as 35 uncorrelated samples. The actual sample size gives an idea about the real sample size we have, since to have a high number of samples highly correlated does not give much information.

In Tables of results

6. *Convergence tests*: When having a positive answer from a convergence test we should say that "no lack of convergence was detected", because as we said before, there is no guaranty about the convergence. Some authors give the values of the tests; I think this is rather irrelevant as far as they did not detected lack of convergence.

7. *Monte Carlo s.e.* (may be accompained by efective sample size and autocorrelation): This is the error produced by the size of the sample. As we said before, it is not the same to estimate the posterior distribution with 500 samples than with 5,000 or 50,000 and the samples can also be correlated. This error is calculated usually in two ways, using groups of samples and examining the sample means, or using temporal series techniques. Current software for MCMC usually gives the MCse. We should augment the sample until this error becomes irrelevant (for example, when it is 10 times lower than the standard deviation of the posterior distribution).

8. *Point estimates*: Median, mean, mode. When distributions are approximately symmetrical we should give one of them. I prefer the median for reasons explained in this book, but the mean and the mode are more frequently given.

9. *Standard deviation of the posterior distribution*: When the distribution is approximately Normal, the s.d. is enough to calculate HPDs; for example, HPD95% will be approximately twice the standard deviation.

10. *Credibility intervals:* HPD, [k, +∞). We can give several intervals in the same paper; for example, [k, +∞) for 80, 90 and 95% of probability.

11. *Probabilities*: P(d)>0, P(d>Relevant), Probability of similarity.

4.3. Other MCMC methods

Not always we will have an algorithm that will provide us random samples of the conditional function. In this case we can apply several methods. Here we can only outline some of the most commonly used methods. There is a full area of research to find more efficient methods and to improve the efficiency of the existent ones. The reader interested in MCMC methods can consult Sorensen and Gianola (2002) or Gelman et al. (2003) for a more detailed account.

4.3.1. *Acceptance – rejection*

The acceptance-rejection method consists in covering the density f(x) from which we want to take random samples with a function g(x) (not a density) from which we have already an algorithm allowing us to take random samples (Figure 4.5). Thus by sampling many times and building a histogram, we can obtain a good representation of g(x). We call g(x) a “proposal function” because is the one we propose to sample.

Let us extract a random sample from g(x). Consider that, as in figure 4.6, the random sample x_0 extracted gives a value for $f(x_0)$ that is ½ of the value of $g(x_0)$. If we take many random samples of g(x) and build a histogram, half of them will be also random

samples of f(x), but which ones? To decide this we can throw a coin: if “face” it is a random sample of f(x), otherwise it is not. If we do this, when sampling many times and building a histogram, we will have a good representation of f(x) at $f(x_0)$.

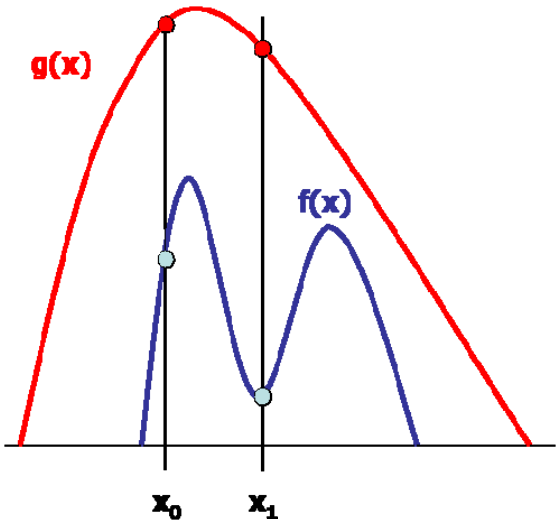

Figure 4.6. Acceptance – Rejection MCMC method. f(x) is the density from which we want to take random samples. g(x) is a function (not a density) from which we know how to take random samples, covering f(x). x_0 is a random sample from g(x) that might also be a random sample from f(x)

We will proceed in a similar manner for the following sample, x_1, examining the ratio between $f(x_1)$ and $g(x_1)$. If this ratio is 1/6, as in figure 4.5, we can throw a dice and decide that if we get a ‘1’ the sample is a random sample from f(x), otherwise it is not. Again, if we do this, when sampling many times and building a histogram, we will have a good representation of f(x) at $f(x_1)$.

The general procedure is as follows:

1) Take a random sample x_0 from g(x)
2) Sample a number k from the uniform distribution U[0,1]
3) If $k < \frac{f(x_0)}{g(x_0)}$ accept x_0 as a random sample of f(x), otherwise, reject it.

For example, we take a random sample of g(x) and we get $x_0 = 7$. We take a random sample from U[0,1] and we get k=0.3. We evaluate the ratio of functions at x_0 and we obtain $\frac{f(7)}{g(7)} = 0.8 > 0.3$, thus we accept that 7 is a random sample of f(x).

How to find a good g(x) is not always easy. We need to be sure it covers f(x), thus we need to know the maximum of f(x). Some g(x) functions can be very inefficient and most samples can be rejected, which obliges to sample many times. For example, in figure 4.7.a we have an inefficient function, x_0 is going to be rejected most of the times. We see in figure 4.7.b a better adapted function.

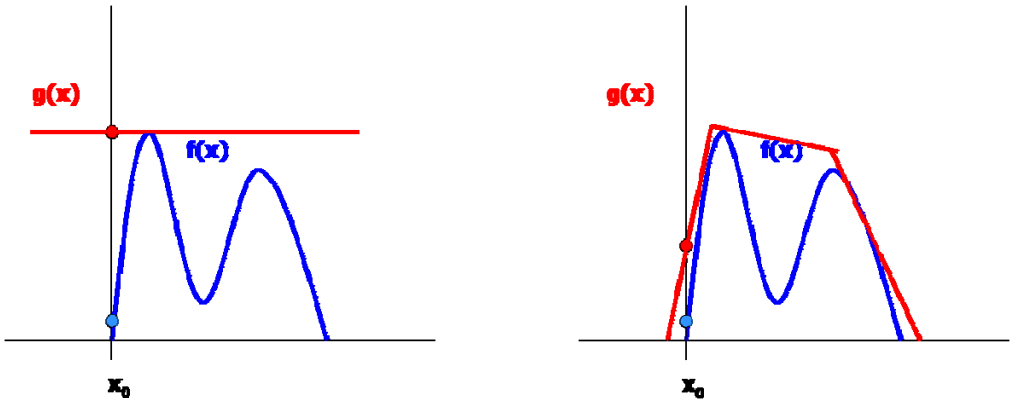

Figure 4.7. a. An easy but inefficient Accepting-Rejection function. b. A better adapted function

There are methods that search a new g(x) according to the success we had in the previous sampling, they are called “adaptive acceptance rejection sampling”. As I said before, there is a whole area of research in these topics.

4.3.2. *Metropolis-Hastings*

The method ([23]) has a rigorous proof based in the theory of Markov chains, that can be found for example in Sorensen and Gianola (2002). We will only expose here how it works.

[23] The following method was developed for symmetric proposal densities by Nicolas Metropolis in 1953, and it was improved for asymmetrical densities for Hastings (1970).

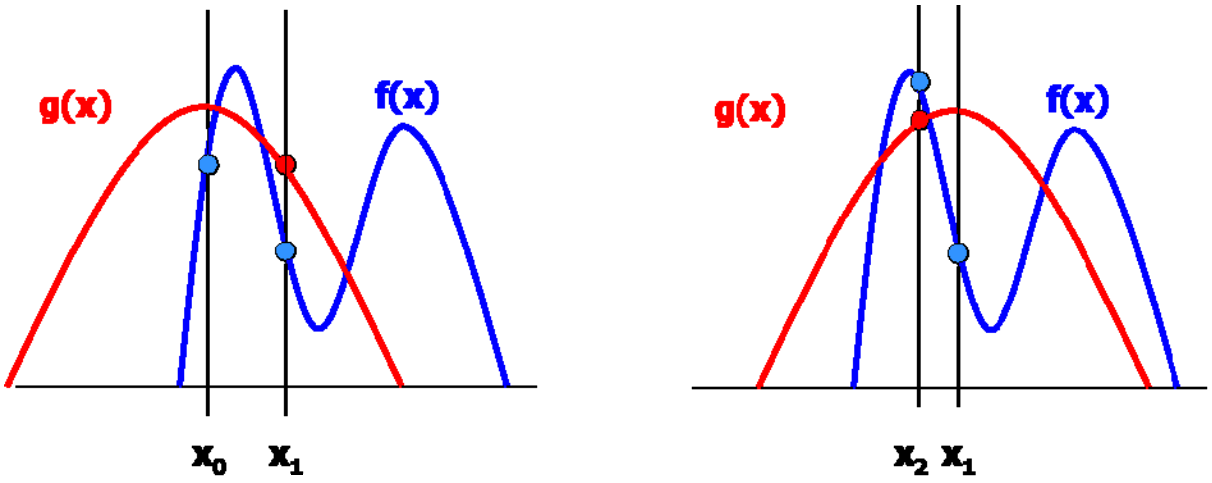

Figure 4.8. A proposal density g(x) from which we know how to take samples and the density function f(x) from which we need to take random samples. a. We start sampling

Our proposal function g(x) is now a density function ([24]), thus it should integrate to 1. We know how to take samples from g(x). The procedure works as follows:

1) Take a number x_0, arbitrary
2) Sample x_1 from g(x)
3) If $f(x_1)>f(x_0)$ accept x_1 as a sample of f(x), otherwise, as in figure 4.7.a, use Acceptance-rejection sampling as in 4.3.1. If x_1 is rejected, sample again from g(x) and repeat the process until you get a sample that is accepted.
4) If x_1 is accepted, *move* the mode of g(x) to x_1 as in figure 4.7.b, and sample again x_2 from g(x). Check whether $f(x_2)>f(x_1)$ and proceed as in 3).

Notice that by accepting x_1 when $f(x_1)>f(x_0)$, we tend to sample more often in the highest probability areas. We ensure we are sampling in the low probability areas by the acceptance-rejection method. When sampling many times and constructing a histogram, it will reflect the shape of f(x).

[24] Although we use the notation f(x) for density functions along this book, we will make an exception for g(x) in order to maintain the nomenclature used before for proposal functions.

A key issue of the method is to find the right proposal density g(x). It should be as similar as possible to the function f(x) from which we want to extract samples, in order to accept as many samples as possible when sampling. If we have a proposal density as in figure 4.8.a, we can accept many samples on the left part of f(x) but never move to the right part of f(x), and therefore we will not find a random sample of f(x) but only of a part of it. Simple functions like the ones in figure 4.7.a cover all the range of f(x) but many samples are rejected, thus it is inefficient. There is research on how to find efficient proposal densities, and there are adaptive Metropolis methods that try to change the proposal density along the sampling process to get more efficient g(x) densities.

Appendix Software for MCMC

There is software that can be used in several mainframes and in Microsoft-Windows PCs and that permits to analyze a large amount of statistical models.

BUGS permits to make inferences from a large amount of models. It is programmed in R-programming language and it allows adding instructions programmed in R. It is not charged by the moment, and it is widely used for Bayesian analyses. It can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs. It does not have the possibility of including the relationship matrix, but recently, Daamgard (2007) showed how to use BUGS for animal models.

TM is a Fortran90 software for multiple trait estimation of variance components, breeding values and fixed effects in threshold, linear and censored linear models in animal breeding. It has been developed by Luis Varona and Andrés Legarra, with intervention of some other people. It is not charged and can be obtained from its authors.

The program computes:

- Posterior distributions for variance components and relevant ratios (heritabilities, correlations).

• Posterior distributions for breeding values and fixed effects with known or unknown variance components.

The program handles:

- Any number of continuous traits.
- Several continuous traits, several polychotomous traits and one binary trait.
- Several binary traits (with some restrictions).
- Censored continuous traits.
- Missing values.
- Sire and animal models.
- Simultaneous correlated animal effects (e.g., sire-dam models for fertility, but not “reduced animal models” or maternal effects).
- Several random environmental effects (permanent effect).
- Different design matrices.
- It is possible to test contrasts of fixed or random effects.

The program does not handle:

- Covariates (neither random regression)
- Heterogeneous variances

MTGSAMTHR, by Van Tassell et al. (1996) **GIBBS90THR1** by Mizstal et al. (2002) are other options.

We will use in this course programs prepared by Wagdy Mekkawy for simple models. They can be obtained from the author.

CHAPTER 5

THE BABY MODEL

"All models are wrong, but some are useful"
George Box and **Norman Drapper**, 1983

5.1. The model
5.2. Analytical solutions
    5.2.1. Marginal posterior distribution of the mean and variance
    5.2.2. Joint posterior distribution of the mean and variance
    5.2.3. Inferences
5.3. Working with MCMC
    5.3.1. Using Flat priors
    5.3.2. Using vague informative priors
    5.3.3. Common misinterpretations
Appendix 5.1
Appendix 5.2
Appendix 5.3

5.1. The model

We will start this chapter with the simplest possible model, and we will see more complicated models in chapter 6. Our model consists only in a mean plus an error term ([25])

$y_i = \mu + e_i$

along the book we will consider that the data are normally distributed, although all procedures and conclusions can be applied to other distributions. All errors have mean zero and are uncorrelated, and all data have the same mean. Thus, to describe our model we will say that

$y_i \sim N(\mu, \sigma^2)$

$\mathbf{y} \sim N(\mathbf{1}\mu, \mathbf{I}\sigma^2)$

where **1'** = [1, 1, 1, ... , 1] and **I** is the identity matrix.

$$f\left(y_i \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right]$$

$$f\left(\mathbf{y} \mid \mu, \sigma^2\right) = f\left(y_1, y_2, \mathrm{K}, y_n \mid \mu, \sigma^2\right) = f\left(y_1 \mid \mu, \sigma^2\right) \cdot f\left(y_2 \mid \mu, \sigma^2\right) \mathrm{L}\ f\left(y_n \mid \mu, \sigma^2\right) =$$

$$= \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right] = \frac{1}{(2\pi)^{\frac{n}{2}} \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[\sum_1^n -\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right] \qquad (5.1)$$

as we saw in 3.5.3. Now we have to establish our objectives. What we want is to estimate the unknowns 'μ' and 'σ^2' that define the distribution.

[25] We call it in humoristic mood "the baby model".

5.2. Analytical solutions

5.2.1. *Marginal posterior density of the mean*

We will try to find the *marginal* posterior distributions for each unknown because this distribution takes into account the uncertainty when estimating the other parameter, as we have seen in chapters 2 and 3. Thus we should find

$$f(\mu \mid \mathbf{y}) = \int_0^{\infty} f(\mu, \sigma^2 \mid \mathbf{y}) d\sigma^2$$

$$f(\sigma^2 \mid \mathbf{y}) = \int_{-\infty}^{\infty} f(\mu, \sigma^2 \mid \mathbf{y}) d\mu$$

We have derived these distributions by calculating the integrals in chapter 3.

$$f(\mu \mid \mathbf{y}) \propto t_{n-1}(\bar{y}, s^2)$$

This is a "Student" t-distribution with parameters $\bar{y}$ and s^2, and n-1 degrees of freedom, where

$$\bar{y} = \frac{1}{n}\sum_1^n y_i \quad ; \quad s^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y})^2$$

The other marginal density we look for is (see Chapter 3)

$$f(\sigma^2 \mid \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n-1}{2}}} \exp\left[-\frac{\sum_1^n (y_i - \bar{y})^2}{2\,\sigma^2} \right] \propto \mathbf{IG}(\alpha, \beta)$$

which is an Inverted Gamma distribution with parameters α, β

$$\alpha = \frac{n-1}{2} - 1 \qquad ; \qquad \beta = \frac{1}{2}\sum_{1}^{n}(y_i - \bar{y})^2$$

5.2.2. *Joint Posterior density of the mean and variance*

We have seen that, using flat priors for the mean and variance,

$$f\left(\mu,\sigma^2 \mid \mathbf{y}\right) = \frac{f\left(\mathbf{y} \mid \mu,\sigma^2\right) f\left(\mu,\sigma^2\right)}{f\left(\mathbf{y}\right)} \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right) f\left(\mu,\sigma^2\right) \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right)$$

$$f\left(\mu,\sigma^2 \mid \mathbf{y}\right) \propto \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_{1}^{n}(y_i - \mu)^2}{2\sigma^2}\right]$$

now both parameters are in red because this is a bivariate distribution.

5.2.3. *Inferences*

We can draw inferences from the joint or from the marginal posterior distributions. For example, if we find the maximum from the joint posterior distribution, this is the most probable value for both parameters μ and σ^2 *simultaneously*, which is not the most probable value for the mean and the variance when all possible values of the other parameter have been weighted by their probability and summed up (i.e.: the mode of the marginal posterior densities of μ and σ^2). We will show now some inferences that have related estimators in the frequentist world.

Mode of the joint posterior density

To find the mode, as it is the maximum value of the posterior distribution, we derive and equal to zero (Appendix 5.2)

$$\text{mode } f\left(\mu,\sigma^2 \mid \mathbf{y}\right) \longrightarrow \begin{cases} \dfrac{\partial}{\partial\mu} f\left(\mu,\sigma^2 \mid \mathbf{y}\right) = 0 \longrightarrow \hat{\mu} = \dfrac{1}{n}\sum y_i = \bar{y} & \text{corresponding to } \hat{\mu}_{ML} \\ \dfrac{\partial}{\partial\sigma^2} f\left(\mu,\sigma^2 \mid \mathbf{y}\right) = 0 \longrightarrow \hat{\sigma}^2 = \dfrac{1}{n}\sum\left(y_i - \bar{y}\right)^2 & \text{corresponding to } \hat{\sigma}^2_{ML} \end{cases}$$

thus the mode of the joint posterior density give formulas that *look like* the maximum likelihood (ML) estimates of the variances, although here the interpretation is different because here they mean that this estimate is the most probable value of the unknowns μ and σ^2, whereas in a frequentist context this means that these values would make the sample most probable if they were the true values. The numeric value of the estimate is the same, but the interpretation is different.

Notice that we will not usually make inferences from joint posterior distributions because when estimating one of the parameters we do no take into account the uncertainty of estimating the other parameter.

Inferences from the marginal posterior density of the mean

As the marginal posterior distribution of the mean is a $t_{n-1}\left(\bar{y},s^2\right)$, the mean, median and mode are the same and they are equal to the sample mean. For credibility intervals we can consult a table of the t_{n-1} distribution.

Mode of the marginal posterior density of the variance

Deriving the marginal posterior density and equating to zero, we obtain (Appendix 5.3)

$$\text{mode } f\left(\sigma^2 \mid \mathbf{y}\right) \longrightarrow \frac{\partial}{\partial\sigma^2} f\left(\sigma^2 \mid \mathbf{y}\right) = 0 \longrightarrow \hat{\sigma}^2 = \frac{1}{n-1}\sum\left(y_i - \bar{y}\right)^2 \text{ corresponding to } \hat{\sigma}^2_{REML}$$

thus the mode of the joint posterior density give formulas that look like the maximum residual likelihood (REML) estimates of the variances, although here the interpretation is different because here this estimate is the most probable value of the unknown σ^2 when the values of the other unknown μ has been considered, weighted

by its probability and integrated out (summed up), whereas in a frequentist context we mean that this value would make the sample most probable if this would be the true value when working in a subspace in which there is no μ (see Blasco, 2001 for a more detailed interpretation). The numeric value of the estimate is the same, but the interpretation is different. Here the use of this estimate is more founded that in the frequentist case, but notice that the frequentist properties are different from the Bayesian ones, thus a good Bayesian estimator is not necessarily a good frequentist estimator, and vice versa.

Mean of the marginal posterior distribution of the variance

To calculate the mean of a distribution is not so simple because we need to calculate an integral. Because of that, modes were more popular before the era MCMC. However, we can have interest in calculating the mean because we do not like the loss function of the mode and we prefer the loss function of the mean as we have seen in chapter 2. To do this, by definition of mean, we have to calculate the integral

$$\text{mean } f\left(\sigma^2 \mid \mathbf{y}\right) = \int_0^{\infty} f\left(\sigma^2 \mid \mathbf{y}\right) f\left(\sigma^2\right) d\sigma^2$$

In this case we know that $f\left(\sigma^2 \mid \mathbf{y}\right)$ is an inverted gamma with parameters α and β that we have seen in 5.2.2. We can calculate the mean of this distribution if we have its parameters, and the formula can be found in several books (see, for example, Bernardo and Smith, 1994). Taking the value of α and β of paragraph 5.2.2, we have

$$\text{Mean}\left[\text{INVERTED GAMMA}\right] = \frac{\beta}{\alpha - 1} = \frac{\frac{1}{2}\sum_1^n \left(y_i - \mu\right)^2}{\frac{n-3}{2} - 1} = \frac{1}{n-5}\sum_1^n \left(y_i - \mu\right)^2$$

which does not have an equivalent in the frequentist world. This gives also a smaller estimate than the mode.

Notice that this estimate does not agree with the frequentist estimate of minimum quadratic risk that we saw in 1.4.4. This estimate had the same expression but dividing by n+1 instead of by n-5. The reason is on one side that we are not minimizing the same risk; in the frequentist case the variable is the estimate $\hat{\sigma}^2$, which is a combination of data, whereas in the Bayesian case the variable is the parameter σ^2, which is not any data combination. Thus when calculating the risk we integrate in one case on this combination of data and in the other vase the parameter.

Bayesian RISK $= E_u(\hat{u} - u)^2 = \int (\hat{u} - u)^2 f(u)\, du$

Frequentist RISK $= E_y(\hat{u} - u)^2 = \int (\hat{u} - u)^2 f(y)\, dy$

The other reason is that these Bayesian estimates have been derived under the assumption of flat (constant) prior information. These estimates will be different if other prior information is used. For example, if we take another common prior for the variance (we will see it in chapter 7)

$$f\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$$

we have

$$f\left(\sigma^2 \mid \mathbf{y}\right) \propto \frac{1}{\sigma^2} \cdot \frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \bar{y}\right)^2}{2\,\sigma^2}\right] \propto \frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}-1}} \exp\left[-\frac{\sum_1^n \left(y_i - \bar{y}\right)^2}{2\,\sigma^2}\right]$$

then, calculating the mean as before

$$\text{Mean}\left[\text{INVERTED GAMMA}\right] = \frac{\beta}{\alpha - 1} == \frac{\frac{1}{2}\sum_1^n \left(y_i - \mu\right)^2}{\frac{n-1}{2} - 1} = \frac{1}{n-3}\sum_1^n \left(y_i - \mu\right)^2$$

which is different from the former estimate. When the number of samples is high all estimates are similar, but if 'n' is low we will get different results, stressing then that prior information is important in Bayesian analyses if the samples are small.

Median of the posterior marginal distribution of the variance

We have stressed in chapter 2 the advantages of the median as an estimator that uses a reasonable loss function and that it is invariant to transformations. The median m_y is

$$\text{median } f(\sigma^2 \mid y) \rightarrow \int_{-\infty}^{m_y} f(\sigma^2 \mid y)dy = \frac{1}{2}$$

thus we must calculate the integral.

Credibility intervals between two values 'a' and 'b'

The probability that the true value lies between 'a' and 'b' is

$$P\left(a < \sigma^2 < b\right) = \int_a^b \mathbf{f}\left(\sigma^2 \mid \mathbf{y}\right) f\left(\sigma^2\right) d\sigma^2$$

thus we must calculate the integral.

5.3. Working with MCMC

5.3.1. *Using flat priors*

To work with MCMC-Gibbs sampling we need o calculate the *conditional* distributions of the parameters $f\left(\mu \mid \mathbf{y}, \sigma^2\right)$ and $f\left(\sigma^2 \mid \mathbf{y}, \mu\right)$. We do not know them, but we can calculate them using Bayes theorem. Using flat priors

$$f\left(\mu \mid \mathbf{y},\sigma^2\right) \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right) f(\mu) \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right)$$

$$f\left(\sigma^2 \mid \mathbf{y},\mu\right) \propto f\left(\mathbf{y} \mid \sigma^2,\mu\right) f\left(\sigma^2\right) \propto f\left(\mathbf{y} \mid \sigma^2,\mu\right)$$

as we know the distribution of the data, we can obtain both conditionals

$$f\left(\mu \mid \mathbf{y},\sigma^2\right) \propto f\left(\mathbf{y} \mid \mu,\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2}{2\sigma^2}\right]$$

$$f\left(\sigma^2 \mid \mathbf{y},\mu\right) \propto f\left(\mathbf{y} \mid \sigma^2,\mu\right) \propto \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2}{2\sigma^2}\right]$$

Notice that the formulae are the same, but the variable is in red, thus the functions are completely different. As we saw in chapter 3, the first is a normal distribution and the second an inverted gamma distribution.

$$f\left(\mu \mid \mathbf{y},\sigma^2\right) \propto N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

$$f\left(\sigma^2 \mid \mathbf{y},\mu\right) \propto \mathbf{IG}(\alpha,\beta) \quad ; \quad \beta = \frac{1}{2}\sum_1^n \left(y_i-\mu\right)^2 \quad ; \quad \alpha = \frac{n}{2}-1$$

We have algorithms to sample from normal and inverted gamma functions, thus we can take random samples of them. We start, for example, with an arbitrary value for the variance and then we get a sample value of the mean. We substitute this value in the conditional of the mean and we get a random value of the variance. We substitute it in the conditional distribution of the mean and we continue the process (figure 5.1)

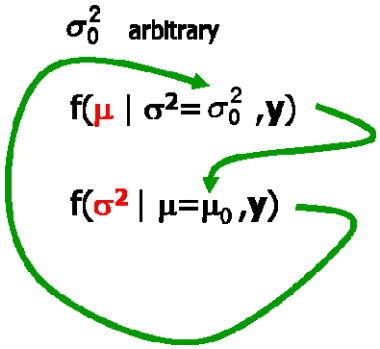

Figure 5.1. Gibbs sampling process for the mean and the variance of the baby model

We will put an example. We have a data vector with four samples

y' = [2, 4, 4, 2]

then we calculate

$$\bar{y} = 3, \quad n = 4, \quad \sum_1^n y_i^2 = 40$$

and we can prepare the first conditional distributions

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto N\left(\bar{y}, \frac{\sigma^2}{n}\right) \equiv N\left(3, \frac{\sigma^2}{4}\right)$$

$$f\left(\sigma^2 \mid \mu, \mathbf{y}\right) \propto \text{lgamma}\begin{cases}\beta = \frac{1}{2}\left[\sum_1^n y_i^2 - \mu^2\right] \\ \alpha = \frac{n}{2} - 1\end{cases} \equiv \text{lgamma}\begin{cases}\beta = \frac{1}{2}\left(40 - \mu^2\right) \\ \alpha = 1\end{cases}$$

Now we start the Gibbs sampling process by taking an arbitrary value for σ^2, for example

$$\sigma_0^2 = 1$$

then we substitute this arbitrary value in the first conditional distribution and we have

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto N\left(3, \frac{1}{4}\right)$$

we sample from this distribution using an appropriate algorithm and we find

$\mu_0 = 4$

then we substitute this sampled value in the second conditional distribution,

$$f\left(\sigma^2 \mid \mu, \mathbf{y}\right) \propto \text{Igamma}\left[12,\ 1\right]$$

now we sample from this distribution using an appropriate algorithm and we find

$$\sigma_1^2 = 5$$

then we substitute this sampled value in the first conditional distribution,

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto N\left(3, \frac{5}{4}\right)$$

now we sample from this distribution using an appropriate algorithm and we find

$\mu_1 = 3$

then we substitute this sampled value in the second conditional distribution, and continue the process. Notice that we sample each time from a different conditional distribution. The first conditional distribution of μ was a normal with mean equal to 3 and variance equal to 1, but the second time we sampled it was a normal with the same mean but with variance equal to 5. The same happened with the different Inverted Gamma distributions from which we were sampling.

We obtain two chains. All the samples belong to different conditional distributions, but after a while they are also samples from the respective marginal posterior distributions (figure 5.2). After rejecting the samples of the "burning period", we can use the rest of the samples for inferences as we did in chapter 4 with the MCMC chains.

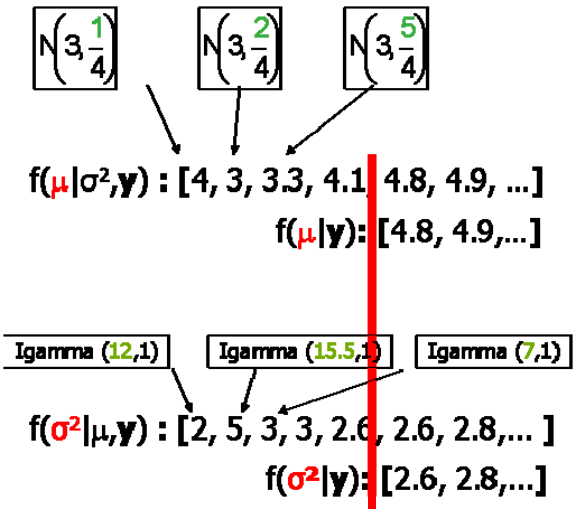

Figure 5.2. A Gibbs sampling process. Samples are obtained from different conditional distributions, but after a "burning" in which samples are rejected, the rest of them are also samples of the marginal posterior distributions.

5.3.2. *Using vague informative priors*

As we saw in chapter 2 and we will see again in chapter 9 with more detail, vague informative priors should reflect the beliefs of the researcher, beliefs that are supposed to be shared by the scientific community to a higher or lesser degree. They need not to be precise, since if they were too precise the data will not bring much information to the experiment and there will not be any need to perform the experiment.

There are many density functions that can express our vague beliefs. For example, we may have an *a priori* expectation of obtaining a difference of 100 g of liveweight between the selected and control rabbit population of our example in chapters 1 and 2. We believe that it is less probable to obtain 50 g, as much as 150 g. We also

believe that it is rather improbable to obtain 25 g, as improbable as to obtain 175 g of difference between the selected and control populations. These beliefs are symmetrical around the most probable value, 100 g, and can be approximately represented by a normal distribution, but also by a t-distribution or a Cauchy distribution. We will choose the most convenient distribution in order to facilitate the way of obtaining the conditional distributions we need for the Gibbs sampling. The same can be said about the variance, but here our beliefs are typically asymmetric because we are using square measurements. For example, we will not believe that the heritability of growth rate is going to be 80% or 90%, even if our expectations are around 40% we will tend to believe that lower values are more probable than higher values. These beliefs can be represented by many density functions, but again we will choose the ones that will facilitate our task of obtaining the conditional distributions we need for Gibbs sampling. Figure 2.2 shows an example of different prior beliefs represented by inverted gamma distributions.

Vague Informative priors for the variance

Let us take an inverse gamma distribution to represent our asymmetric beliefs for the variance. This function can show very different shapes by changing its parameters α and β.

$$f\left(\sigma^2\right) = \frac{1}{\left(\sigma^2\right)^{\alpha+1}} \exp\left[-\frac{\beta}{\sigma^2}\right]$$

again, as in 5.3.1, we do not the conditional distribution of the variance, but we can calculate it using Bayes theorem.

$$f\left(\sigma^2 \mid \mathbf{y}, \mu\right) \propto f\left(\mathbf{y} \mid \sigma^2, \mu\right) f\left(\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right] \cdot \frac{1}{\left(\sigma^2\right)^{\alpha+1}} \exp\left[-\frac{\beta}{\sigma^2}\right]$$

$$\propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}+\alpha+1}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2 - 2\beta}{2\sigma^2}\right]$$

which is an inverted gamma with parameters ‘a’ and ‘b’

$$a = \frac{n}{2} + \alpha$$

$$b = \frac{1}{2}\sum_1^n \left(y_i-\mu\right)^2 - \beta$$

Vague Informative priors for the mean

Now we use an informative prior for the mean. As we said before, our beliefs can be represented by a normal distribution. Thus we determine the mean and variance of our beliefs, ‘m’ and ‘v’ respectively

$$f\left(\mu\right) \propto \frac{1}{\left(v^2\right)^{\frac{1}{2}}} \exp\left[-\frac{\left(\mu-m\right)^2}{2v^2}\right]$$

We do not know the conditional mean but we can know it by using Bayes theorem as before,

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mu, \sigma^2\right) f\left(\mu\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i-\mu\right)^2}{2\sigma^2}\right] \cdot \frac{1}{\left(v^2\right)^{\frac{1}{2}}} \exp\left[-\frac{\left(\mu-m\right)^2}{2v^2}\right]$$

after some algebra gymnastics (Appendix 5.3) this becomes

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto \exp\left[-\frac{1}{2} \cdot \frac{\left(\mu-w\right)^2}{d^2}\right] \propto N\left(w, d^2\right)$$

the values of w and d^2 can be found in Appendix 5.4. This is a normal function with mean w and variance d^2 and we know how to sample from a normal distribution. Thus, we can start with the Gibbs sampling mechanism as in 5.3.1.

5.2.3. Common misinterpretations

The parameters of the inverted gamma distribution are degrees of freedom: The concept of degrees of freedom was developed by Fisher (1922), who represented the sample in a space of n-dimensions (see Blasco 2001 for an intuitive representation of degrees of freedom). This has no relationship with what we want. We manipulate the parameters in order to change the shape of the function, and it is irrelevant whether these "hyper parameters" are natural numbers or fractions. For example, Blasco et al. (1998) use fractions for these parameters.

One of the parameters of the inverted gamma distribution represents credibility and the other represents the variance of the function: Both parameters modify the shape of the function in both senses: dispersion and sharpness showing more credibility, thus it is incorrect to name one of them as the parameter of credibility and to use standard deviations or variances coming from other experiments as a parameter of the inverted gamma distribution. Both parameters should be manipulated in order to obtain a shape that will show our beliefs, and it is irrelevant which values they have as far as the shape of the function represents something similar to our state of beliefs.

Appendix 5.1

$$\frac{\partial}{\partial\mu} f\left(\mu,\sigma^2 \mid \mathbf{y}\right) = \frac{\partial}{\partial\mu} \quad \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n (y_i-\mu)^2}{2\sigma^2}\right] =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \cdot \frac{2\sum_1^n (y_i-\mu)}{2\sigma^2} \cdot \exp\left[-\frac{\sum_1^n (y_i-\mu)^2}{2\sigma^2}\right] = 0$$

$$\sum_1^n (y_i-\hat{\mu}) = 0 \longrightarrow \sum_1^n y_i - n\hat{\mu} = 0 \longrightarrow \hat{\mu} = \frac{1}{n}\sum_1^n y_i$$

$$\frac{\partial}{\partial\sigma^2} f\left(\mu,\sigma^2 \mid \mathbf{y}\right) = \frac{\partial}{\partial\sigma^2} \quad \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n (y_i-\mu)^2}{2\sigma^2}\right] =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}}} \cdot \frac{\sum_1^n (y_i-\mu)}{2\left(\sigma^2\right)^2} \cdot \exp\left[-\frac{\sum_1^n (y_i-\mu)^2}{2\sigma^2}\right] + \frac{-\frac{n}{2}}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}+1}} \cdot \exp\left[-\frac{\sum_1^n (y_i-\mu)^2}{2\sigma^2}\right] = 0$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}+2}} \cdot \frac{\sum_1^n (y_i-\mu)}{2} + \frac{2\left(-\frac{n}{2}\right)\sigma^2}{2(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n}{2}+2}} = 0$$

$$\sum_1^n (y_i-\hat{\mu})^2 - n\cdot\hat{\sigma}^2 = 0 \longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_1^n (y_i-\hat{\mu})^2$$

Appendix 5.2

$$\frac{\partial}{\partial\sigma^2}\mathbf{f}\left(\sigma^2\,|\,\mathbf{y}\right)\propto\frac{\partial}{\partial\sigma^2}\quad\frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}}}\exp\left[-\frac{\sum_1^n\left(y_i-\bar{y}\right)^2}{2\,\sigma^2}\right]=$$

$$=\frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}}}\cdot\frac{\sum_1^n\left(y_i-\mu\right)}{2\left(\sigma^2\right)^2}\cdot\exp\left[-\frac{\sum_1^n\left(y_i-\mu\right)^2}{2\sigma^2}\right]+\frac{-\frac{n-1}{2}}{\left(\sigma^2\right)^{\frac{n-1}{2}+1}}\cdot\exp\left[-\frac{\sum_1^n\left(y_i-\mu\right)^2}{2\sigma^2}\right]=0$$

$$=\frac{1}{\left(\sigma^2\right)^{\frac{n-1}{2}+2}}\cdot\frac{\sum_1^n\left(y_i-\mu\right)}{2}+\frac{2\left(-\frac{n-1}{2}\right)\sigma^2}{2\left(2\pi\right)^{\frac{n}{2}}\left(\sigma^2\right)^{\frac{n-1}{2}+2}}=0$$

$$\sum_1^n\left(y_i-\hat{\mu}\right)^2-\left(n-1\right)\cdot\hat{\sigma}^2=0\quad\longrightarrow\quad\hat{\sigma}^2=\frac{1}{n-1}\sum_1^n\left(y_i-\hat{\mu}\right)^2$$

Appendix 5.3

$$\mathrm{f}\left(\mu\,|\,\sigma^2,\mathbf{y}\right)\propto\mathrm{f}\left(\mathbf{y}\,|\,\mu,\sigma^2\right)\mathrm{f}\left(\mu\right)\propto\frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}}\exp\left[-\frac{\sum_1^n\left(y_i-\mu\right)^2}{2\sigma^2}\right]\cdot\frac{1}{\left(v^2\right)^{\frac{1}{2}}}\exp\left[-\frac{\left(\mu-m\right)^2}{2v^2}\right]$$

but we saw in chapter 3, Appendix 3.2, that

$$\mathrm{f}\left(\mathbf{y}\,|\,\mu,\sigma^2\right)\propto\frac{1}{\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}}\exp\left[-\frac{\left(\mu-\bar{y}\right)^2}{2\frac{\sigma^2}{n}}\right]$$

then, substituting, we have

$$f\left(\mu \mid \sigma^2, \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mu, \sigma^2\right) f(\mu) \propto \frac{1}{\left(\frac{\sigma^2}{n}\right)^{\frac{1}{2}}} \exp\left[-\frac{(\mu-\bar{y})^2}{2\frac{\sigma^2}{n}}\right] \cdot \frac{1}{\left(v^2\right)^{\frac{1}{2}}} \exp\left[-\frac{(\mu-m)^2}{2v^2}\right] \propto$$

$$\propto \frac{1}{\left(\frac{\sigma^2}{n} \cdot v^2\right)^{\frac{1}{2}}} \exp\left[-\frac{(\mu-\bar{y})^2}{2\frac{\sigma^2}{n}} - \frac{(\mu-m)^2}{2v^2}\right] \propto \exp\left[-\frac{v^2(\mu-\bar{y})^2 + \frac{\sigma^2}{n}(\mu-m)^2}{2\frac{\sigma^2}{n}v^2}\right]$$

Now, the exponential can be transformed if we take into account that

$$v^2(\mu-\bar{y})^2 + \frac{\sigma^2}{n}(\mu-m)^2 = \left(v^2 + \frac{\sigma^2}{n}\right)\mu^2 - 2\left(v^2\bar{y} + \frac{\sigma^2}{n}m\right)\mu + v^2\bar{y}^2 + \frac{\sigma^2}{n}m^2 \quad \propto$$

$$\propto \left(v^2 + \frac{\sigma^2}{n}\right)\mu^2 - 2\left(v^2\bar{y} + \frac{\sigma^2}{n}m\right)\mu = \frac{\mu^2 - 2\frac{\left(v^2\bar{y} + \frac{\sigma^2}{n}m\right)}{\left(v^2 + \frac{\sigma^2}{n}\right)}\mu}{\frac{1}{\left(v^2 + \frac{\sigma^2}{n}\right)}}$$

calling

$$w = \frac{\left(v^2\bar{y} + \frac{\sigma^2}{n}m\right)}{\left(v^2 + \frac{\sigma^2}{n}\right)}$$

and substituting in the expression, it becomes

$$\frac{\mu^2 - 2w\mu}{\frac{1}{v^2 + (\sigma^2/n)}} = \frac{\mu^2 - 2w\mu + w^2 - w^2}{\frac{1}{v^2 + (\sigma^2/n)}} \propto \frac{(\mu - w)^2}{\frac{1}{v^2 + (\sigma^2/n)}}$$

$$f(\mu \mid \sigma^2, \mathbf{y}) \propto \exp\left[-\frac{(\mu - w)^2}{2\frac{(\sigma^2/n)\cdot v^2}{(\sigma^2/n) + v^2}} \right]$$

calling

$$\frac{1}{d^2} = \frac{1}{\sigma^2/n} + \frac{1}{v^2} = \frac{(\sigma^2/n) + v^2}{(\sigma^2/n)\cdot v^2}$$

we have

$$f(\mu \mid \sigma^2, \mathbf{y}) \propto \exp\left[-\frac{(\mu - w)^2}{2\cdot d^2} \right] \quad \propto N(w, d^2)$$

CHAPTER 6

THE LINEAR MODEL

> "If any quantity has been determined by several direct observations, made under the same circumstances and with equal care, the arithmetic mean of the observed values gives the most probable value".
>
> **Carl Friedrich Gauss** (1809).

6.1. The "fixed" effects model
- 6.1.1. The model
- 6.1.2. Marginal posterior distributions via MCMC using Flat priors
- 6.1.3. Marginal posterior distributions via MCMC using vague informative priors
- 6.1.4. Least Squares as a Bayesian Estimator

6.2. The "mixed" model
- 6.2.1. The model
- 6.2.2. Marginal posterior distributions via MCMC
- 6.2.3. BLUP as a Bayesian estimator
- 6.2.4. REML as a Bayesian estimator

6.3. The multivariate model
- 6.3.1. The model
- 6.3.2. Data augmentation

Appendix 6.1

6.1. The “fixed” effects model

6.1.1. *The model*

The model corresponds, in a frequentist context, to a “fixed effects model” with or without covariates. In a Bayesian context all effects are random, thus there is no distinction between fixed models, random models or mixed models. We will describe here the Normal linear model, although other distributions of the data can be considered, and the procedure will be the same. Our model consists in a set of effects and covariates plus an error term. For example, if we measure the weight at weaning of a rabbit and we have a season effect (with two levels) and a parity effect (with two levels) plus a covariate ‘weight of the dam’, the model will be

$$y_{ijkl} = \mu + S_i + P_j + b \cdot A_{ijk} + e_{ijkl}$$

where S is the season effect, P the parity effect and A the age of the dam. As there are several piglets in a litter, y_{ijk} is the weight of the piglet l of a dam that belongs to the herd i, the piglet was born in the parity j, the dam had an age A_{ijk} that was the same for all piglets born in the same litter. For example, in the following equations we have two rabbits of the same litter weighting 520 and 430 grams, born in season 1 and the second parity of the dam 111 that weighted 3400 grams. Then we have a rabbit that weighted 480 grams, born in the second season and the first parity of dam 221 that weighted 4200 grams, and finally we have a rabbit weighting 550 grams, born in season 1 and parity 1 of the dam 112 that weighted 4500 grams.

$$520 = \mu + S_1 + P_2 + b \cdot 3400 + e_{1111}$$
$$430 = \mu + S_1 + P_2 + b \cdot 3400 + e_{1112}$$
$$480 = \mu + S_2 + P_2 + b \cdot 4200 + e_{2211}$$
$$550 = \mu + S_1 + P_1 + b \cdot 4500 + e_{1121}$$

In matrix form,

$$\begin{bmatrix} 520 \\ 430 \\ 480 \\ 550 \\ L \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 3400 \\ 1 & 1 & 0 & 0 & 1 & 3400 \\ 1 & 0 & 1 & 0 & 1 & 4200 \\ 1 & 1 & 0 & 1 & 0 & 4500 \\ L & & L & L & L & L \end{bmatrix} \cdot \begin{bmatrix} \mu \\ S_1 \\ S_2 \\ P_1 \\ P_2 \\ b \end{bmatrix} + \begin{bmatrix} e_{1111} \\ e_{1112} \\ e_{2211} \\ e_{1121} \\ L \end{bmatrix}$$

And, in general form,

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

where **y** contains the data, **X** is a matrix containing the covariates and the presence (1) or absence (0) of the levels of the effects. **b** is a vector of all unknowns and **e** is a vector with the errors. We consider the errors having mean zero and being independently normally distributed, all of them with the same variance,

$$\mathbf{e} \mid \sigma^2 \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$$

as in a Bayesian context there are not fixed effects, the correct way of expressing the distribution of the data is

$$\mathbf{y} \mid \mathbf{b} \sim N(\mathbf{Xb}, \mathbf{I}\sigma^2)$$

$$f\left(\mathbf{y} \mid \mathbf{b}, \sigma^2\right) \propto \frac{1}{\left|\mathbf{I}\sigma^2\right|^{\frac{1}{2}}} \exp\left[-\frac{\left(\mathbf{y}-\mathbf{Xb}\right)'\left[\mathbf{I}\sigma^2\right]^{-1}\left(\mathbf{y}-\mathbf{Xb}\right)}{2}\right] \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\left(\mathbf{y}-\mathbf{Xb}\right)'\left(\mathbf{y}-\mathbf{Xb}\right)}{2\sigma^2}\right]$$

In a Bayesian context, to completely specify the model we also need the prior distribution of the unknowns. We will consider below two cases as in the Baby model: flat and conjugated priors.

Our objective is to find the marginal posterior distributions of all unknowns; in our example

$f(\mu|\mathbf{y})$, $f(S_1|\mathbf{y})$, $f(S_2|\mathbf{y})$, $f(P_1|\mathbf{y})$, $f(P_2|\mathbf{y})$, $f(\sigma^2|\mathbf{y})$

([26]) or combinations of effects, for example

$$f\left(\frac{\mu + S_1}{\mu + S_2} \mid \mathbf{y}\right)$$

Although there are analytical solutions as in the Baby model, we only will develop the Gibbs sampling procedure. We need to obtain samples of the joint posterior distribution

$f(\mathbf{b}, \sigma^2 \mid \mathbf{y})$

We will obtain a matrix of chains, in which each row is a random sample of the joint distribution,

$$\begin{matrix} \mu & S_1 & S_2 & E_1 & E_2 & \beta & \sigma^2 \end{matrix}$$
$$\begin{bmatrix} 478 & -15 & 10 & -45 & 39 & 0.21 & 140 \\ 501 & -10 & 2 & -87 & 102 & 0.12 & 90 \\ 523 & 3 & 51 & -12 & 65 & 0.15 & 120 \\ L & L & L & L & L & L & L \end{bmatrix}$$

each column is a random sample of the marginal posterior distribution of each element of **b**, and the last chain is a random sample of the marginal posterior distribution of σ^2

[26] In a Bayesian context it is possible to estimate μ and the other effects because the colinearity can be broken by using the appropriate priors. In general, however, we will estimate combinations of effects as in the frequentist case: for example we will introduce the usual restriction that the sum of all levels of an effect is zero.

6.1.2. *Marginal posterior distributions via MCMC using Flat priors*

To work with MCMC-Gibbs sampling we need the *conditional* distributions of the unknowns $f(\mathbf{b} \mid \mathbf{y}, \sigma^2)$ and $f(\sigma^2 \mid \mathbf{y}, \mathbf{b})$. We do not know them, but we can calculate them using Bayes theorem. Using flat priors

$$\mathbf{b} \sim U[\mathbf{b}_1, \mathbf{b}_2] \longrightarrow f(\mathbf{b}) = \textbf{constant}$$
$$\sigma^2 \sim U[0, s] \longrightarrow f(\sigma^2) = \text{constant}$$

where U is the uniform function with its bounds. Now, the posterior distributions are

$$f(\mathbf{b} \mid \mathbf{y}, \sigma^2) \propto f(\mathbf{y} \mid \mathbf{b}, \sigma^2) f(\mathbf{b}) \propto f(\mathbf{y} \mid \mathbf{b}, \sigma^2)$$

$$f(\sigma^2 \mid \mathbf{y}, \mathbf{b}) \propto f(\mathbf{y} \mid \sigma^2, \mathbf{b}) f(\sigma^2) \propto f(\mathbf{y} \mid \sigma^2, \mathbf{b})$$

as we know the distribution of the data, we can obtain both conditionals

$$f(\mathbf{b} \mid \mathbf{y}, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})}{2\sigma^2} \right]$$

$$f(\sigma^2 \mid \mathbf{y}, \mu) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})}{2\sigma^2} \right]$$

Notice that the formulae are the same, but the variable is in red, thus the functions are completely different.

The conditional distribution of **b** is a multinormal distribution (Appendix 6.1).

$$f(\mathbf{b} \mid \mathbf{y}, \sigma^2) \propto \frac{1}{\left|\mathbf{X}'\mathbf{X}\sigma^2\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \cdot (\mathbf{b} - \hat{\mathbf{b}})' \left[(\mathbf{X}'\mathbf{X})^{-1} \sigma^2 \right]^{-1} (\mathbf{b} - \hat{\mathbf{b}}) \right] \sim N\left[\hat{\mathbf{b}}, (\mathbf{X}'\mathbf{X})^{-1} \sigma^2 \right]$$

where $\hat{\mathbf{b}}$ has the same form as the minimum least square estimator ([27])

$$\hat{\mathbf{b}} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$$

and the conditional distribution of σ^2 is an inverted gamma distribution with parameters

$$\beta = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

.

$$\alpha = \frac{n}{2} - 1$$

$$f\left(\sigma^2 \mid \mathbf{y}, \mathbf{b}\right) \sim IG(\alpha, \beta)$$

We have algorithms to extract random samples of both functions, thus we can start with the Gibbs sampler as we have seen in chapter 4 and in the particular case of the baby model in chapter 5.

We start with an arbitrary value for the variance (for example) and then we get a multiple sample value of **b**. We substitute this value in the conditional of the variance and we get a random value of the variance. We substitute it in the conditional distribution of the **b** and we continue the process (figure 6.1), as we have seen in chapter 5 for the baby model.

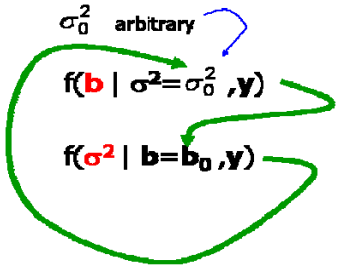

Figure 6.1. Gibbs sampling process for the b and the variance of the Normal linear model

[27] We do not say that $\hat{\mathbf{b}}$ is the minimum least square estimator but that it has "the same form" to stress that we are not estimating anything by least squares in a frequentist way, although the first proof of the least square method given by Gauss was a Bayesian one (Gauss, 1809).

6.1.3. *Marginal posterior distributions via MCMC using vague informative priors*

Vague informative priors for the variance

Let us take an inverse gamma distribution to represent our asymmetric beliefs for the variance. This function can show very different shapes by changing its parameters α and β (see figure 2.2).

$$\sigma^2 \mid \alpha,\beta \sim \mathrm{IG}(\alpha,\beta)$$

$$f\left(\sigma^2\right) = \frac{1}{\left(\sigma^2\right)^{\alpha+1}} \exp\left[-\frac{\beta}{\sigma^2}\right]$$

we do not the conditional distribution of the variance, but we can calculate it using Bayes theorem.

$$f\left(\sigma^2 \mid \mathbf{y},\mu\right) \propto f\left(\mathbf{y} \mid \sigma^2,\mu\right) f\left(\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{2\sigma^2}\right] \cdot \frac{1}{\left(\sigma^2\right)^{\alpha+1}} \exp\left[-\frac{\beta}{\sigma^2}\right] \propto$$

$$\propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}+\alpha+1}} \exp\left[-\frac{(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})-2\beta}{2\sigma^2}\right]$$

which is an inverted gamma with parameters ‘a’ and ‘b’

$$a = \frac{n}{2} + \alpha$$

$$b = \frac{1}{2}(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) - \beta$$

***Vague informative priors for* b**

Our beliefs, like in the baby model, can be represented by a multinormal normal distribution. Thus we determine the mean and variance-covariance of our beliefs, '**m**' and '**V**' respectively.

$\mathbf{b} \mid \mathbf{m}, \mathbf{V} \sim N(\mathbf{m}, \mathbf{V})$

$$f(\mathbf{b}) \propto \frac{1}{|\mathbf{V}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{b}-\mathbf{Xm})' \mathbf{V}^{-1}(\mathbf{b}-\mathbf{Xm})\right]$$

In the particular case in which all our believes have the same mean and variance 'v' and are uncorrelated, the prior is

$$f(\mathbf{b}) \propto \frac{1}{v^{\frac{n}{2}}} \exp\left[-\frac{1}{2v}(\mathbf{b}-\mathbf{Xm})'(\mathbf{b}-\mathbf{Xm})\right]$$

We do not know the conditional mean but we can know it by using Bayes theorem as before,

$$f\left(\mathbf{b} \mid \sigma^2, \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mathbf{b}, \sigma^2\right) f(\mathbf{b})$$

$$f\left(\mathbf{b} \mid \mathbf{y}, \sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})}{2\sigma^2}\right] \cdot \frac{1}{|\mathbf{V}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{b}-\mathbf{Xm})' \mathbf{V}^{-1}(\mathbf{b}-\mathbf{Xm})\right]$$

that with some algebra gymnastics can be transformed in a multinormal distribution, as we did for the baby model y appendix 5.3.

Now, we can start with the Gibbs sampling mechanism as in 5.3.1 (Figure 5.1).

6.1.4. *Least Squares as a Bayesian Estimator*

The least square estimator was developed by Legendre under intuitive bases. Later ([28]), Gauss found the first statistical justification of the method, developing least squares first as the mode of the conditional posterior distribution and later under frequentists bases. In a Bayesian context, we have seen that, under flat priors,

$$f\left(\mathbf{b} \mid \mathbf{y}, \sigma^2\right) \sim N\left[\hat{\mathbf{b}}, \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2\right]$$

As in a Normal distribution the mean, mode and median are the same, the least square estimator can be interpreted as

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = mode\, f\left(\mathbf{b} \mid \sigma^2, \mathbf{y}\right) = median\, f\left(\mathbf{b} \mid \sigma^2, \mathbf{y}\right) = mean\, f\left(\mathbf{b} \mid \sigma^2, \mathbf{y}\right)$$

Notice that this is a conditional distribution; i.e., the least square estimator needs a value for σ^2 to be calculated. In a frequentist context, we estimate σ^2, then we take this as the true value for the variance and we calculate $\hat{\mathbf{b}}$. We do not take into account the error committed when estimating the variance. However, in a Bayesian context, we calculate the *marginal* posterior distribution for **b**

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

we take into account all possible values of σ^2 and multiply by their probabilities, integrating afterwards; i.e., we take into account the error of estimating the variance.

In more philosophical grounds, as we need the *true* value of σ^2 to find the least square estimate $\hat{\mathbf{b}}$, we know that we never find a real least square estimate because we do not know the true value of σ^2. Bayesian theory does not require true values to work. The Bayesian interpretation of $\hat{\mathbf{b}}$ is the mean, mode and median of a

[28] Gauss insisted in that he found the method before Legendre, but he did not publish it. See Hald (1998) if you are interested in the controversy or you have French or German nationalistic feelings.

conditional distribution, in which σ^2 is given. We do the same, but at least now we know what we are doing.

6.2. The "mixed" model

6.2.1. *The model*

The model corresponds, in a frequentist context, to a "mixed model". As we said before, in a Bayesian context all effects are random, thus there is no distinction between fixed models, random models or mixed models. We will also consider here that the data are normally distributed, although other distributions of the data can be considered, and the procedure will be the same. Our model consists in a set of effects and covariates plus what in a frequentist model is a "random" effect, plus an error term. We can add an individual genetic effect to the model of our former example, and we have in this case

$$y_{ijkl} = \mu + S_i + P_j + b \cdot A_{ijk} + u_{ijkl} + e_{ijkl}$$

where u_{ijkl} is the individual genetic effect. In matrix form

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

where **y** contains the data, **X** is a matrix containing the covariates and the presence (1) or absence (0) of the levels of the effects. **b** is a vector of what in a frequentist context are "fixed effects" and covariates, **u** is a vector with the individual genetics effects that in a frequentist context are considered "random", and **e** is a vector of the residuals. **Z** is a matrix containing the covariates and the presence (1) or absence (0) of the levels of the individual genetics effects. If all individuals have records, **Z** = **I**, but if some individuals do not have records (for example the parental population, or traits in which data are recorded in only one sex), **Z** is not the identity matrix; for example: if individuals 1,3, 4, 5 and 6 have records but individuals 2 has no records,

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

$$\begin{bmatrix} 520 \\ 430 \\ 480 \\ 550 \\ 500 \end{bmatrix} = \mathbf{Xb} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} + \mathbf{e}$$

Thus we will consider that we have 'n' data and 'q' genetic effects to be estimated. We consider the residuals normally independently distributed all of them with mean zero and the same variance σ^2. We will take bounded flat priors for the variance

$$\mathbf{e} \mid \sigma^2 \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$$

$$\sigma^2 \sim U[0,p]$$

where U[0,p] is the uniform distribution between 0 and p, both included, and p is subjectively chosen.

The genetic effects are normally distributed with a variance-covariance matrix that depends on the additive genetic variance σ_u^2 and the relationship matrix **A**. This last matrix has the relationship coefficients between genetic effects and it is a known matrix that is calculated according to the parental relationships between individuals, based in Mendel's laws. We will also take a bounded uniform distribution for the genetic variance.

$$\mathbf{u} \mid \sigma_u^2, \mathbf{A} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$$

$$\sigma_u^2 \sim U[0,p]$$

where U[0,p] is the uniform distribution between 0 and p, both included, and p is subjectively chosen.

We will also consider a uniform distribution for the other effects

$\mathbf{b} \sim U[\mathbf{0},\mathbf{w}]$

where **0** and **w** are vectors and **w** is subjectively chosen. U[**0**,**w**] is the uniform distribution between **0** and **w**, both included.

Now the model is completely specified and we can write the distribution of the data

$\mathbf{y} \mid \mathbf{b}, \mathbf{u}, \sigma^2 \sim N(\mathbf{Xb}+ \mathbf{Zu}, \mathbf{I}\sigma^2)$

Notice that this is a simplified notation. We should have written

$\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \mathbf{Z}, \mathbf{u}, \mathbf{A}, \sigma_u^2, \sigma^2, \mathcal{H} \sim N(\mathbf{Xb}+ \mathbf{Zu}, \mathbf{I}\sigma^2)$

where $\mathcal{H}$ is the set of hypothesis we also need to define the data distribution (for example; the hypothesis that the sample has been randomly collected). We simplify the notation because **X**, **Z**, **A** and $\mathcal{H}$ are always known and because given **u** we do not need **A** and σ_u^2 any more, since the genetic effects are yet determined. We can also write, in simplified notation,

$\mathbf{y} \mid \mathbf{b}, \sigma_u^2, \sigma^2 \sim N(\mathbf{Xb}, \mathbf{Z'AZ}\sigma_u^2 + \mathbf{I}\sigma^2)$

We have now new unknowns and combination of unknowns to estimate. Our objective is to find the marginal posterior distributions of all unknowns

$f(\mu|\mathbf{y}), f(S_1|\mathbf{y}), f(S_2|\mathbf{y}), f(P_1|\mathbf{y}), f(P_2|\mathbf{y}), f(u_1|\mathbf{y})\ f(u_2|\mathbf{y}),\ldots, f(\sigma_u^2|\mathbf{y}), f(\sigma^2|\mathbf{y})$

or combinations of them, for example we can be interested in estimating the marginal posterior distribution of the response to selection, which can be defined as the distribution of the average of the genetic values of the last generation.

6.2.2. *Marginal posterior distributions via MCMC*

To work with MCMC-Gibbs sampling we need the *conditional* distributions of the unknowns.

$$f\left(\mathbf{b} \mid \mathbf{u}, \sigma_u^2, \sigma^2, \mathbf{y}\right)$$
$$f\left(\mathbf{u} \mid \mathbf{b}, \sigma_u^2, \sigma^2, \mathbf{y}\right)$$
$$f\left(\sigma_u^2 \mid \mathbf{u}, \mathbf{b}, \sigma^2, \mathbf{y}\right)$$
$$f\left(\sigma^2 \mid \mathbf{u}, \mathbf{b}, \sigma_u^2, \mathbf{y}\right)$$

We will write first the joint distribution

$$f\left(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma^2 \mid \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mathbf{u}, \mathbf{b}, \sigma_u^2, \sigma^2\right) \cdot f\left(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma^2\right)$$

Some unknowns are not independent “a prior”; for example **u** and σ_u^2. In this case we know by the theory of probability that P(A,B)=P(A|B)·P(B), thus

$$f\left(\mathbf{u}, \sigma_u^2\right) = f\left(\mathbf{u} \mid \sigma_u^2\right) \cdot f\left(\sigma_u^2\right)$$

but if some unknowns are independent “a priori”, for example **u** and **b** we know that

$$f(\mathbf{b}, \mathbf{u}) = \cdot f(\mathbf{b}) \cdot f(\mathbf{u})$$

then, assuming independence “a priori” between some unknowns, we have

$$f\left(\mathbf{b}, \mathbf{u}, \sigma_u^2, \sigma^2 \mid \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mathbf{u}, \mathbf{b}, \sigma_u^2, \sigma^2\right) \cdot f\left(\mathbf{b}\right) \cdot f\left(\mathbf{u} \mid \sigma_u^2\right) \cdot f\left(\sigma_u^2\right) \cdot f\left(\sigma^2\right)$$

This supposition sometimes holds and sometimes not. For example, it is well known that the best dairy cattle farms have the best environment and also buy the best semen, thus their genetic level is higher and this generates a positive covariance

between **b** and **u**. In the case of pigs in a model with only season effects, for example, it is not expected that the genetically best pigs come in summer or in winter, thus it seems that we can assume independence between **b** and **u**. It is less clear the independence between the genetic and environmental variances. Usually the literature in genetics offers heritabilities, which is a ratio between the genetic variance and the sum of the genetic and environmental variances, thus it Is not easy to have a subjective opinion of the genetic variance independent of our opinion about the environmental variance. However it is even more complicated to assess our opinion about the covariances, thus Bayesian geneticists prefer, with some exception (Blasco et al., 1998) to consider prior independence, hoping that the data will dominate the final result and this assumption will not have any consequence in the results.

Considering the prior distributions we have established in 6.2.1., we have

$$f\left(\mathbf{b},\mathbf{u},\sigma_u^2,\sigma^2 \mid \mathbf{y}\right) \propto f\left(\mathbf{y} \mid \mathbf{b},\mathbf{u},\sigma^2\right) \cdot f\left(\mathbf{u} \mid \mathbf{A}\sigma_u^2\right) \propto$$

$$\propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)'\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)\right] \frac{1}{\left(\sigma_u^2\right)^{\frac{q}{2}}} \exp\left[-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\right]$$

Now, all the conditionals are based in this function, but changing the red and black parts.

$$f\left(\sigma^2 \mid \mathbf{b},\mathbf{u},\sigma_u^2,\mathbf{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)'\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)\right] \frac{1}{\left(\sigma_u^2\right)^{\frac{q}{2}}} \exp\left[-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\right] \propto$$

$$\propto \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)'\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)}{2\sigma^2}\right]$$

This is an inverted gamma, as we have seen in 6.2.1 and chapter 3, with parameters

$$\beta = \frac{1}{2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})$$

$$\alpha = \frac{n}{2} - 1$$

$$f(\sigma_u^2 \mid \mathbf{b}, \mathbf{u}, \sigma^2, \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})\right] \frac{1}{(\sigma_u^2)^{\frac{q}{2}}} \exp\left[-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\right] \propto$$

$$\propto \frac{1}{(\sigma_u^2)^{\frac{q}{2}}} \exp\left[-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\right]$$

This is also an inverted gamma distribution, with parameters

$$\beta = \frac{1}{2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}$$

$$\alpha = \frac{q}{2} - 1$$

$$f(\mathbf{b} \mid, \mathbf{u}, \sigma_u^2, \sigma^2, \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})\right] \exp\left[-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\right] \propto$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})\right] \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}^* - \mathbf{Xb})'(\mathbf{y}^* - \mathbf{Xb})\right]$$

where $\mathbf{y}^* = \mathbf{y} - \mathbf{Zu}$. We have seen in 6.2.1 that this can be transformed in a multinormal distribution.

$$f(\mathbf{u} \mid \mathbf{b}, \sigma_u^2, \sigma^2, \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})'(\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})\right] \frac{1}{(\sigma^2)^{\frac{q}{2}}} \exp\left[-\frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\right]$$

that can also be converted after some algebra in a multinormal distribution.

After having the conditionals identified as functions from which we have algorithms for taking random samples from them, we can start with the Gibbs sampling procedure (Figure 6.2) as we did in the case 6.2.1.

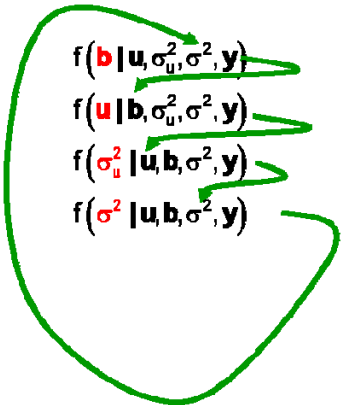

Figure 6.2. Gibbs sampling process for the components of the model y=Xb+Zu+e and the variance components

6.2.3. *BLUP as a Bayesian estimator*

BLUP in a frequentist context

BLUP are the initials of “Best Linear Unbiased Predictor”, a name that is somewhat misleading because BLUP is only the best estimator (minimum risk) within the class of the unbiased estimators ([29]). Moreover, BLUP is not unbiased in the sense in which this word is used for fixed effects, as we saw in 1.5.2. BLUP is just a linear predictor with some good properties. Henderson (1976) discovered that BLUP and the corresponding estimates for the fixed effects (BLUE, Best linear Unbiased Estimators) could be obtained by solving the equations

[29] In a frequentist context, the word ‘estimator’ is used only for fixed effects, that remain constant in each conceptual repetition of the experiment; for random effects the most common word is ‘predictor’, since the effect changes in each repetition of the experiment. In a Bayesian context there are neither fixed effects nor conceptual repetitions of the experiment, thus we will only use the word estimator.

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\frac{\sigma^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

These are called the “mixed model equations” and they are particularly popular among animal breeders.

To solve these equations we need the true value of σ_u^2 and σ^2, but we do not know them, thus we will never get real BLUP, but pseudo-BLUP in which the true values of the variance components will be substituted by our estimations. If we have good estimations, we hope our pseudo-BLUP to be close to the real BLUP, but in a frequentist context we do not have any method to include the error of estimation of the variance components in the pseudo-BLUP we obtained, thus we will underestimate our standard errors.

How Henderson derived BLUP

Henderson (1973) explained how he derived BLUP as a result of a mistake. He was learning with Jay Lush how to predict genetic values, and was also learning with Mood (one of the authors of the well known textbook of statistics Mood & Graybill, 1963) how to estimate fixed effects. Due to an error in the mood’s book, he thought he was using the maximum likelihood method for the mixed model what was not actually true. He presented the method in a local dairy cattle meeting, but when he realized he was wrong, he did not publish the method (only a summary of the paper presented in the dairy cattle meeting appeared in the Journal of Animal Science in 1949). Later, with the help of Searle, he discovered that the method had good statistical properties, and it became the most popular method between animal breeders when Henderson (1976) discovered how to calculate $\mathbf{A}^{-1}$ directly and at a low computational cost.

When working by maximum likelihood, we try to find the value of the parameter that, if true, will lead to having our data the maximum probability to be sampled (see chapter 1). If we want to find the maximum likelihood of **u**, we will try to find the value

of **u** that, if true, will lead to a maximum probability of **y**. This cannot be done with random effects because if we fix the value of **u**, it is not random any more. Henderson ignored this, and treated **u** as a random effect. As we have said, we try to find the value of **u** that, *if true*, will lead to a maximum probability of **y**, thus it is natural to multiply by the probability of this value to be true. Henderson maximized

$$f(\mathbf{y} \mid \mathbf{u}) \cdot f(\mathbf{u})$$

which, incidentally, is the joint distribution of the data and the random values f(**y**, **u**). We know the distribution of the data; they are normally distributed with a mean **Xb** and a variance $\mathbf{I}\sigma^2$, but when the random values **u** are 'given', the mean is **Xb** + **Zu**. Thus, Henderson maximized the quantity φ

$$\varphi = f(\mathbf{y} \mid \mathbf{u}) \cdot f(\mathbf{u}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu})'\left(\mathbf{I}\sigma^2\right)^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu})\right]\exp\left[-\frac{1}{2}\mathbf{u}'\left(\mathbf{A}\sigma_u^2\right)^{-1}\mathbf{u}\right]$$

to facilitate this, we will take logarithms, because the maximum of a quantity is at the same point as the maximum of its logarithm.

$$\log \varphi \propto (\mathbf{y}-\mathbf{Xb}-\mathbf{Zu})'\left(\mathbf{I}\sigma^2\right)^{-1}(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}) + \mathbf{u}'\left(\mathbf{A}\sigma_u^2\right)^{-1}\mathbf{u}$$

$$\frac{\partial}{\partial \mathbf{b}} \ln\varphi \propto -2\left(\sigma^2\right)^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{Xb})$$

$$\frac{\partial}{\partial \mathbf{u}} \ln\varphi \propto -2\left(\sigma^2\right)^{-1}\mathbf{Z}'(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}) + 2\left(\mathbf{A}\sigma_u^2\right)^{-1}\mathbf{u}$$

equating to zero to find the maxima, we obtain

$$\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} + \mathbf{X}'\mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}'\mathbf{y}$$
$$\mathbf{Z}'\mathbf{X}\hat{\mathbf{b}} + \mathbf{Z}'\mathbf{Z}\hat{\mathbf{u}} + \sigma^2\mathbf{A}\left(\sigma_u^2\right)^{-1}\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{y}$$

that in matrix form leads to the mixed model equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma^2}{\sigma_u^2}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

This procedure is confusing, since it is not maximum likelihood and we are also estimating **b** without applying the same logic to **b** than to **u**.

What Henderson really did

In a Bayesian context it is clearer what Henderson really did. Henderson maximized

$$\varphi = f\left(\mathbf{u},\mathbf{b} \mid \mathbf{y},\sigma_u^2,\sigma^2\right) \propto f\left(\mathbf{y} \mid \mathbf{u},\mathbf{b},\sigma_u^2,\sigma^2\right) f\left(\mathbf{u},\mathbf{b} \mid \sigma_u^2,\sigma^2\right)$$

Assuming prior independence between **u** and **b**, and assuming a flat prior for **b**

$$f\left(\mathbf{u},\mathbf{b} \mid \sigma_u^2,\sigma^2\right) = f\left(\mathbf{u} \mid \sigma_u^2,\sigma^2\right) \cdot f\left(\mathbf{b} \mid \sigma_u^2,\sigma^2\right) \propto f\left(\mathbf{u} \mid \sigma_u^2,\sigma^2\right)$$

Then we arrive to the formula that Henderson maximized:

$$\varphi = f\left(\mathbf{u},\mathbf{b} \mid \mathbf{y},\sigma_u^2,\sigma^2\right) \propto f\left(\mathbf{y} \mid \mathbf{u},\mathbf{b},\sigma_u^2,\sigma_e^2\right) f\left(\mathbf{u} \mid \sigma_u^2,\sigma^2\right) \propto$$

$$\propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})'\left(\mathbf{I}\sigma^2\right)^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})\right] \exp\left[-\frac{1}{2}\mathbf{u}'\left(\mathbf{A}\sigma_u^2\right)^{-1}\mathbf{u}\right]$$

Now it is clear what Henderson did. He found the mode of the joint posterior distribution of **b** and **u**, conditioned not only to the data, but to the values of the variance components. From this point of view, BLUP does not need true values any more, but given values, and the joint mode is the most probable posterior value of the unknowns. Here **b** and **u**, are treated in the same way, **b** has a flat prior and **u** has a Normal prior distribution.

BLUP as a Bayesian estimator

In a Bayesian context there are no differences between fixed and random effects, thus we do not need mixed models and we can work with the same model as in 6.1.2. We will use vague priors for **b**. We assume independence between **b** and **u**.

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

$\mathbf{W} = [\,\mathbf{X}\ \mathbf{Z}\,] \qquad \mathbf{t}' = [\,\mathbf{b}'\ \mathbf{u}'\,]$

$\mathbf{b} \mid \mathbf{m}_b, \mathbf{S} \sim N(\mathbf{m}_b, \mathbf{S})$

$\mathbf{u} \mid \sigma_u^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$

$$\mathbf{t} \mid \mathbf{m}, \mathbf{S}, \sigma_u^2 \sim N\left(\begin{bmatrix}\mathbf{m}_b \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\sigma_u^2\end{bmatrix}\right) \equiv N(\mathbf{m}, \mathbf{V})$$

where $\mathbf{m} = \begin{bmatrix}\mathbf{m}_b \\ \mathbf{0}\end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix}\mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\sigma_u^2\end{bmatrix}$

$\mathbf{y} \mid \mathbf{t}, \sigma^2 \sim N(\mathbf{Wt}, \mathbf{I}\sigma^2)$

Now we will find the mode of the posterior distribution of **t** given the data, *but also conditioned to the variance components*. Applying Bayes theorem, we have

$$f(\mathbf{t} \mid \mathbf{y}, \mathbf{V}, \sigma^2) \propto f(\mathbf{y} \mid \mathbf{t}, \mathbf{V}, \sigma^2) \cdot f(\mathbf{t} \mid \mathbf{V}) \propto$$

$$\propto \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{Wt})'(\mathbf{I}\sigma^2)^{-1}(\mathbf{y} - \mathbf{Wt})\right] \exp\left[-\frac{1}{2}(\mathbf{t} - \mathbf{m})'\mathbf{V}^{-1}(\mathbf{t} - \mathbf{m})\right]$$

$$\ln f(\mathbf{t} \mid \mathbf{y}, \mathbf{m}, \mathbf{V}, \sigma^2) \propto \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{Wt})'(\mathbf{y} - \mathbf{Wt}) + (\mathbf{t} - \mathbf{m})'\mathbf{V}^{-1}(\mathbf{t} - \mathbf{m})$$

$$\frac{\partial}{\partial \mathbf{t}} \ln f\left(\mathbf{t} \mid \mathbf{y}, \mathbf{V}, \sigma^2\right) \propto -\frac{1}{\sigma^2}\mathbf{W}'\left(\mathbf{y}-\mathbf{W}\mathbf{t}\right)+\mathbf{V}^{-1}\left(\mathbf{t}-\mathbf{m}\right)$$

equating to zero this leads to

$$\mathbf{W}'\mathbf{W}'\hat{\mathbf{t}}+\sigma^2\mathbf{V}^{-1}\hat{\mathbf{t}}=\mathbf{W}'\mathbf{y}-\sigma^2\mathbf{V}^{-1}\mathbf{m}$$

$$\left[\mathbf{X}\ \mathbf{Z}\right]'\begin{bmatrix}\mathbf{X}\\ \mathbf{Z}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}+\sigma^2\begin{bmatrix}\mathbf{S} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\sigma_u^2\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}=\left[\mathbf{X}\ \mathbf{Z}\right]'\mathbf{y}-\sigma^2\begin{bmatrix}\mathbf{S} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\sigma_u^2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{m}_b\\ \mathbf{0}\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{X}'\mathbf{X}+\sigma^2\mathbf{S}^{-1} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\dfrac{\sigma^2}{\sigma_u^2}\mathbf{A}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{X}'\mathbf{y}-\sigma^2\mathbf{S}^{-1}\mathbf{m}_b\\ \mathbf{Z}'\mathbf{y}\end{bmatrix}$$

These equations are very similar to the mixed model equations. In fact, if $\mathbf{S}^{-1}=\mathbf{0}$ they are identical to the mixed model equations. This condition only holds if the prior variance of **b** is infinite; i.e., if we use unbounded flat priors for **b**. Therefore, in a Bayesian context, the difference between what in a frequentist context are called "fixed" or "random" effects is only the type of prior they have. A "fixed" effect in a frequentist context is just a random effect with an unbounded flat prior in a Bayesian context. The mystery of the difference between fixed and random effects has now been solved.

In a Bayesian context, BLUP is the mode of the joint posterior distribution of **b** and **u**, conditioned not only to the data, but to the values of the variance components, when we use an unbounded flat prior for **b**. Notice that in a Bayesian context BLUP is not biased or unbiased, since there are not repetitions of the experiment. We can be interested in what will happen in repetitions of the experiment, but our inferences are based only in our sample and the priors, not in the information of the sampling space.

6.2.4. *REML as a Bayesian estimator*

We have seen in 5.2.3 that the mode of the marginal posterior distribution of the variance gives an expression that is the same we obtain in a frequentist context for the REML estimate. The same happens in the linear model, the REML estimators of σ_u^2 and σ^2 are coincident with the mode of the joint marginal posterior density

$$f\left(\sigma_u^2,\sigma^2|\mathbf{y}\right)=\iint f\left(\mathbf{b},\mathbf{u},\sigma_u^2,\sigma^2|\mathbf{y}\right)d\mathbf{b}\,d\mathbf{u}=$$

$$=\iint f\left(\sigma_u^2,\sigma^2|\mathbf{b},\mathbf{u},\mathbf{y}\right)\cdot f\left(\mathbf{b}\right)\cdot f\left(\mathbf{u}\right)d\mathbf{b}\,d\mathbf{u}\propto\iint f(\sigma_u^2,\sigma^2|\mathbf{b},\mathbf{u},\mathbf{y})\cdot f\left(\mathbf{u}\right)d\mathbf{b}\,d\mathbf{u}$$

when prior values are assumed to be flat for **b** and normal for **u**, as in the case of BLUP. Notice that this is not the best Bayesian solution; we usually will prefer the mean or the median of each marginal distribution for each variance component instead of the mode of the joint distribution of both variance components.

6.3. The multivariate model

6.3.1. *The model*

When several correlated traits are analysed together, we use a multivariate model. Sometimes the models for each trait may be different. For example, when analyzing litter size and growth rate, a dam may have several litters and consequently several records for litter size, but only one data for growth rate. Moreover, many animals will have one data for growth rate but no data for litter size, because they were males or they were not selected to be reproductive stock. We will put an example in which one trait has several records and the other trait only one record: for example, in dairy cattle we have several records for milk production but only one record for type traits. This means that we can add an environmental effect that is common for all lactation records, but we do not have this effect in the type traits because we only have one record for animal. The multivariate model is

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{e}_1$$

$$\mathbf{y}_2 = \mathbf{X}_2\mathbf{b}_2 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{W}_2\mathbf{p}_2 + \mathbf{e}_2$$

where $\mathbf{b}_1$ and $\mathbf{b}_2$ are environmental effects (season, herd, etc.), $\mathbf{u}_1$ and $\mathbf{u}_2$ are genetic effects, $\mathbf{p}_2$ is the common environmental effect to all records of trait 2, and $\mathbf{e}_1$ and $\mathbf{e}_2$ are the residuals. We assume

$$\mathbf{y}_1 \mid \mathbf{b}_1, \sigma_1^2 \sim N(\mathbf{X}\mathbf{b}_1, \mathbf{I}\sigma_1^2)$$

$$\mathbf{y}_2 \mid \mathbf{b}_2, \mathbf{u}_2, \mathbf{p}_2, \sigma_2^2 \sim N(\mathbf{X}\mathbf{b}_2 + \mathbf{Z}\mathbf{u}_2 + \mathbf{W}\mathbf{p}_2, \mathbf{I}\sigma_2^2)$$

$$\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix} \sim \text{Uniform bounded}$$

$$\mathbf{p}_2 \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2)$$

$$\sigma_p^2 \sim \text{Uniform, } \geq 0, \text{ bounded}$$

$$\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}_{\text{SORTED BY INDIVIDUAL}} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{A})$$

$$\mathbf{G} = \begin{bmatrix}\sigma_{u_1}^2 & \sigma_{u_1u_2}\\ \sigma_{u_1u_2} & \sigma_{u_2}^2\end{bmatrix} \sim \text{Uniform bounded}$$

$$\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}_{\text{SORTED BY INDIVIDUAL}} \sim N(\mathbf{0}, \mathbf{R} \otimes \mathbf{I})$$

$$\mathbf{R} = \begin{bmatrix}\sigma_{e_1}^2 & \sigma_{e_1e_2}\\ \sigma_{e_1e_2} & \sigma_{e_2}^2\end{bmatrix} \sim \text{Uniform bounded}$$

when vague priors are used, a multinormal distribution is often used for the priors of $\mathbf{b}_1$ and $\mathbf{b}_2$, and Inverted Whishart distributions (the equivalent to the inverted Gamma for the multivariate case) are used for **G** and **R**.

It is also assumed prior independence between some unknowns. As most priors are constant,

$$f\left(\mathbf{u}_1,\mathbf{u}_2,\mathbf{b}_1,\mathbf{b}_2,\mathbf{p},\sigma_p^2,\mathbf{G},\mathbf{R}\right) \propto f\left(\mathbf{u}_1,\mathbf{u}_2 \mid \mathbf{G}\right)\cdot f\left(\mathbf{p} \mid \sigma_p^2\right)$$

This model has the great problem in order to be managed that the design matrixes are different and we have an effect more in trait 2. If all design matrixes would be the same, we could write the model with both traits as

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

where **y, b, p** and **u** are the data and the effects of both traits. The only new distributions we need is

$$\begin{bmatrix}\mathbf{p}_1\\ \mathbf{p}_2\end{bmatrix}_{\text{SORTED BY INDIVIDUAL}} \sim N\left(\mathbf{0},\mathbf{P}\otimes\mathbf{I}\right)$$

$$\mathbf{P} = \begin{bmatrix}\sigma_{p_1}^2 & \sigma_{p_1p_2}\\ \sigma_{p_1p_2} & \sigma_{p_2}^2\end{bmatrix} \sim \text{Uniform bounded}$$

([30]) In next paragraph we will see how we can write the multivariate model as if all design matrixes were the same and all traits would have the same effects. This technique is known as “data augmentation”.

[30] We have to express **u**, **e** and **p** “sorted by individual” in order to use this synthetic notation, as the reader can easily find by putting an example.

6.3.2. *Data Augmentation*

Data augmentation is a procedure to augment the data base filling the gaps until all traits have the same design matrixes. Thus, if some traits have several season effects and one of the traits has only one season effect, new data come with several season effects for this trait until it has the same **X** matrix as the others. If one trait has only one record, new records are added until we have also a common environmental effect for this trait. The conditions that these new records added must follow are:

1. The new records are added to fill the gaps until all traits have the same design matrixes

2. The new records must not be used for inferences, since they are not real records.

The second condition is important. Inferences are only based on the sample **y**, the augmented data must not be used for inferences, and they will not add any information or modify the result of the analyses.

Let us call $\mathbf{z'} = [\mathbf{z'}_1, \mathbf{z'}_2]$ the vector of augmented data for trait 1 and 2, and let us call the new data vector with the recorded and the augmented data

$$\mathbf{y}_1^* = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{z}_1 \end{bmatrix} \qquad \mathbf{y}_2^* = \begin{bmatrix} \mathbf{y}_2 \\ \mathbf{z}_2 \end{bmatrix}$$

the new multivariate model is now

$$\mathbf{y}_1^* = \mathbf{X}\mathbf{b}_1 + \mathbf{Z}\mathbf{u}_1 + \mathbf{W}\mathbf{p}_1 + \mathbf{e}_1$$
$$\mathbf{y}_2^* = \mathbf{X}\mathbf{b}_2 + \mathbf{Z}\mathbf{u}_2 + \mathbf{W}\mathbf{p}_2 + \mathbf{e}_2$$

which can be written as

$$\mathbf{y}^* = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{p} + \mathbf{e}$$

and solved as in 6.2.2. Now we should find a way to generate the augmented data to avoid that they will take part in the inference.

Let us call $\boldsymbol{\theta}$ all unknowns

$$\boldsymbol{\theta} = \mathbf{u},\mathbf{b},\mathbf{p},\sigma_p^2,\mathbf{G},\mathbf{R}$$

We should generate the augmented data **z** that are also unknown and should be treated as unknowns. As with the other unknowns $\boldsymbol{\theta}$, we should estimate the posterior distribution conditioned to the data $f(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y})$. We do not know this distribution, but we know the distribution of the data and we can apply Bayes theorem. The joint prior, according to the laws of probability, is

$$f(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{y}) = f(\boldsymbol{\theta} \mid \mathbf{y}) \cdot f(\mathbf{z})$$

Applying Bayes theorem, we have

$$f(\boldsymbol{\theta},\mathbf{z} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta},\mathbf{z}) \cdot f(\boldsymbol{\theta},\mathbf{z}) = f(\mathbf{y} \mid \boldsymbol{\theta},\mathbf{z}) \cdot f(\mathbf{z} \mid \boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})$$

but, according to the laws of probability, we have

$$f(\mathbf{y},\mathbf{z} \mid \boldsymbol{\theta}) = f(\mathbf{y} \mid \boldsymbol{\theta},\mathbf{z}) \cdot f(\mathbf{z} \mid \boldsymbol{\theta})$$

thus, substituting, we have

$$f(\boldsymbol{\theta},\mathbf{z} \mid \mathbf{y}) \propto f(\mathbf{y},\mathbf{z} \mid \boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}) = f(\mathbf{y}^* \mid \boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})$$

and now we can start with the Gibbs sampling because we know the distribution of $\mathbf{y}^*$ and the conditionals (Figure 6.3). The only new conditional is the conditional of the augmented data, but the augmented data are distributed as the data, thus

$\mathbf{z} \mid \mathbf{b}, \mathbf{u}, \mathbf{p}, \mathbf{R} \sim N(\mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp}, \mathbf{I}\sigma_2^2)$

in each case, the data are sampled from the corresponding distribution

$\mathbf{z}_1 \mid \mathbf{b}_1, \mathbf{u}_1, \mathbf{p}_1, \sigma_1^2 \sim N(\mathbf{Xb}_1 + \mathbf{Zu}_1 + \mathbf{Wp}_1, \mathbf{I}\sigma_1^2)$

$\mathbf{z}_2 \mid \mathbf{b}_2, \mathbf{u}_2, \mathbf{p}_2, \sigma_2^2 \sim N(\mathbf{Xb}_2 + \mathbf{Zu}_2 + \mathbf{Wp}_2, \mathbf{I}\sigma_2^2)$

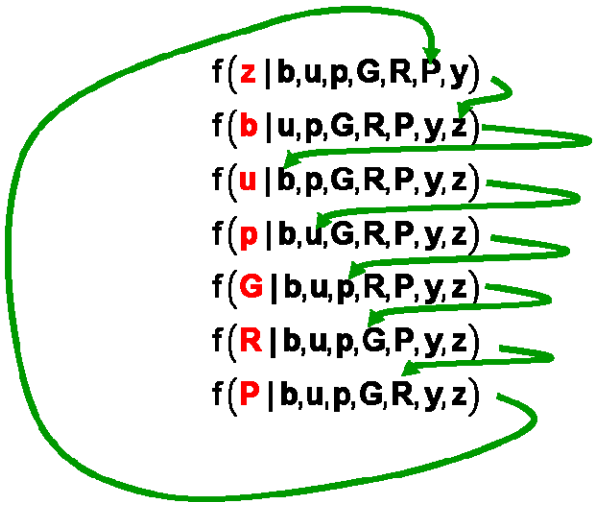

Figure 6.3. Gibbs sampling process for the components of the multivariate model with augmented data

At the end of the process we will have chains for all unknowns and for the augmented data **z**. We will ignore the augmented data and we will use the chains of the unknowns for inferences. This is a legitimate procedure, because we have sampled the distribution f(**θ**, **z** | **y**), which depends on the data **y** but not on the augmented data.

In practice, it is more convenient to do a similar process augmenting residuals instead of data. See Sorensen and Gianola (2004) for details.

Appendix 6.1

First, consider the following product:

$$\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{b}-\hat{\mathbf{b}}\right)=\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}-2\mathbf{b}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}+\hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}$$

where

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

and substituting, we have

$$\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{b}-\hat{\mathbf{b}}\right)=\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}-2\mathbf{b}'\mathbf{X}'\mathbf{y}+\text{constant}$$

Now consider the numerator in the exp of $f\left(\mathbf{b}\,|\,\mathbf{y},\sigma^2\right)$

$$(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})=\mathbf{y}'\mathbf{y}-2\mathbf{b}'\mathbf{X}'\mathbf{y}+\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{y}'\mathbf{y}+\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{b}-\hat{\mathbf{b}}\right)+\text{constant}$$
$$=\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{b}-\hat{\mathbf{b}}\right)+\text{constant}$$

Substituting,

$$f\left(\mathbf{b}\,|\,\mathbf{y},\sigma^2\right)\propto\frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}}\exp\left[-\frac{(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})}{2\sigma^2}\right]\propto\frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}}\exp\left[-\frac{\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\mathbf{X}'\mathbf{X}\left(\mathbf{b}-\hat{\mathbf{b}}\right)}{2\sigma^2}\right]\propto$$

$$\propto\frac{1}{\left|\mathbf{X}'\mathbf{X}\sigma^2\right|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\cdot\left(\mathbf{b}-\hat{\mathbf{b}}\right)'\left[(\mathbf{X}'\mathbf{X})^{-1}\sigma^2\right]^{-1}\left(\mathbf{b}-\hat{\mathbf{b}}\right)\right]$$

where we have added $|\mathbf{X}'\mathbf{X}|^{½}$ from the proportional constant, and organized the exp to put it as a Normal distribution.

CHAPTER 7

PRIOR INFORMATION

> "We were certainly aware that inferences must make use of prior information, but after some considerable thought and discussion round these matters we came to the conclusion, rightly or wrongly, that it was so rarely possible to give sure numerical values to these entities, that our line of approach must proceed otherwise"
>
> **Egon Pearson**, 1962.

7.1. Exact prior information
- 7.1.1. Prior information
- 7.1.2. Posterior probabilities with exact prior information
- 7.1.3. Influence of prior information in posterior probabilities

7.2. Vague prior information
- 7.2.1. A vague definition of vague prior information
- 7.2.2. Examples of the use of vague prior information

7.3. No prior information
- 7.3.1. Flat priors
- 7.3.2. Jeffrey's priors
- 7.3.3. Bernardo's "Reference" priors

7.4. Improper priors

7.5. The Achilles heel of Bayesian inference

7.1. Exact prior information

7.1.1. *Prior information*

When there is exact prior information there is no discussion about Bayesian methods and it can be integrated using the rules of probability. The following example is based on an example prepared for Fisher, who was a notorious anti-Bayesian, but he never objected the use of prior probability when clearly established.

There is a type of laboratoy mouse whose skin colour is controlled by a single gene with two alleles 'A' and 'a' so that when the mouse has two copies of the recessive allele (aa) its skin is brown, and it is black in the other cases (AA and aa). We cross two heterozygous([31]) and we have a descent that is black coloured. We want to know whether this black mouse is homozygous (AA) or heterozygous (Aa or aA). In order to know this, we cross the mouse with a brown mouse (aa) and examine the offspring (Figure 4.1).

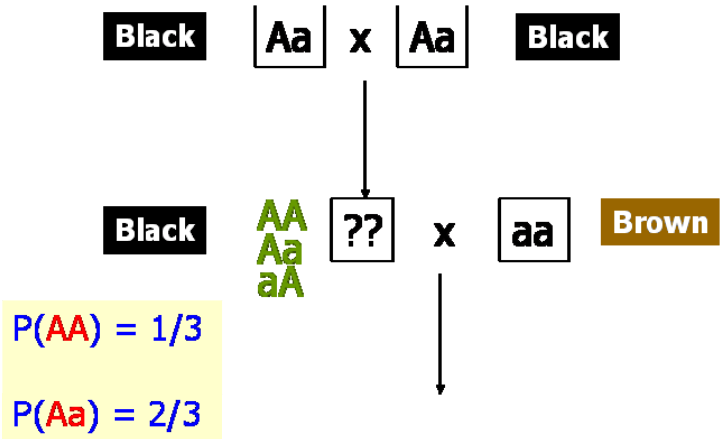

Figure 4.1. Experiment to determine wheter a parent is homocygous (AA) ot heterozygous (Aa or aA)

If we get a brown mouse in the offspring, we will know that the mouse is heterozygous, but if we only get black offspring we still will doubt about wheter it is

[31] We know they are heterozygous because they are offspring of black and brown mice.

homo- or heterozygous. If we get many black offspring, it will be unlikely that the mouse is heterozygous, but *before making the experiment* we have some probabilities of obtaining black or brown offspring. We know that the mouse to be tested cannot be 'aa' because otherwise it would be brown, thus it received both alleles 'A' from its mother and father, or an allele 'A' from the father and an allele 'a' from the mother to become 'Aa', or the opposite, to become 'aA'. We have three possibilites, thus the probability of being 'AA' is 1/3 and the probabilities of being heterocygous ('Aa' or 'aA', both are genetically identical ([32])) is 2/3. This is what we expect before having any data from the experiment. Notice that these expectations are not merely 'beliefs', but quantified probabilities. Also notice that they come from our knowledge of the Mendel laws and from the knowledge that our mouse is the son of two heterozygous.

7.1.2. *Posterior probabilities with exact prior information*

Now, the experiment is made and we obtain three offspring, all black (figure 4.2). They received for sure an allele 'a' from the mother and an allele 'A' from our mouse, but our mouse still can be homozygous (AA) or heterozygous (Aa). Which is the probability of being each type?

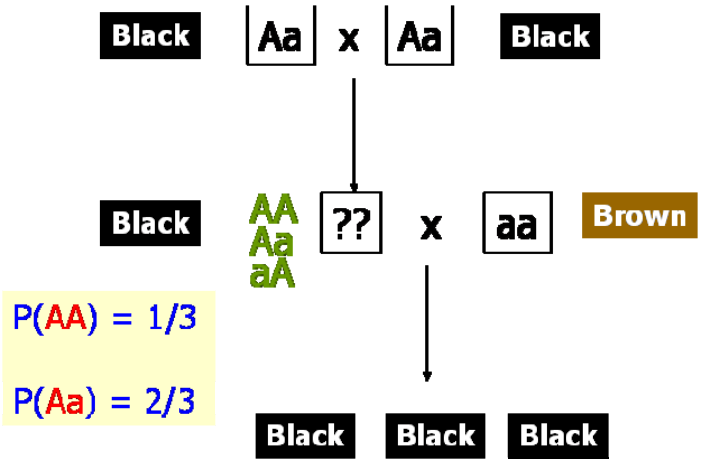

Figure 4.2. Experiment to determine wheter a parent is homocygous (AA) ot heterozygous (Aa or aA)

[32] We will no make any difference from Aa and aA in the rest of the chapter, thus 'Aa' will mean both 'Aa' and 'aA' from now.

To know this we will apply Bayes Theorem. The probability of being homozygous (AA) given that we have obtained three offspring black is

$$P(AA \mid \mathbf{y} = 3 \text{ black}) = \frac{P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA)}{P(\mathbf{y} = 3 \text{ black})}$$

We know that *if it is true that our mouse is AA*, the probability of obtaining a black offspring is 1, since the offspring will always have an allele 'A'. Thus,

$$P(\mathbf{y} = 3 \text{ black} \mid AA) = 1$$

We also know that the prior probability of being AA is 1/3, thus

$$P(AA) = 0.33$$

Finally, the probability of the sample is the sum of the probabilities of two excluding events: having a parent homozygous (AA) or having a parent heterozygous (Aa) (see footnote 2).

$$P(\mathbf{y} = 3 \text{ black}) = P(\mathbf{y} = 3 \text{ black } \& \ AA) + P(\mathbf{y} = 3 \text{ black } \& \ Aa) =$$

$$= P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA) + P(\mathbf{y} = 3 \text{ black} \mid Aa) \cdot P(Aa)$$

to calculate it we need the prior probability of being heterozygous, that we know it is

$$P(Aa) = \frac{2}{3}$$

and the probability of obtaining our sample *if it is true that our mouse is Aa.* If our mouse would be Aa, the only way of obtaining a black offspring is that this offspring get his allele A from him, thus the probability of obtaining one black offspring will be ½. The probability of obtraining three black offspring will be ½ x ½ x ½ , thus

$$P(\mathbf{y} = 3 \text{ black} \mid Aa) = \left(\frac{1}{2}\right)^3$$

Now we can calculate the probability of our sample:

$$P(\mathbf{y} = 3 \text{ black}) = P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA) + P(\mathbf{y} = 3 \text{ black} \mid Aa) \cdot P(Aa) =$$

$$= \quad 1 \quad \cdot \quad \frac{1}{3} \quad + \quad \left(\frac{1}{2}\right)^3 \quad \cdot \quad \frac{2}{3} \quad = 0.42$$

Then, applying Bayes theorem

$$P(AA \mid \mathbf{y} = 3 \text{ black}) = \frac{P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA)}{P(\mathbf{y} = 3 \text{ black})} = \frac{1 \times 0.33}{0.42} = 0.80$$

The probability of being heterozygous can be calculated again using Bayes theorem, or simplily as

$$P(Aa \mid \mathbf{y} = 3 \text{ black}) = 1 - P(AA \mid \mathbf{y} = 3 \text{ black}) = 1 - 0.80 = 0.20$$

Thus, we had a prior probability, before obtaining any data, and a probability after obtaining three black offspring

prior $P(AA) = 0.33$	posterior $P(AA \mid \mathbf{y}) = 0.80$
prior $P(Aa) = 0.67$	posterior $P(Aa \mid \mathbf{y}) = 0.20$

before the experiment was performed it was more probable that our mouse was heterozygous (Aa), but after the experiment it is more probable that it is homozygous (AA).

Notice that the sum of both probabilities is 1

$P(AA \mid \mathbf{y}) + P(Aa \mid \mathbf{y}) = 1.00$

thus the posterior probabilitites give a relative measure of uncertainty (80% and 20% respectively). However, the sum of the likelihoods is not 1 because they come from different events

P(**y** |AA) + P(**y** |Aa) = 1.125

thus the likelihoods do not provide a *measure* of uncertainty.

7.1.3. *Influence of prior information in posterior probabilitites*

If instead of using exact prior information we had used flat priors, repeating the calculus, we will obtain

$$\text{prior } P(AA) = 0.50 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.89$$
$$\text{prior } P(Aa) = 0.50 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.11$$

we can see that flat prior information had an influence in the final result. When having exact prior information it is better to use it.

If we have exact prior information and we have a large amount of information, for example P(AA) = 0.002, computing the probabilities again we obtain

$$\text{prior } P(AA) = 0.002 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.02$$
$$\text{prior } P(Aa) = 0.998 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.98$$

thus despite of having evidence from the data in favour of AA, we decide that the mouse is Aa because prior information dominates and the posterior distribution favours Aa. This has been a frequent criticism to Bayesian inference, but one wonders why an experiment should be performed when the previous evidence is so strong in favour of Aa.

What could have happened if instead three black offspring we had obtained seven black offspring?

Repeating the calculus for **y** = 7 black, we obtain

$$\text{prior } P(AA) = 0.33 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.99$$
$$\text{prior } P(Aa) = 0.67 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.01$$

If flat priors were used, we obtain

$$\text{prior } P(AA) = 0.50 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.99$$
$$\text{prior } P(Aa) = 0.50 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.01$$

in this case the evidence provided by the data dominates over the prior information. However, if prior information is very large

$$\text{prior } P(AA) = 0.002 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.33$$
$$\text{prior } P(Aa) = 0.998 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.67$$

thus even having more data, prior information dominates the final result when it is very large, which should not be normally the case. In general, prior information loses importance whith larger samples. For example, if we have n uncorrelated data,

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta) f(\theta) = f(y_1, y_2, \mathrm{K}, y_n|\theta) f(\theta) = f(\mathbf{y}_1|\theta) f(\mathbf{y}_2|\theta) \mathrm{K}\, f(\mathbf{y}_n|\theta) f(\theta)$$

taking logarithms

$$\log f(\theta|\mathbf{y}) \propto \log f(\mathbf{y}_1|\theta) + \log f(\mathbf{y}_2|\theta) + \mathrm{K} + \log f(\mathbf{y}_n|\theta) + \log f(\theta)$$

we can see that prior information has les and less importance as the number of data augments.

7.2. Vague prior information

7.2.1. *A vague definition of vague prior information*

It is infrequent to find exact prior information. Usually there is prior information, but it is not clear how to formalize it in order to describe this information using a prior distribution. For example, if we are going to estimate the heritability of litter size of a rabbit breed we know that this heritability has been also estimated in other breeds and it has given often values between 0.05 and 0.11. We have a case in which the estimate was 0.30, but the standard error was high. We have also a high realized heritability in an experiment performed in Ghana, but our prejudices prevent us to take this experiment too seriously. A high heritability was also presented in a Congress, but this paper did not pass the usual peer review filter and we tend to give less credibility to this result. Moreover, some of the experiments are performed in situations that are more similar to our experiment, or with breeds that are closer to ours. It is obvious that we have prior information, but, how can we manage all of this?

One of the disappointments the student that has arrived to Bayesian inference attracted by the possibility of profiting prior information for his experiments receives is that modern Bayesians tend to avoid the use of prior information due to the difficulties of defining it properly. A solution for this problem was offered in the decade of the 30s by the British philosopher and by the Italian mathematician Bruno de Finetti, but the solution is unsatisfactory in many cases as we will see. They propose, in the words of De Finetti that “Probability does not exist”. Thus what we call probability is just a state of beliefs. This definition has the advantage of including events like the probability of obtaining a 6 when throwing a dice and the probability of Scotland becoming an independent republic in this decade. Of course, in the first case we have some mathematical rules that will determine our beliefs and in the second we do not have these rules, but in both cases we can express sensible beliefs about the events.
Transforming probability, which looks as a concept external to us, into beliefs, that looks like an arbitrary product of our daily mood, is a step that some scientists refuse to walk. Nevertheless, there are three aspects to consider:

1. It should be clear that although beliefs are subjective, this does not mean that they are arbitrary. Ideally, the previous beliefs should be expressed by experts and there should be a good agreement among experts on how prior information is evaluated.

2. Prior beliefs should be vague and contain little information; otherwise there is no reason to perform the experiment, as we have seen in 7.1.3. In some cases an experiment may be performed in order to add more accuracy to a previous estimation, but this is not normally the case.

3. Having data enough, prior information loses importance, and different prior beliefs can give the same result, as we have seen in 7.1.3 ([33]).

There is another problem of a different nature. In the case of multivariate analyses, it is almost impossible to determine a rational state of beliefs. How can we determine our beliefs about the heritability of first trait, when the second trait has a heritability of 0.2, the correlation between both traits is -0.7, the heritability of the third trait is 0.1, the correlation between the first and the third traits is 0.3 and the correlation between the second and the third trait is 0.4; then our beliefs about the heritability of the first trait when the heritability of the second trait is 0.1, …etc.? Here we are unable of represent any state of beliefs, even a vague one.

7.2.2. *Examples of the use of vague prior information*

To illustrate the difficulties of working with vague prior information we will show here two examples of attempts of constructing prior information for the variance components. In the first attempt, Blasco et al. (1998) try to express the prior information about variance components for ovulation rate in pigs. As the available information is on heritabilities, they consider that the phenotypic variance is estimated without error, and express their beliefs on additive variances as if they were

[33] Bayesian statisticians often stress that having data enough the problem of the prior is irrelevant. However, having data enough, Statistics is irrelevant. The science of Statistics is useful when we want to determine which part of our observation is due to random sampling and what is due to a natural law.

heritabilities. The error variance priors are constructed according to the beliefs about additive variances and heritabilities.

> "Prior distributions for variance components were built on the basis of information from the literature. For ovulation rate, most of the published research shows heritabilities of either 0.1 or 0.4, ranging from 0.1 to 0.6 (Blasco et al. 1993). Bidanel et al. (1992) reports an estimate of heritability of 0.11 with a standard error of 0.02 in a French Large White population. On the basis of this prior information for ovulation rate, and assuming [without error] a phenotypic variance of 6.25 (Bidanel et al. 1996), three different sets of prior distributions reflecting different states of knowledge were constructed for the variance components. In this way, we can study how the use of different prior distributions affects the conclusions from the experiment. The first set is an attempt to ignore prior knowledge about the additive variance for ovulation rate. This was approximated assuming a uniform distribution, where the additive variance can take any positive value up to the assumed value of the phenotypic variance, with equal probability. In set two, the prior distribution of the additive variance is such that its most probable value is close to 2.5 [corresponding to a heritability of 0.4], but the opinion about this value is rather vague. Thus, the approximate prior distribution assigns similar probabilities to different values of the additive variance of 2.5. The last case is state three, which illustrates a situation where a stronger opinion about the probable distribution of the additive variance is held, a priori, based on the fact that the breed used in this experiment is the same as in Bidanel et al. (1992). The stronger prior opinion is reflected in a smaller prior standard deviation. Priors describing states two and three are scaled inverted chi-square distributions. The scaled inverted chi-square distribution has two parameters, v and S^2 . These parameters were varied on a trial and error basis until the desired shape was obtained. Figure 1 [Figure 2.1 in this book] illustrates the three prior densities for the additive variance for ovulation rate."
>
> **Blasco et al.**, 1998

In the second example, Blasco et al. (2001) make an attempt of drawing prior information on uterine capacity in rabbits. Here phenotypic variances are considered to be estimated with error, and the authors argue about heritabilities using the transformation that can be found in Sorensen and Gianola (2002).

> "Attempts were made to choose prior distributions that represent the state of knowledge available on uterine capacity up to the time the present experiment was initiated. This knowledge is however extremely limited; the only available information about this trait has been provided in rabbits by BOLET *et al.* (1994) and in mice by GION *et al*. (1990), who report heritabilities of 0.05 and 0.08 respectively… Under this scenario of uncertain prior information, we decided to consider

> three possible prior distributions. State 1 considers proper uniform priors for all variance components. The (open) bounds assumed for the additive variance, the permanent environmental variance and the residual variance were (0.0, 2.0), (0.0, 0.7) and (0.0, 10.0), respectively. Uniform priors are used for two reasons: as an attempt to show prior indifference about the values of the parameters and to use them as approximate reference priors in Bayesian analyses. In states 2 and 3, we have assumed scaled inverse chi-square distributions for the variance components, as proposed by SORENSEN *et al.* (1994). The scaled inverse chi-square distribution has 2 parameters, ν and S^2, which define the shape. In state 2, we assigned the following values to these two parameters: (6.5, 1.8), (6.5, 0.9) and (30.0, 6.3) for the additive genetic, permanent environmental and residual variance, respectively. In state 3, the corresponding values of ν and S^2 were (6.5, 0.3), (6.5, 0.2) and (20.0, 10.0). The implied, assumed mean value for the heritability and repeatability under these priors, approximated …[Sorensen and Gianola, 2002], is 0.15 and 0.21, respectively for state 1, 0.48 and 0.72 for state 2, and 0.08 and 0.16 for state 3".
>
> **Blasco et al.** ,2001.

The reader can find more friendly examples about how to construct prior beliefs for simpler problems in Press (2002).

In the multivariate case, comparisons between several priors should be made with caution. We find often authors that express their multivariate beliefs as inverted Wishart distributions (the multivariate version of the inverted gamma distributions), changing the hyper parameters arbitrarily and saying that "as results almost do not change, this means that we have data enough and prior information does not affect the results". If we do not know the amount of information we are introducing when changing priors, this is nonsense, because we can always find a multivariate prior sharp enough to dominate the results. Moreover, inverted Wishart distributions without bounds can lead to sample covariance components that are clearly outliers. There is no clear solution for this problem. Blasco et al. (2003) uses priors based on two different reasons: flat priors and vague priors constructed considering that one of the parameters of the Wishart distribution is similar to a matrix of (co)variances:

> We can then compare the two possible states of opinion, and study how the use of the different prior distributions affects the conclusions from the experiment. We first used flat priors (with limits that guarantee the property of the distribution) for two reasons: to show indifference about their value and to use them as reference priors, since they are usual in Bayesian analyses. Since prior opinions are difficult to draw in the multivariate case, we chose the second prior by substituting a (co)variance matrix of the components in the hyper parameters

> $\mathbf{S}_R$ and $\mathbf{S}_G$ and using $n_R = n_G = 3$, as proposed by Gelman *et al.* [11] in order to have a vague prior information. These last priors are based on the idea that **S** is a scale-parameter of the inverted Wishart function, thus using for $\mathbf{S}_R$ and $\mathbf{S}_G$ prior covariance matrixes with a low value for *n*, would be a way of expressing prior uncertainty. We proposed $\mathbf{S}_R$ and $\mathbf{S}_G$ from phenotypic covariances obtained from the data of Blasco and Gómez [5].
>
> Blasco et al.. 2003.

The reader probably feels that we are far from the beauty initially proposed by the Bayesian paradigm, in which prior information was integrated with the information of the experiment to better asses the current knowledge. Therefore, the reader would not be probably surprised by knowing that modern Bayesian statisticians tend to avoid vague prior information, or to use it only as a tool with no particular meaning. As Bernardo and Smith say:

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

7.3. No prior information

7.3.1. *Flat priors*

Since the origins of Bayesian inference (Laplace, 1774) and during its development in the XIX century, Bayesian inference was always performed under the supposition of prior ignorance represented by flat priors. Laplace himself, Gauss, Pearson and others suspected that flat priors did not represent prior ignorance, and moved to examine the properties of the sampling distribution. Integrating prior information was not proposed until the work of de Finetti quoted before.

It is quite easy to see why flat priors cannot represent ignorance: Suppose we think that we do not have any prior information about the heritability of a trait. If we represent this using a flat prior (figure 4.3), the event A "the heritability is lower than

0.5" has a probability of 50% (blue area). Take now the event B "the square of the heritability is lower than 0.25". This is exactly the same event ([34]) as event A, thus its probability should be the same, also a 50%.

We are as ignorant about h^2 as about h^4, thus we should represent the ignorance about h^4 also with flat priors if they represent ignorance. However, if we do this and we also maintain that $P(h^4<0.25) = 50\%$ we arrive to an absurd conclusion: we do not know nothing about h^2 but we know that h^4 is closer to zero than h^2 (figure 4.3).

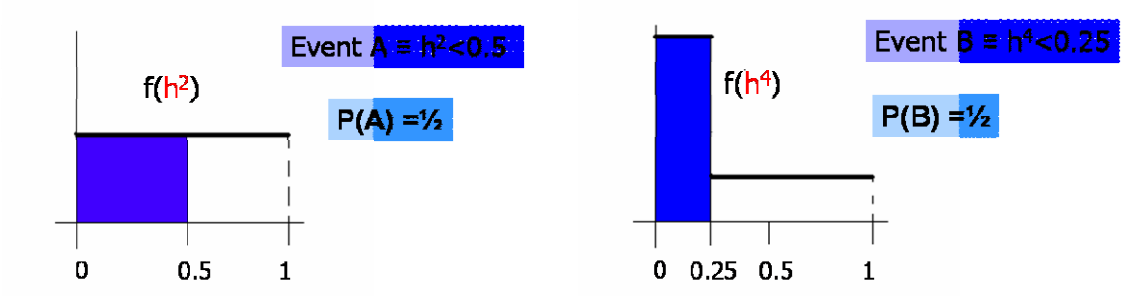

Figure 4.3. Flat priors are informative

To avoid this absurd conclusion we have to admit that flat priors do not represent ignorance, but they are informative. The problem is that we do not know what this prior information really means. However, this information is very vague and should not cause problems; in most cases that data will dominate and the results will not be practically affected by the prior.

7.3.2. *Jeffreys priors*

Ideally, the estimate of a parameter should be invariant to transformations. It is somewhat annoying that an estimate of the variance (for example, the mean of the marginal posterior distribution of the variance) is not the square of the same estimate of the standard deviation (the mean of the marginal posterior distribution of the standard deviation). If we want to be coherent, our prior information of a function of a parameter should be obtained as we showed in 3.3.2. For example, is we have the prior information on the standard deviation, the prior information of the variance should be

[34] If the reader does not like squares, take half of the heritability or whatever other transformation.

$$f\left(\sigma^2\right) = f\left(\sigma\right)\left|\frac{d\sigma^2}{d\sigma}\right|^{-1} = f\left(\sigma\right)\left(2\sigma\right)^{-1} = \frac{1}{2\sigma}f\left(\sigma\right)$$

Harold Jeffreys proposed to use a prior invariant to transformations, so that if f(θ) is a Jeffreys prior for θ then $[f(\theta)]^2$ is a prior for θ^2. For example, Jeffreys prior for the standard deviation is

$$\text{Jeffrey's prior } f\left(\sigma^2\right) \propto \frac{1}{\sigma}$$

Then, the prior for the variance is

$$f\left(\sigma^2\right) = \frac{1}{2\sigma}f\left(\sigma\right) \propto \frac{1}{2\sigma}\cdot\frac{1}{\sigma} \propto \frac{1}{\sigma^2}$$

which is the square of the Jeffreys prior for the standard deviation.

The Jeffrey's prior of a parameter θ is:

$$\text{Jeffreys prior } f\left(\theta\right) \propto \sqrt{E_y\left(\frac{\partial \log f(\mathbf{y}\,|\,\theta)}{\partial\theta}\right)^2}$$

For example, the Jeffreys prior of the variance is

$$f\left(\sigma^2\right) \propto \sqrt{E_y\left(\frac{\partial \log f(\mathbf{y}\,|\,\sigma^2)}{\partial\sigma^2}\right)^2} \propto \frac{1}{\sigma^2}$$

The deduction of this prior is in appendix 7.1. In Appendix 7.2 we show that these priors are invariant to transformations.

Jeffreys priors are widely used in univariate problems, but they lead to some paradoxes in multivariate problems.

7.3.3. *Bernardo's "Reference" priors*

If we do not have prior information and we know that all priors are informative, a sensible solution may be to use priors with minimum information. Bernardo (1979) proposed to calculate posterior distributions in which the amount of information provided by the prior is minimal. To build these posterior densities we need:

1. To use some definition of information. Bernardo proposes the definition of Shanon (1948), an engineer of the Bells laboratory who defined the information transmitted through a channel in a way that was useful for statisticians.

2. To define the amount of information provided by an experiment: This is defined as the distance between the prior and he posterior information, averaging for all possible samples.

3. To use some definition of distance between distributions. Bernardo uses the Kullback's divergence between distributions, based on Bayesian arguments.

4. To find a technically feasible way for solving the problem and deriving the posterior distributions. This is technically complex and reference priors are not easy to derive.

These priors have the great advantage of being invariant to reparametrization.

In the multivariate case we should transform the multivariate problem in univariate ones taking into account the parameters that are not of interest. This should be made, by technical reasons, conditionalising the other parameters in some order. The problem is that the reference prior obtained differs depending on the order of conditionalization. This is somewhat uncomfortable, since would oblige to the scientist to consider several orders of conditionalization to see whether the results differ. This can be done with few parameters, but not in cases in which the problems

are highly parametrized. For the later cases, the strategy would be to use sensible vague informative priors when possible or to nest the problems as geneticist do when they are considering that all genetic effects are distributed as a Normal (0, $\mathbf{A}\sigma_u^2$) where only σ_u^2 is unknown.

Bernardo's reference priors are obviously out of the scope of this book. However, José-Miguel Bernardo is developing in Valencia University a very promising area in which the same "reference" idea has been applied to hypothesis test, credibility intervals and other areas, creating a "reference Bayesian statistics" that will be probably used in next years when software will be available.

7.4. *Improper priors*

Some priors are not densities, for example: f(θ) = k, where k is an arbitrary constant, is not a density because $\int$ f(θ) dθ = ∞. However, improper priors lead to proper posterior densities when

$$f(y) = \int f(y|\theta)\, f(\theta)\, d\theta < \infty$$

Sometimes they are innocuous and they are not used in the inference, for example,

y~N(μ,1)
μ ~ k

$$f(y) = \int f(y \mid \mu) f(\mu) d\mu = \int f(y \mid \mu) \cdot k \cdot d\mu = k \cdot \int f(y \mid \mu) \cdot d\mu =$$

$$= k \int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2} \right] d\mu = k \int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(\mu-y)^2}{2} \right] d\mu = k$$

$$f(\mu \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \mu) f(\mu)}{f(\mathbf{y})} = \frac{f(\mathbf{y} \mid \mu) \cdot k}{k} = f(\mathbf{y} \mid \mu)$$

thus in this case the posterior density of μ does not take into account the prior.

In general, it is recommended to use always proper priors, to be sure that we always obtain proper posterior densities. When using Gibbs sampling, some densities look as proper ones and they may be improper. Although when using MCMC all densities are in practice proper ones (we never sample in the infinite), samples can have very long burning periods and can lead to chains that only apparently have converged. The recommendation is always to used proper priors (bounded priors with reasonable limits, for example), unless it has been proved that they are innocuous (Hobert and Casella, 1992).

7.5. *The Achilles heel of Bayesian inference*

Bayesian inference, or Inverse probability, as it was always called before and should still be called, is extremely attractive because of the use of probabilities and the possibility of integrating prior information. However, integrating prior information is much more difficult that the optimistic Bayesians of the fifties thought. This lead to use several artefacts in order to make possible the use of probability. Some statisticians think that an artefact multiplied by a probability will give an artefact and not a probability, and consequently they are reluctant to use Bayesian inference. There is not a definitive answer to this problem, and it is a matter of opinion to use Bayesian or frequentist statistics, both are now widely used and no paper will be refused by a publisher because it uses a type or the other type of statistics.

Many users of statistics, like the author of this book, are not “Bayesians” or “frequentists”, but just people with problems. Statistics is a tool to help in solving these problems, and users depend more on the existence of easy solutions and friendly software than in the background philosophy. I use Bayesian statistics because I understand probability better than significance levels and because it permits to me to express my results in a more clear way for later discussion. Some other users prefer Bayesian statistics because there is a route for solving their problems: to make a joint posterior distribution, to find the conditionals and to use MCMC to find the marginal distributions. We *behave as* if we were working with real probabilities, (this should not be objected by frequentists). To know the true probabilities drives us to the PROBLEM OF INDUCTION, a very difficult problem that

we cannot expose in this lecture notes. The Appendix “Three new dialogues between Hylas and Filonus” tries to expose this problem and its difficulty in a literary form.

> “*Felix qui potuit rerum cognoscere causas*” (Happy the man that can know the causes of the things!
>
> **Virgil**, Egloga IX, 63

Appendix 7.1

$$f\left(\sigma^2\right) \propto \sqrt{I\left(\sigma^2 \mid \mathbf{y}\right)} = \sqrt{E_y\left(\frac{\partial \log f(\mathbf{y} \mid \sigma^2)}{\partial \sigma^2}\right)^2}$$

$$f(\mathbf{y} \mid \sigma^2) = \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sigma^2\right)^{\frac{n}{2}}} \exp\left[-\frac{\sum_1^n \left(y_i - \mu\right)^2}{2\sigma^2}\right]$$

$$\log f(\mathbf{y} \mid \sigma^2) = k - \frac{n}{2}\log\sigma^2 - \frac{\sum\left(y_i - \mu\right)^2}{2\sigma^2}$$

$$\frac{\partial \log f(y \mid \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum\left(y_i - \mu\right)^2}{2\sigma^4}$$

$$E_y\left(\frac{\partial \log f(\mathbf{y} \mid \sigma^2)}{\partial \sigma^2}\right)^2 = E_y\left(\frac{n^2}{4\sigma^4} + \frac{\sum (y_i - \mu)^2}{2\sigma^8} - \frac{2n}{2\sigma^4}\right)$$

$$E_y\left[\sum_1^n \left(y_i - \mu\right)^2\right] = nE_y\left(y_i - \mu\right)^2 = n\sigma^2$$

$$E_y\left(\frac{\partial \log f(\mathbf{y} \mid \sigma^2)}{\partial \sigma^2}\right)^2 = \frac{n^2 - 4n}{4\sigma^4} + \frac{n\sigma^2}{2\sigma^8} \propto \frac{1}{\sigma^4}$$

$$f\left(\sigma^2\right) \propto \sqrt{I\left(\sigma^2 \mid \mathbf{y}\right)} = \sqrt{E_y\left(\frac{\partial \log f(\mathbf{y} \mid \sigma^2)}{\partial \sigma^2}\right)^2} \propto \sqrt{\frac{1}{\sigma^4}} \propto \frac{1}{\sigma^2}$$

Appendix 7.2

$$I(\lambda \mid \mathbf{y}) = E_y\left(\frac{\partial \log f(y \mid \lambda)}{\partial \lambda}\right)^2 = E_y\left(\frac{\partial \log f\left(y \mid g(\theta)\right)}{\partial \theta} \cdot \frac{\partial \theta}{\partial \lambda}\right)^2 = \left(\frac{\partial \theta}{\partial \lambda}\right)^2 E_y\left(\frac{\partial \log f\left(y \mid g(\theta)\right)}{\partial \theta}\right)^2 =$$

$$= \left(\frac{\partial \theta}{\partial \lambda}\right)^2 I(\theta \mid \mathbf{y})$$

$$f(\theta) \propto \sqrt{I(\theta \mid \mathbf{y})}$$

$$f(\lambda) \propto \sqrt{I(\theta \mid \mathbf{y})} \cdot \left|\frac{\partial \theta}{\partial \lambda}\right| = \sqrt{I(\theta \mid \mathbf{y}) \cdot \left|\frac{\partial \theta}{\partial \lambda}\right|^2} = \sqrt{I(\lambda \mid \mathbf{y})}$$

FURTHER READING

Blasco A. 2001. The Bayesian controversy in Animal Breeding. J. Anim. Sci. 79: 2023-2046.

Sorensen D.A., Gianola D. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer. New York.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2003. Bayesian Data Analysis. (2nd ed.). Chapman and Hall.

Press S.J. 2002. Subjective and Objective Bayesian Statistics: Principles, Models, and Applications. Wiley.

Robert C., Casella G. 2004. Monte Carlo Statistical Methods (2nd ed.). New York: Springer-Verlag, 2004.

REFERENCES

REFERENCES

Bayarri M.J., Berger J.O. 2004. The Interplay of Bayesian and Frequentist Analysis. Statistical Science19: 58–80.

Berger J.O., Wolpert R.L. 1984. The likelihood principle. Institute of Mathematical Statistics. Lecture Notes - Monograph Series. Purdue University.

Bernardo J. M., 1979. Reference posterior distributions for Bayesian inference. J. Royal Statist. Soc. B 41:113-147.

Bernardo J. M., Smith F.M. 1994. Bayesian theory. Wiley.

Blasco A. 2001. The Bayesian controversy in Animal Breeding. J. Anim. Sci. 79: 2023-2046.

Blasco A., Piles M., Varona L. 2003. A Bayesian analysis of the effect of selection for growth rate on growth curves in rabbits. Genetics Selection Evolution.35: 21-42.

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

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

Blasco A., Argente M.J., Santacreu M.A., Sorensen D., Bidanel J.P. 2001. Bayesian analysis of response to selection for uterine capacity in rabbits. J. of Anim. Breed. And Genet. 118: 93-100.

Box G.E.P., Draper N.R. 1987. Empirical Model-Building and Response Surfaces. Wiley.

Damgaard L. H. 2007. Technical note: How to use Winbugs to draw inferences in animal models. J. Anim. Sci. 85:1363–1368

De Finetti B. 1937. La prévision: ses lois logiques, ses sources subjectives. Annales de l'Institut Henri Poincairé 7 :1-68. Translated in H.E. Kyburg and H.E. Smokler, 1964, Studies in subjective probability. Wiley.

Edgeworth F.Y. 1908. On the probable error of frequency constants. J. R. Stat. Soc. 71:381-397, 499-512, 651-678, Addendum in1908, 72:81-90.

Fienberg S.E. 2006. When Did Bayesian Inference Become “Bayesian"?. Bayesian Analysis 1: 1-40.

Fisher R.A. 1912. On an absolute criterion for fitting frequency curves. Messenger Math. 41: 155-160. Reprinted in Stat Sci. 1997, 12: 39-41.

Fisher R. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1:3-32.

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

Fisher R. 1925. Statistical Methods for research workers. Oliver and Boyd.

Fisher R. 1935. The logic of inductive inference. J. R. Stat. Soc. 98:39-82.

Fisher R. 1936. Uncertain inference. Proc. Am. Acad. Arts Sci. 71:245-258.

Fisher R. 1950. Contributions to Mathematical Statistics. Wiley.

Fisher R. 1956. Statistical methods and scientific inference. Oliver and Boyd.

Gauss C.F. 1809. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Translated by C.E. Davis. 1963. Dover. New York.

Gelfand A.E., Smith F.M. Sampling-Based Approaches to Calculating Marginal Densities. 1990. J. American Statistical Association, 85:398-409.

Gelman, A., and D. B. Rubin, 1992 Inference for iterative simulation using multiple sequences. Stat. Sci. 7**:** 457–472.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2003. Bayesian Data Analysis. Chapman and Hall.

Geman S., Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEE Transactions on pattern analysis and machine intelligence 6:721-742.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments In Bayesian Statistics 4. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. Smith, eds. Oxford University Press.

Gilks, W. R; Richardson, S.; Spiegelhalter, D. J. (Eds), 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall

Hald A. 1998. A history of mathematical statistics from 1750 to 1930. Wiley.

Hastings W.K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57:97-109.

Henderson C.R. 1973. Sire evaluation and genetic trends. In: Proc. Anim. Breed. and Genet. Symp. in honor of Dr. J. L. Lush. Blacksburg, Virginia. pp.10-41.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69.

Hobert J.P., Casella G. 1996. The effect of improper priors on Gibbs sampling in Hierarchical Linear Mixed Models. J. of the Amer. Stat. Assoc. 436:1461-1473.

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

Johnson, V. E. 1996. Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. J. Am. Stat. Assoc. 91:154-166.

Jeffreys H. 1961. Theory of probabilities. 3rd. ed. Clarendon Press.

Kant E. 1781. Critique of pure reason. Reprinted in translation. Orbis S.A.

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

Kendall M., Stuart A., Ord K. 1998. Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory. Arnold.

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.

Kendall 1961. Daniel Bernouilli on Maximum Likelihood. Biometrika 48:1-8.

Keynes J.M. 1921. A treatise on probability. Macmillan Publ. Co. London.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller N.H., Teller E. 1953. Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21:1087-1092.

Metropolis N., Ulam S. 1949. The Monte Carlo method. J. Amer. Stat. Assoc. 44: 335-341.

Mill J.S. On Liberty. 1848. Reprinted in 2006 by Longman.

Misztal I., Tsuruta S., Strabel T., AuvrayB. , Druet T., Lee D.H. Blupf90 and related programs (bgf90). In 7th World Congress on Genetics Applied to Livestock Production, pages CD–ROM Communication N 28–07, 2002.

Mood A.M., Graybill F.A. 1963. Introduction to the theory of Statistics. MacGraw Hill.

Neyman J., Pearson E. 1933. On the problem of the most efficient test of statistical hypotheses. Phil Trans. of the Royal Soc. 231A: 289-337.

Pearson E. 1962. Some thoughts on statistical inference. Ann. Math. Stat. 33: 394-403.

Pearson K. 1920. The fundamental problems of practical statistics. Biometrika, 13:1-16.

Popper K. 1936. The logic of Scientific discovery. Reprinted by Routledge. 2002.

Press S.J. 2002. Subjective and Objective Bayesian Statistics: Principles, Models, and Applications. Wiley.

Raftery, A. E., and S. M. Lewis, 1992 How many iterations in the Gibbs sampler?, In Bayesian Statistics, Vol. 4, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. (Ed.) Oxford Univ. Press.

Ramsey F.P.1926. In: The foundation of athematics and other logical essays R.B. Braithwaite (Ed.) Adams &Co. 1965.

Robert C. P. 1992. L'Analyse statistique bayesienne. Economica. Paris. France.

Robert C., Casella G. 2004. Monte Carlo Statistical Methods (2nd ed.). Springer-Verlag.

Robinson G. K. 1991. That BLUP is a good thing: the estimation of random effects. Stat. Sci. 6:15-51.

Shanon C.E. 1948. A mathematical theory of communication. Bell System Tech. J. 27:379-423, 623-656.

Sorensen D. A., Wang C.S., Jensen J., Gianola D. 1994. Bayesian analysis of genetic change due to selection using Gibbs sampling. Genet. Sel. Evol. 26:333-360.

Sorensen D.A., Gianola D. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer. New York.

Stigler S.M. 1983. Who discovered Bayes Theorem? Am. Stat. 37: 290-296.

Stigler S.M. 1986. The History of Statistics: The measurement of uncertainty before 1900. Harvard Univ. Press.

Student. 1908. On the probable error of a mean. Biometrika, 6: 1-25.

Van Tassell C.P. Van Vleck L.D. 1996. Multiple-trait gibbs sampler for animal models: flexible programs for bayesian and likelihood-based (co)variance component inference. J Anim Sci, 74:2586–2597.

Visscher P.M. 1998. On the Sampling Variance of Intraclass Correlations and Genetic Correlations. Genetics 149: 1605–1614.

Von Mises R. 1957. Probability statistics and truth. Macmillan Publ. Co. London, UK.

Wang C.S., Gianola D., Sorensen D.A., Jensen J., Christensen A., Rutledge J.J. 1994. Response to selection for litter size in Danish Landrace pigs: a Bayesian analysis. Theor. Appl. Genet. 88:220-230.

APPENDIX

THREE NEW DIALOGUES BETWEEN HYLAS AND FILONUS

PRELIMINARY NOTE

The dialogues between Hylas and Filonus were discovered by the Bishop Berkeley (to which a famous University owes its name), at the beginning of the XVIII century. Berkeley was so impressed by Filonus' arguments (without doubt a pseudonym, which I will explain later), that he based his philosophy on the reasoning he found in the dialogues. His book "A Treatise concerning the principles of Human knowledge" is no more than a development of the reasons Filonus expresses for doubting an ability to attain real and true knowledge of the world, and even about the existence of a world beyond that which is perceived through sensory impressions. For many years a favourite task of philosophers from all schools has been to refute Filonus, but the arguments of this protagonist of the dialogues are much more difficult to refute than might appear at first glance. Curiously, a refutation of Filonus is offered by Hylas himself and is recorded in three other dialogues that the well-known philosopher of science Mario Bunge brought to light, although, for reasons that I shall give later, I suspect they are somewhat apocryphal. I wondered about trying to discern which parts of the dialogues published by Bunge could be considered as an authentic transcription of the conversations between Hylas and Filonus and which should be cast aside as having been added subsequently by other copyists, when I came upon a number of notes in diverse handwriting reconstructing other dialogues, which took place around the same time. The difficulty in reconstructing them lies in the fact that the recuperated fragments are not in order, which is why I propose trying to follow the arguments as I find it more logical. However, there is room for other researchers to propose not only a different order, but also to add a few paragraphs to those I could not fit into the right place or to interpret the gaps and intertwine them more precisely with the discussions that have already been preserved.

In the first dialogue Filonus studies the problem of induction, starting with one of Hylas' arguments (contained in the third dialogue recorded by Bunge), which hardly seems convincing to him. In Hylas' opinion induction is possible, for if it were not, science and technique would not be possible either. Filonus makes him see that the certainty of his scientific predictions is impossible, to which Hylas agrees unwillingly, although he offsets Filonus' argument quoting that at least our knowledge of things is probable. Here follows an interjection about what we can assume is probability, and Filonus, whose name means "linked to the mind"([35]), makes Hylas see that induction is only possible by considering probability as a degree of personal belief. This surprises Hylas (whose name means 'linked to matter'), who thinks that a degree of personal belief is unacceptable to infer something about Nature, but as the dialogue develops he realises the drawbacks of considering probability in a way that is external to the investigator. Finally he accepts that probability can be defined in the broad sense as this state of belief, but he is worried about the consequences that might arise for science and the interpretation of the world from transforming a tool, he thought was objective, into a scale of doubt of a subjective kind. In the second dialogue, Hylas confesses to his friend Filonus his uneasiness at not being able to escape from subjectivity when occupied with something so apparently objective as science, and in particular, by the need to refer to previous opinions about the results we are going to obtain, often expressed in a vague way. However, Filonus makes him see that it is precisely this use of previous information, even though it is irrelevant, which allows us to use probability to determine the uncertainty associated with experiments, something that is not possible using other theories. In the third dialogue, when it seems as though all has been concluded, the induction problem solved, Hylas alerts Filonus to the fact that methods of belief always end up by indirectly assuming untested hypotheses. Filonus observes that

[35] Which leads us to the assumption that both names are mere pseudonyms.

Hylas' reasons are valid, but also that, in admitting them, although scientific activity would still be possible on a rational basis, he still would not be able to resolve the induction problem. On parting, they are both worried and promise to meet again to continue investigating the matter.

I have arbitrarily divided the dialogues into three groups, because Berkeley did it that way and so did Bunge, but this responds more to a thematic order than to the development of the discussion the way I infer it to have taken place. The reader who wants to learn more about the protagonists' arguments can refer to previous dialogues, found in Alianza and Ariel publications, collected by Berkeley and Bunge respectively. In some places I have had to fill in some gaps found in the manuscripts, making note of the topic that was introduced in the discussion, although on very few occasions was I obliged to interrupt the conversation. The footnotes are mine entirely.

FIRST DIALOGUE

HYLAS.- Good morning, Filonus! What is the result of your meditations? Now do you admit that we can perceive more than what our senses reveal to us? Or on the contrary, do you still think that the material world is unknown to us?

FILONUS.- I admit it, Hylas, and I also admit that we can conclude that many years ago events of which no one was conscious took place, that we can build entities just like irrational numbers, which no one has perceived [36], and that we have innate [37] information that does not come from our perceptions. But yesterday you used, my dear friend, an argument that made me so uncomfortable that I have not been able to get it out of my mind from the moment we parted. You told me that we could infer a law of Nature through the induction of a great number of specific cases.

HYLAS.- That is what I believe. How would science work otherwise? How would it be possible for us to build bridges without uncertainty about whether they are going to collapse, causing everyone on them to fall, were it not that our experience showed us the conditions that made them safe? Could you live in the doubt of whether the sun will shine tomorrow [38] ?

FILONUS.- I understand, Hylas, the difficulties implied in denying your reasoning, but I do not find any problem in asserting that the sun may have already disappeared and that we shall not know it until the few minutes its light takes to reach us have expired [39]. My faith in the new bridge not collapsing has not prevented the falling of others, and I can only tell you

[36] These arguments are collected in the first and second dialogues published by M. Bunge.

[37] These argument is not included in any of the published dialogues. It probably belongs to a lost fragment.

[38] This example is often associated to Hume (1739), who published his Treatise of Human Nature later to the conversations between Hylas and Filonus, but can be found yet in Sixtus Empiricus, in the 2nd century A.D.

[39] The speed of light was not estimated until the middle of the 19th Century, thus this fragment should be apocryphal.

that if I live without this sort of preoccupations it is due to the fact that my Nature does not make me think of them, and I act as if a knew for sure what I don't know, not even as a probability, and I do not get upset by thinking the Universe could end today or tomorrow [40]. I do not understand what rule, according to you, God has created so that upon listing several particular cases we can make inferences about cases we do not even know about.

HYLAS.- I admit, Filonus, that we cannot speak with the certainty of those who have perfect knowledge of all the facts Nature provides, but you cannot deny that we can qualify certain events as highly improbable.

FILONUS.- I do not understand, Hylas, what you mean by the word "improbable". Are you referring to the number of bridges that collapse with respect to the ones that remain solid?

HYLAS.- I find your proposal quite shocking, and I think it is inspired mostly by a game of irony, more than by a discussion that tries to illuminate the darkest corners of these concepts; but even so I could accept what you say as an expression of uncertainty regarding the doubt that a bridge will collapse.

FILONUS.- And how many times has the sun exploded in order for us to be able to calculate the probability that it will go up in smoke again?

HYLAS.- You are joking Filonus, and I beg you to centre our conversation on concrete examples to which I can give you an answer.

FILONUS.- Following your reasoning, Hylas, few are the examples of doubt to which we could apply probability. If we do an experiment to infer something or other, how can we associate what is probable to the results if we have never seen in how many cases our predictions have failed?

HYLAS.- You can imagine you repeated that experiment an infinite number of times and that your result is one derived from one of the possible repetitions.

FILONUS.- You ask for too much imagination, my friend Hylas. I beg you, please give me an example of your mode of action because I find myself a little disoriented and I do not know how to infer general laws from things I have not yet seen.

HYLAS.- Let us suppose, Filonus, that I toss a coin into the air and I record the results.

FILONUS.- I accept the example, Hylas.

HYLAS.- After repeating the same experiment several times, half the time you will have got heads and the other half you will have got tales, to which I can say that we have a 50% probability of getting heads the next time we throw. There you can see how I infer about things that I have not seen.

FILONUS.- But if after one thousand repetitions, 60% of the time you got heads and the other 40% you got tails, what would you say then?

HYLAS.- I would without any doubt postulate the fact that the coin is rigged.

---

[40] This reminds me the "animal faith" of my namesake Agustín Santayana, professor of philosophy in Harvard at the end of the 19th Century and beginning of the 20th Century.

FILONUS.- What would you say if I tossed the coin a million times and it resulted in heads 55% of the time and in tails the rest of the time?

HYLAS.- I would also say the coin is rigged.

FILONUS.- Then the only possible definition for a coin is 'that object which is thrown in the air and falls on the head's side approximately half the time', is that not right?

HYLAS.- I see Filonus that you are accusing me of introducing the defined into the definition, but it is not like that. What I am saying is that with my rigged coin, getting heads 60% of the time and tails 40%, I can deduce what results I may get in subsequent tosses, so I will then say that the probability of getting tales is only 0.4 if I obstinately continue to use it.

FILONUS.- I have a feeling we shall reach a poor conclusion, Hylas, but I would like you to please give me yet another example related to the laws of Nature.

HYLAS.- Well now, suppose we are walking about the countryside, let us imagine that I want to know how much a new-born piglet from Hampshire[41] weighs. I would get a good number of them, weigh them, and I would infer a general law of Nature.

FILONUS.- And how would you know that the real value is not very different from the one you obtained?

HYLAS.- Because if I repeated collecting and weighing the piglets an infinite number of times, all the values obtained would revolve around the true value.

FILONUS.- Hylas, you are asking me to make conclusions based not only on the values you obtained but also on those you could have obtained but did not. Do you think it is wise and safe to draw conclusions about laws basing yourself on something that you have not even done?

HYLAS.- I draw no conclusions from experiments I have never done, but rather from those I have done and my knowledge of what would happen if I repeated them.([42]).

FILONUS.- Let us then suppose that you have used scales that can weigh up to ten kig([43]), and let us suppose that your values are all around one kilo; let us say they were 0.6, 0.8, 1, 1.2 and 1.4 kg.

HYLAS.- I would then conclude that the average weight is 1kg.

FILONUS.- But, what would you say if I told you that the scales are broken and you cannot weigh anything heavier than 1.4kg?

---

[41] My friend Luis Silió has called my attention to the fact that the Hampshire bred was not called by this name until the beginning of the 20th Century. He suggests this to be an apocryphal fragment, but I doubt it because it fits too well into the context. It might be or that they talk about pigs from Hampshire (not the Hampshire breed) or just an error of transcription, and they talk about the Berkshire pig, a well documented breed since the middle of the 18th Century.

[42] Which today we would say is "of our knowledge of the distribution of the estimator in the sampling space".

[43] Of course the measurements of weight in the original are in eighteenth century English pounds; I have adapted them to the current European units.

HYLAS.- Then I would have to increase my estimation, because if I had repeated the experiment many times, I would have got values higher than 1.4kg which I would have been unable to detect.

FILONUS.- But then if our friend Bishop Berkeley were to tell us that he had fixed the scales without me noticing, what value would you assign to the weight of our new - born in Hampshire?

HYLAS.- Again, without doubt, 1 kg.

FILONUS.- Which means that you are modifying the value of your inference, not based on facts from your experiment, which never exceeded the value that was apparently on the broken scales, but rather on the basis of the experiments that you could have carried out but never did. And do you think that any wise man should proceed by basing his inference on the imaginary results of experiments that have never taken place?

HYLAS.- I have seen very talented men proceeding in that way, and you cannot blame anyone if they have been informed by an insuperable error as to what results they can expect.

FILONUS.- Let's set another example: How can you associate probabilities to facts like, for example: it might rain tomorrow; new worlds may be discovered; the enterprises in which we invest will prosper? How can we know if the decisions taken at a crucial moment are going to be the right ones?

HYLAS.- I have no clue, and I don't think anybody on this earth has a solution to that.

FILONUS.- Consider, Hylas, that we can determine the uncertainty by expressing our degree of belief that an event will occur, or our opinion about the value that such a character as the weight of the newly-born piglets will receive.

HYLAS.- I do not understand what it is you are proposing, Filonus. Do you want me to assign values to my beliefs in such a way that they obey the laws of probability?

FILONUS.- That is what I pretend, and nothing else. I am sure that before weighing those piglets you did not believe they could weigh ten kilos, nor even ten grams. Assign, then, a value to your beliefs and we shall then see the results of the experiment.

HYLAS.- I find what you propose highly irregular, Filonus. I may have some beliefs whilst you have others, and how ever many times we mix them with the results from an experiment they shall always end up being our personal beliefs, nothing solid on which to base an inference.

FILONUS.- I want to make you realise, Hylas, that the experiments are not so conclusive, at least not usually, as to confirm or refute a theory, or even in many cases to be able to achieve a precise measurement. And I also remind you that we submit every result to discussion and to personal interpretations.

HYLAS.- But at least we can separate what facts we have from what we later add. ([44]).

FILONUS. And what attitude would you adopt if the facts told you that the piglets of your experiment weighed thirty kilos when born? Wouldn't you at least be doubtful of the

[44] Here we find a precedent to the well-known aspiration "let the facts speak for themselves".

prediction of your scales?

HYLAS.- Your question has an implicit answer, but I do not know in what way that modifies my argument.

FILONUS.- And what if you were to comment on the weight of the piglets with the farmers of the region and they were to tell you that they weigh much less than what has been seen on other occasions, what would you say then?

HYLAS.- I do not have an answer for I did not do the experiment, but I suppose that I would conclude to the fact that the piglets I weighed were ill, or I would reach some similar sort of conclusion that would permit me to explain the anomaly in my facts.

FILONUS.- And what if I were to tell you that the information you got from some of the farmers isn't reliable, because they are all a bunch of rascals and like to laugh at city people; or because they are all completely stupid and would not know a scale any better than their own pigs would?

HYLAS.- It is obvious, in that case, that I would give very little credit to their affirmations, whilst I would do the contrary with those farmers whom I could trust.

FILONUS.- And wouldn't you then, my good Hylas, be concluding from earlier information, subjectively evaluated, and not from your facts?

HYLAS.- That's how I would be doing it, but I still insist that I would then separate what my opinions are from what I add, meaning the facts that I obtain from my personal interpretation.

FILONUS.- I do not understand very clearly what is the purpose of this separation if you are not going to use it later on. But let us return to my original argument: if you want to know how very probable your results are, you need to use some sort of subjective evaluation of that probability.

HYLAS.- I have already told you before what I understand to be probability and how I would evaluate the uncertainty of my experiment, and I have not needed to recur to my personal world for it, but instead I have talked to you about objective measures.

FILONUS.- I see that my criticising your method has made very little impression. But what would you say if I concluded that your probability is a particular case of this much bigger one that I propose?

Here there is a caesura that I have been unable to reconstruct. It is obvious that Filonus proposes to Hylas something similar to the modern concept of the 'interchangeability' of facts. When the order in which the facts from a sample appear does not affect the probability of this sample, the facts are 'interchangeable', and in that case the theory of probability as frequency is a specific case of the subjective theory ([45]). This only affects Filonus' arguments in a marginal way, but it also should be indicated to conserve the connection in the dialogue.

HYLAS.- They worry me, Filonus, all these things you tell me, and you must let me reflect upon them, for it is the first time I am faced with such a surprising fact as that

[45] That is, of one of the versions of the subjective theory (of the "objective" version, if you will excuse the play on words). The concept was proposed by Bruno de Finetti, and it can be found in any textbook on Bayesianism.

probability could be a state of my opinion and not something that remains detached from me, which is was I have thought up to now.

FILONUS.- I find it more surprising the way you mentally repeat an experiment. And what to say about your probabilities when I apply them to astronomy or to other events that tomorrow brings.

HYLAS.- But it is true that we accept in a more comfortable manner the difficulties in a theory we were taught during our youth, and that we live with them with the same ease as we do with problems that our relatives cause, considering them an inevitable product of fate and accepting them with resignation and patience, whilst we are more intolerant with the difficulties that new theories cause, however much they are going to improve our life and our understanding of the laws that govern Nature.

FILONUS.- What you say is exact, Hylas, and I think it would be wise to leave the discussion here so you that your understanding may become accustomed to the new concepts with which you have challenged it today. Contrary to what philosophers would have us believe, reason is not a perfect machine that admits anything new if the syllogism is correct, and I have seen many a wise man wondering around stubborn, ignoring good reasoning, until they were later permitted to change their way of thinking; and even then I am only referring to those cases in which they did.

SECOND DIALOGUE

FILONUS.- Good Morning Hylas, what has been the result of your reflections? When I left you yesterday you seemed worried and I did not want to bother you during the evening, engrossed as you were in your meditations. Do you now find Nature a little more intelligible?

HYLAS.- Say no more, my friend, for I am making myself ill thinking that I cannot infer anything real from the world and that all I have are the opinions of the people I have spoken with! I thought that, if not with the certainty to which every man aspires, I could at least assert that I observe things that exist and are not a part of me. I thought that by means of experiment and skill I would be able to deduce the laws that govern the world, and that way I would also understand more clearly the Creator's work and what His means and purposes were. Now I feel dismayed, for I can only count on the beliefs of my neighbours, without anything outside of each man to help me explain the causes of things.

FILONUS.- Hylas, speaking to you tires me (and I beg you, do not feel offended, for you know I speak to you with the love of a friend) , because you consistently continue to doubt about the reality that surrounds you([46]). Do you think by chance that your neighbours are demented and they are going to tell you things that make no sense as soon as they meet you? Do you not see that one thing is that probability may be subjective and another that it may not be reasonable?

HYLAS.- But before I used to believe in the reality of things that I had to estimate, and I knew a true and immutable value existed around which the data obtained in my experiments

[46] This probably refers to the doubts expressed by Hylas at the beginning of the second of the dialogues recorded by Bishop Berkeley.

would revolve. Now that value has disappeared, it no longer exists, I no longer mention it, my affirmations are now about the probability you get one value or another, whilst the real value has disappeared!

FILONUS.- It has not disappeared, although you are right when you say that you no longer mention it. And why would you want to mention something that not only is unknown, but that you are never going to know about anyway? What is the use of referring to a value that is not only unknown but also unknowable? Yes, that value exists, and it continues existing while you speak as I do, and it remains solid and changeless in the same world as the Hampshire pig, the ideal jug, the young man and the horses ([47]). I am simply speaking about the probability of a determined weight when the pigs are born, and not about what would be this unknowable value.

HYLAS.- But earlier on I also wanted to talk about probabilities and you forbade me.

FILONUS.- You did not speak to me about the probability of the piglets weight being one or another, in light of the sample, but instead you spoke to me about what probability there was that this sample would take on this or that value. And if I have to be frank, then I do not understand why you have so much interest in knowing what the probability of your sample is instead of worrying about the essence and the core of the problem; meaning what probability there is for the weight of the new-born piglet to adopt this or that value.

Here there is an important caesura in the original papers. Hylas and Filonus have obviously been arguing about a form of what we know today as *inverse probability,* made popular by the theorem attributed to Bayes (1780)([48]).

FILONUS.- And I made you realise, Hylas, that the only way you can talk about probability being associated to your facts is by considering previous probability, the one that your observations generate, and finally the one that results from combining both([49]). Your affirmations cannot be linked to probability in any other way.

HYLAS.- Well, then what value will my predictions have? Do you not realise that if I gave my opinion *a priori* a high category, the result would be determined by that opinion which I have introduced? Do you not see that in this manner I could communicate to the Royal Society whatever result I wished?

FILONUS.- I operate on the premise that you are an honest man and that your previous opinion is not influenced by your interests. But in any case, Hylas, tell me, if you are so sure of your previous opinion, why do you want to do this experiment?

---

[47] This last affirmation is a reference to a passage in the Phaedo, Plato's dialogue in which Socrates establishes the world of ideas where the universals lie and of which the objects of the world are imperfect copies. The only thing that does not fit the passage is the ideal type of Hampshire pig, which makes me suspect that this is an apocryphal addition.

[48] This attribution is doubtful; I will go into detail later. Bayes was an obscure clergyman who did not have any mathematical work published in his lifetime. He was a member of the Royal Society owed to some studies on metaphysics. We do not have the original manuscript of his article and we must confide in his friend, R. Price, who presented the publication to the Royal Society attributing it to Bayes.

[49] This is an expression of the Bayes Theorem that seems surprising coming from Filonus, since this theorem was published posthumously in 1780. However, Stigler (1983) has pointed out that the theorem in question was already included in a book published in 1750, in which the author indicates that it is borrowed from his friend. Stigler also makes us aware of how improbable it is that this friend was Bayes, which is why we must attribute this theorem to an unknown mathematician. Perhaps the author of this theorem is Filonus himself.

HYLAS.- I do not know what it is you are trying to say.

FILONUS.- What I want to say is that if your previous opinion has a high probability, there is no reason for you to carry out an experiment, for the only thing you would get out of it would be to reduce your property without your knowledge growing or being any different from that which you had before beginning. Only when your opinion *a priori* is uncertain is when it is worth looking for evidence in Nature that corroborates or disproves the ideas that roam around in your head.

HYLAS.- It might be that I am interested in confirming something that I am already very sure of.

FILONUS.- You contradict yourself, Hylas; if you are already very sure about it then no experiment on earth can take that security away from you, and if there is an experiment capable of it, then your previous opinion is not as firm as you tried to convince me it was.

HYLAS.- I accept what it is you are telling me, but I still have another objection. If I should only confirm my opinion when it is indistinct, then what do I get out of that opinion? What is the use of me having a theory that unites previous opinions to facts, if I am later only going to use it if the previous opinion lacks in value?

FILONUS.- You are exaggerating, Hylas. It is one thing that previous opinion dominates and another that it cannot be combined with the one that the facts give you, for what it is worth. In any case, if this makes you feel uncomfortable, I would suggest that you collect as many facts as possible so that the previous opinion does not have any influence on your result.

HYLAS.- What you propose, Filonus, is not consistent with your approaches. For, what advantages will I obtain from your ways of procedure if I then have to ignore my previous opinions? Why all the commotion if I then have to do all that is possible for my previous opinion not to be taken into account? You speak first ([50]) of the superior coherence of your system, of its beauty, of its simplicity; and you then propose that I ignore the principal part on which its beauty it based, that is, the combination of probabilities pre and post experiment.

FILONUS.- I am not telling you to abandon these previous probabilities, I'm only telling you that if you feel insecure about how you should express them and you prefer to trust the information that your facts give you, then you have a way of doing it.

HYLAS.- I confess to you, Filonus, that this way your system looses a good part of its charm. I then do not know what I am to gain, if in the end the facts are going to be the ones who drag me out of my ignorance.

FILONUS.- Facts are always the ones to drag us out of our ignorance, whether they be facts that we obtain or facts that were obtained by others before us. Our opinion *a priori* should not be partial, unfounded or moved by interest, but rather it should be such that various wise men could embrace it, within the uncertainty with which we must necessarily express it.

HYLAS.- And isn't it better to be dragged out of your ignorance by the facts of your experiment than by that vague idea of probability *a priori,* which appears to be nothing more than the blurry outline of an intangible spirit?

[50] This passage in which Filonus speaks of the coherence of his system has been lost.

FILONUS.- I insist, Hylas, that probability *a priori* is necessary to construct the inference. You cannot say that you choose the most probable hypothesis without even speaking about it. This is not a perversion of my system but rather an obligation of the laws of probability.

HYLAS.- I can choose the hypothesis which makes my facts more probable, so I do not need to talk about probabilities *a priori.*

FILONUS.- I do not understand what it is you mean by this.

HYLAS.- What I'm trying to say is that among two or more alternative hypotheses, I would choose the one that made it more probable for me to obtain the facts that I had already obtained[51].

FILONUS.- And what are the advantages you see in this procedure? Why do you think it is more reasonable to choose the hypothesis that makes the facts you are going to obtain more probable?

HYLAS.- I see it as intuitively reasonable. It doesn't seem adequate to sustain it was highly improbable that the facts I obtained were going to appear.

FILONUS.- Hylas, you are not choosing the hypothesis that makes your sample more probable; you are choosing the hypothesis that if *it were right,* it would give you the data you obtained with the maximum probability, which is not the same thing.

HYLAS.- I do not see the importance of such a nuance.

FILONUS.- I shall set an º. Suppose you want to estimate men's stature in Scotland, and for this you take the height of just one Scotsman, and his height is 1.60 metres[52].

HYLAS.- I think your sample is a little lacking, but let us continue. I'd say that the hypothesis that would make that fact the most probable would be that the average height of the Scottish was 1.60. If it were like that, then to obtain this data would be the most probable fact. If the Scottish were dwarfs whose height only reached a metre, obtaining a Scotsman of 1.60 metres would be highly improbable, and I don't know why I should sustain that I obtained a highly improbable fact if the way I moved was by choosing a man at random.

FILONUS.- I want to make you realise that you have once again said *if it were like that,* which means that in reality you don't know if it is that value which makes your fact more probable or not.

HYLAS.- I am conscious of this limitation, but even so, I do not understand why I should sustain that the average height of Scottish men is two metres, or a metre-fifty.

FILONUS.- Hylas, the hypothesis that makes your only fact the most probable is that ALL Scottish are 1.60 metres tall. This and no other is the hypothesis that makes your sample more probable, for sustaining that there are Scottish with a height different from 1.60 makes it more improbable to find a Scottish man of that stature than if the totality of them are this height.

HYLAS.- Are you joking, Filonus? Saying that the Scottish are not men but rather machines designed by some evil spirit, enemy of delicacy in the arts and the perfection in

[51] Hylas proposes deciding on the basis of what we call today the "Bayes factor".

[52] I translate feet into metres as I did pounds into kilos previously.

music? [53] How can you expect me to sustain something so absurd as that every Scotsman is the same height as his neighbour?

FILONUS.- And how do you know that the Scottish are men, and that their stature should vary, and that it is not to be expected to find in Nature the uniformity that can't be found in other places? You are not, by any chance, using any information *a priori?*

HYLAS.- And even if it is so, which it obviously is, it only intervenes in my inference by making me lay aside any absurd hypotheses.

FILONUS.- Making you discard them, or giving them zero *a priori* probability, which is the same thing. It is an extreme example, but it can make you realise that only by choosing the hypothesis that, if it were right, makes your facts more probable, is sufficient for inference, and only when this hypothesis is considered by probability *a priori* can you choose what you really desire; that is, the most probable hypothesis given the facts you have.

HYLAS.- What you say is all right for choosing among conflicting hypotheses, but only once we have established the frame we move within, this meaning that the Scottish are people and that people's height varies, and it is as likely to find Scottish people that are taller and others that are shorter than the average, I think I can choose for the height of the Scottish the value that makes it more probable that I found my facts and not others.

FILONUS.- You mean the value that, *in the case that this value is the true one,* would make it more likely for you to find the data you obtained[54].

HYLAS.- Yes, Filonus, that is what I want to say; you have made me realise this on various occasions.

FILONUS.- But then you find yourself in the same situation. Imagine that the Scottish man you found was two metres tall, would it be reasonable to assume they are giants?

HYLAS.- It is very improbable that I were to find a value like that picking a Scottish man at random.

FILONUS.- That is precisely what I want to make you realise. Your system has the defect of it only making your sample more probable *in the case of* the true value being the one you obtained. To avoid the terrible starting point of your inference, *"in the case of",* there is a need to weigh that *"case"* for its probability, and only then will we have a sensible inference.

HYLAS.- I would agree if, in fact, we were weighing each case for its probability, but this is not your procedure, that probability which you are considering is not really the probability in each case but rather your previous opinion. I do not want my experiment to be influenced by my prejudgements to such an extreme. I am willing to ignore from the outset how tall I expect the Scottish to be, and for me whatever value that can be imagined has in principle the same probability.

FILONUS.- I do not find that way of acting intelligent, but even so I could still tell you that thanks to that constant probability that you have marked *a priori*, you can talk about how probable it is that the Scottish are or are not two metres tall on the average.

---

[53] We infer that Hylas is English.
[54] This is what today is called the "method of maximum likelihood".

HYLAS.- I do not want absurd conclusions, Filonus. I would say that in this case a single Scottish man is not enough for my purposes and I need to record a larger number so as to not depend on my prejudgement.

FILONUS.- That is why earlier on I was telling you that if you were so worried about not depending on your previous opinions, you should get hold of as many facts as possible so the previous opinion does not have any influence on the result.

HYLAS.- And if I have to ask you once again what it is I have to gain by considering these previous opinions, when precisely what I am trying to do is to get rid of them?

FILONUS.- Well, even though you lose all previous opinions because they are your facts and not those of others which are going to govern your results, you gain a lot, Hylas, a lot more than you suspect. Now you can talk about what can be the probability of the piglets' weight being this or that value, of what probability they have of being bigger than the piglets of any other breed; you can say that your hypothesis A is twice more probable than your hypothesis B but five times less probable than hypothesis C; in conclusion, you can express the uncertainty by means of probability and not by means of dark mechanisms that call on the place occupied by ghostly samples that you never took and will never take.

HYLAS.- I understand the advantages your method offers, but I still have some doubts. What if the correct hypothesis were not to be found among the ones I am now testing?

FILONUS.- Hylas, we do not know which one is correct, nor can we know.

HYLAS.- I understand you, but how can I assign a probability to other hypotheses when the most highly probable one is not found among them?

FILONUS.- Because the probabilities that you are comparing are relative. You express your belief that one is more probable than the other, but your results do not make any reference to other possible hypotheses that you have not yet considered. I see that your tendency to imagine what you don't do; to take imaginary samples, I'd say, makes you think that there are an infinite amount of possible hypotheses that you should consider. No, Hylas, you can only speak about what you are actually speaking about, not about something you are not. Your comparisons always refer to the hypotheses which concern you at that very moment, and the probability they have refer to how much more probable one of them is with respect to the other.

HYLAS.- But if I were to consider a new hypothesis, then the probabilities of the rest would be modified, isn't that right?

FILONUS.- That is so, although I do not know why this disturbs you.

HYLAS.- The reason is obvious, Filonus, how is it possible that a hypothesis has a probability of 60% , and that it turns into 40% once the other alternatives are considered?

FILONUS.- Once again you believe, Hylas, that probability is something external, belonging to things, and not, as we had already agreed, the expression of your opinion on this or that hypothesis. When you consider new alternatives, it is logical that your degree of belief regarding the truth of one or another hypothesis is modified.

HYLAS.- But then I can't establish my absolute probability, even though it is subjective, of any hypothesis. When I say that its probability is 60% it is only an illusion, for in reality my

beliefs would be modified if I had sufficient information, and I would choose another hypothesis more appropriate to the state of things.

FILONUS.- I don't understand what you mean by absolute probability, nor about not having enough information. That is not the way we behave when we want to discover the state of Nature. If we had such sufficient information, I can't see the reason for doing experiments.

HYLAS.- You do not convince me, Filonus, but attribute that to the limits that God imposed on my knowledge or to the difficulty with which things new to us settle on our spirit. I do not wish to carry on with this problem because I want you to give me an answer to yet another difficulty I have seen in your doctrine. What happens when we have a complete lack of previous information? If I understand correctly, whether it be a little or a lot, it is essential if we wish to talk about probability.

FILONUS.- You put me in an awkward position, Hylas, since I do not know when you have heard of absolute ignorance with respect to the laws of Nature. Quite a novel experiment this is! In that case I would recommend that you say that all previous alternatives have the same probability.

HYLAS.- It doesn't seem to me, Filonus, as if it is the same thing to say that I do not know something as to say that all the alternatives have the same previous probability. If I have a bag in which my servant has put black and white balls, not knowing in what proportion, I do not believe that taking out a white ball or taking out a black ball has the same probability. That will depend on what the true proportion is.

FILONUS.- You are not talking about what the real proportion is, but rather about your opinion *a priori.*

HYLAS.- Even so I do not see why my opinion *a priori* should be the one that considers all probabilities to be equal. This procedure looks more to be indifference than ignorance. ([55]).

FILONUS.- In that case try various previous opinions, and if you have sufficient facts, then they will not affect the final result, which means you rid yourself of the problem because you know then that the opinions *a priori* are not going to influence your declarations.

HYLAS.- I am still left with one doubt, Filonus. When I want to discover various things at the same time (for example, what the value of this or that character is, what the relation between the two is) and I apply this to many characteristics of an animal, how shall I be able to express my previous opinion? My brain is insufficient for imagining how much a piglet could weigh if the mother had a small or large litter and at the same time they were fed in a special way, one way or another, and examining all the variants and relationships among characters, giving my opinions for each case. This is too greater a demand on me and I am afraid that it would easily result in contradictions.

FILONUS.- I understand you Hylas, and in this case I cannot think of any other means except trying various opinions once again and waiting for the facts to make them unnecessary.

[55] Hylas is not far off. Although in the literature we see "non-informative priors" referring to those in which the alternatives have the same probabilities (flat priors in the continuous case), all of them are informative. Ignorance cannot be expressed by means of probabilities.

HYLAS.- But in this case I find myself incapable of expressing those opinions, Filonus!. It is not that I refuse to compare various previous opinions, it is that I do not how to express them with exactitude..

FILONUS.- This is a difficult problem, and I can only recommend you do what you can starting from different bases. For example, on one hand you can sustain that all the values have the same probability *a priori*, and on the other you can establish separately the previous opinions for each character, as if they were unrelated. You can sustain that the characters are related in a somewhat vague and more or less arbitrary way. If the results in the end are the same ones, it is difficult to sustain that the previous opinion had any value, for you started with different bases to construct each opinion *a priori*.

HYLAS.- Once again I feel your theory loses its charm, Filonus. Although I admit, and this is how I mean it, that is not just a small advantage that we can still operate within this world of probabilities to express uncertainty.

FILONUS.- This is also how I see it. And let me show my satisfaction for having persuaded you of the great compensation we obtain, even though we pay for it with lack of precision in our procedure.

HYLAS.- So then it seems like God did not create a perfect world in which we could decide without erring in how to behave. But I feel satisfied with this, Filonus, so let us retire now, for the sun is about to set and we are still quite a walk from our homes.

THIRD DIALOGUE

FILONUS.- Good morning, Hylas. I hope you are feeling better and that you can enjoy this clean daybreak, so unusual in the hills of Cotswold, and the beauty of our cathedral ([56]) lighted by the first rays of the sun whose loyalty we have so much doubted in these last few days.

HYLAS.- Good morning, Filonus. I have brought something positive with me taken from our conversations, and it is the thankfulness to God for allowing the sun to shine as we expected, in spite of our doubts regarding it. Although I must confess that it worries me to not be able to know the nature of things as something certain, but rather as something probable. It is as if reality were seen straining through a mist, or as if we were blind and needed a guide in order to situate ourselves in the world that surrounds us.

FILONUS.- But you must think, Hylas, that that probability is continually nearing certainty. Once you have carried out your experiment, your state of beliefs is new and different from the one you had before starting the task. You have less important doubts and your knowledge is more precise. And that is the state of opinion that you or others shall use in future experiments, if they feel the need to better specify their uncertainty.

HYLAS.- You mean that with time, I and others more fortunate than myself will improve our knowledge and extinguish our ignorance little by little. If I understand you correctly, after

[56] He is referring to the Lincoln Cathedral, a magnificent example of English flamboyant Gothic.

each experiment I can make the state of my beliefs more accurate, and whenever I want to get closer to the truth I will make use of that new state of beliefs as *'a priori'* as in the process of inquiry, with which, God willing, we shall be nearer the truth and the knowledge of the laws that govern Nature.

FILONUS.- That is the way I see it, and that is the way we can acquire general knowledge from the particular facts we examine.

HYLAS.- If you are right, you would have resolved the problem of induction, which has preoccupied so many philosophers for so many centuries.

FILONUS.- Well, now that we have resolved the problem of induction, let us take pleasure today in delights nearer our senses and let us save our lucubration for moments in which melancholy dominates our spirit. There is time for everything, and good days not as many, so I propose, since you have already pacified your understanding, that we now excite what least depends on it.

HYLAS.- You are right, but for me to fully enjoy the pleasures that you are suggesting, the premise that my understanding is pacified must be an accurate one.

FILONUS.- And is it not so? Or do you feel new doubts with respect to all the things that we have so much discussed? Do you for any reason think you can be an impartial judge of what you experience?

HYLAS.- Have no fear, Filonus, for all of that is clear to me. What now worries me is the way in which we have resolved the problem of induction. We sustain that we can know the probability of an event, or that a variable takes on a value, given on one hand the facts obtained in the experiment and on the other the probability derived from our previous experience.

FILONUS.- That is right, and even though this probability is subjective, when the number of facts is large enough, your previous opinion must be irrelevant.

HYLAS.- But there is still something that doesn't quite fit in the process. What would happen if I had obtained the sample incorrectly? What if the facts did not reflect the state of the population correctly? What would happen if they were not distributed in the way I presumed? What would happen if my previous opinion, shared by many other scientists, were to differ so greatly from the truth because of my prejudgements, so common in men, that not even with a reasonably high number of facts could I offset the mistaken opinion given by the experts?

FILONUS.- Look, Hylas, you are letting yourself slide down the hill that leads to scepticism, and from there to solipsism and later to melancholy there is a short stretch.

HYLAS.- Should I, then fool myself and say that I understand what I do not, and admit that I find reasonable what seems to me has faulty foundations?

FILONUS.- No, no you should not, but neither should you ignore the limits of human understanding. Probabilities *a priori* do not exist alone, floating on the waves created by seas of opinions. All probability is based on certain hypotheses, and so is this previous probability. In reality we should call it 'probability *a priori* given a series of hypotheses', for only with regard to them is this probability feasible.

HYLAS.- But then the result of our inference is not a probability 'given experimental evidence', but a probability 'given experimental evidence and *a group of attached hypotheses'.*

FILONUS.- That is the way I understand it.

HYLAS.- And how do you then say you have resolved the problem of induction, if you have left it as intact as it was before starting your lucubration? It has maintained its virginity if I am correct! How can you assert that Science already has a firm base if it all depends of the truth of previous hypotheses, and this is a truth that you are unaware of? How can you justify that your knowledge progresses if its foundation is as fragile as the clay of the idol's feet? Or is it that you have some means of estimating the certainty of the hypotheses that accompany all your reasoning?

FILONUS.- Calm down, Hylas, for your own health comes before all else in the world. I do not know how certain those hypotheses may be. However, if I suspected that the sample was not taken properly or that the worker who weighed the pigs mixed up the facts, or did it badly out of spite because he had an unsettled debt with me, then in such cases, I wouldn't give credit to my results. If I did, it would be because I was convinced that everything was done in an honest and reasonable way. And science progresses that way, because I can choose between rival theories even though they are conditioned by the veracity of those hypotheses, for in the end the experimental evidence will make me lean towards one theory or the other.

HYLAS.- They seem to me like poor means for reaching such high objectives. If science should progress on the basis of good faith, of the absence of unavoidable error, of the correct interpretation of Nature, and of all that your hypotheses contain, I wouldn't tell you that I expect quick and constant progress, but rather ups and downs, false starts and errors that should remain unsolved because we trust in the fact that we did things correctly.

FILONUS.- Having a way of discerning which theory we prefer is not a poor result, although it is true that science is always conditioned by the veracity of those hypotheses. In reality the sword of Damocles is always present, providing us with mistaken hypotheses, but once we admit that they are the most reasonable we can find, we can activate our decision mechanism make our estimations about all we find in Nature more precise, say how probable it is for one character to have this or that value, and prefer this theory to that one based on how much more probable this one is to its alternatives.

HYLAS.- In how much more probable *you think it is,* you mean.

FILONUS.- Of course Hylas, we have already agreed that probability is nothing other than my state of beliefs.

HYLAS.- That's it, but I have the impression that it is so much our custom to believe that probability is something objective and that it is outside us that upon talking about how probable this or that hypothesis is, we will end up believing that we are truly talking about objective probability and not about the state of our opinion.

FILONUS.- And what do you suggest? That scientists send their declarations to the Royal Society using the phrase 'my probability'? It would be annoying and it would emphasise the author's ego, something that a man of science should always avoid.

HYLAS.- Yes, but it wouldn't give the impression of a detached objectivity.

FILONUS.- Nor does it intend to. Insisting upon subjectivity would also give the impression of something arbitrary being done, and we have already agreed that subjectivity does not in any way imply arbitrariness if the scientist isn't mad. And this without taking into account that most of the time the facts dominate that subjective previous opinion that so disgusts you.

HYLAS.- I admit that science can progress, although I doubt that progress can be so rapid and constant as the application of your principle seems to suggest , but we have left the induction principle untouched. In reality no experimental evidence can guarantee us proximity to the truth, for we depend on other hypotheses which we do not know to be true.

FILONUS.- But you won't deny that we have established a reasonable behaviour that can be followed by all those men that are interested in learning about the laws of Nature.

HYLAS.- A behaviour that will produce good results because God won't allow it to be any other way, but not because it intrinsically guarantees an approximation to the truth.

FILONUS.- Hylas, you are adducing a comparison that could be quite fortunate. Ethics without God are based on the behaviour that is most beneficial for the individual and the community. In the same way, scientific progress without God would be based on the premise that a scientist's behaviour will give results which we consider, under certain hypotheses, to be more probable and better adjusted to what we already know. In both cases the individual's behaviour is determined on uncertain, but experimental, bases.

HYLAS.- I do not understand the comparison, Filonus, and I think it is completely forced. With regard to ethics, I don't know how you fuse the individual's and society's benefit, when they so often come into conflict; and with regard to science, I can't see what relation there is between the behaviour of the scientist and what it is I am asking you: how can you guarantee that experimental evidence will lead you nearer to the truth?

FILONUS.- Hylas, If you are so extreme then I shall say no, no I cannot assure that experimental evidence will lead me nearer to the truth; nor can I assure that you are here now and that I am not speaking to a ghost or I am immersed in a dream without knowing it. I do not even know if I can talk appropriately like myself, since I only have experimental evidence about myself!

HYLAS.- I beg you not to feel offended, Filonus, my questions arise from my perplexity and not from my desire to ridicule your answers. I wanted certainty and I exchanged it with you for probability, and now I don't even have the consolation of reaching probable knowledge. How happy the monks must be, not having to doubt anything and trusting in God to govern and resolve the course of their lives! Is it that blind faith that definitively guides our belief that we know something?

FILONUS.- Hylas, to begin to discuss something you must at least first assume that your interlocutor is truly present. To add hypotheses to our experiments, taking for granted they hadn't been sabotaged, the samples were taken at random and no neglected risk was introduced, does not seem to me to be paying a high price for the conclusions at which we will arrive. And if the result permits you to act upon Nature and benefit from it, if you can build bridges that don't collapse, if your crops yield more, if you can cure the ill, all of this indicates that you cannot be too mistaken.

HYLAS.- Yes, but ordinary practice does not require speculative knowledge([57]). You can believe that you are ill because bad humours take over your blood, and that you shall recover by taking certain herbs; and if you take the right ones, health shall return to your body, even though the causes of your illness might not have been the ones you suspected. Using theories with scarce foundation because they work well is like making children believe that the bogeyman exists just so they will eat their soup. An engineer might not need anything more than rules he can keep in his memory, but a scientist is obliged to think in a different way and ask himself the reasons for things he infers and for the decisions he takes.

FILONUS.- I do not know if this world is made for you, Hylas, but I think we have good reasons for behaving the way we do and believing in all that we believe.

HYLAS.- And what is the nature of those good reasons?

FILONUS.- I couldn't answer you now, Hylas. The happiness with which I began the day has vanished, and I see that new matters must be submitted to our reflection.

HYLAS.- Then let us go down to the city, and we shall continue to deal with these things another day. Let us free our bodies from their tensions since we cannot do the same for our spirit. It has been said that the thinker should be frugal, and that the intensity of thought always made the great geniuses forget their sustenance. I am not one of them and I do not encourage you to be either.

FILONUS.- Let us go then, Hylas, and leave the discussion for tomorrow.

---

[57] Hylas is referring to a commentary recorded in the third of the dialogues published by Bishop Berkeley: *"Hence ordinary people persist in their errors and, in spite of this, they manage to function in everyday life."*