CHAPTER 1

DO WE UNDERSTAND CLASSIC 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. Definitions
   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. Shrinkage of random effects estimates
   1.5.3. Bias, variance and Risk of an estimator when the effect is fixed or random
   1.5.4. 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
Appendix 1.4
1.1. Historical introduction

The Bayesian School was, in practice, founded by the French aristocrat and
politician Marquis Pierre Simon Laplace via several works published from 1774
to 1812, and it had a preponderant role in scientific inference during the
nineteenth century (Hald, 1998). A few 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 had never published a mathematical paper in 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, Bayesian techniques were commonly
used along the 19th and the first few decades of the 20th century. For example,
the first deduction of the least square method made by Gauss in 1795 (although
it was published in 1809) was made using Bayesian theory. At that time these
techniques were known as “inverse probability”, because their objective was to
estimate the parameters from the data (i.e. to find the probability of the causes
from their effects); direct probability takes place when the probability distribution
originating the data is known and the probability of a sample is calculated from it
(throwing a dice, for example). 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. Given that Laplace ignored the work
of Bayes, the correct name for this school should be “Laplacian”, or perhaps we
should preserve the old and explicit name of “inverse probability”. Nevertheless,
as it is commonplace today, we will use the name “Bayesian” throughout this
book.

Bayesian statistics uses probability to express the uncertainty about the
unknowns that are being estimated. The use of probability is more efficient than
any other method of expressing uncertainty (Ramsay, 1926). Unfortunately, to
make it possible, inverse probability ‘needs’ the knowledge of some prior
information. For example, we know through published literature that in many
countries the pig breed Landrace has a litter size of around 10 piglets. We then perform an experiment to learn about Spanish Landrace litter size, and a sample of five sows is evaluated for litter size in their first parity. Let us say an average of six born piglets is observed, this seems a very unlikely outcome a priori, and we should not put much trust in our sample. However, it is not entirely 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 9. Laplace became aware of this problem, and in his later works, he examined the possibility of making inferences based in the distribution of the samples rather than in the probability of the unknowns, founding in practice the frequentist school (Hald, 1998).

Fisher’s work on the likelihood distribution in the 20’s and the frequentist work of Neyman and Pearson in the 30’s eclipsed Bayesian inference. The reason being that they offered inferences and measured uncertainty about these inferences without the need of prior information. Fisher developed the properties of the method of maximum likelihood, a method attributed to him, although Daniel Bernouilli proposed it as early as 1778 (Kendall, 1961) and Johan Heinrich Lambert in 1760 (Hald, 1998). When Fisher discussed this method (Fisher, 1935) one of the discussants of his paper noticed that the statistician and economist of Irish-Catalan origin Isidro Francis Edgeworth had proposed it in 1908. It is likely that Fisher did not know of this article when he proposed to use the likelihood in an obscure paper published in 1912 (1), when he was 22 years old, but it is remarkable that Fisher never cited in his life any precedent of his work about likelihood. 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 and an alternative, a procedure that Fisher disliked. Fisher considered that when a null hypothesis is not rejected we

---

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”.
cannot assume straight away that this hypothesis is true, but instead just take this hypothesis as provisional (Fisher 1925, 1935), very much in the same sense as the Popper theory of refutation (Popper, 1934). It is interesting to note that Popper’s famous theory of conjectures and refutations was published independently and much later than Fisher’s one.

Although some researchers still used Bayesian methods in the 30’s, like the geologist and statistician Harold Jeffreys did, the classical Fisher-Neyman-Pearson school dominated the statistical world until the 60’s, when a ‘revival’ started that 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 and independently to consider probability as a “belief”. Prior information should then be evaluated by experts and the probabilities assigned to different events according to the experts’ opinion. This procedure can work for a few traits or effects, but it has serious problems in the multivariate case. The other theoretical problem is how to represent “ignorance” when there is no prior information or when 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, with the exception of very simple inferences, this leads to multiple integrals that cannot be solved even using approximate methods. This supposed a big problem in order to apply Bayesian techniques until the 90’s, when a numerical method was applied in order to find practical solutions to all these integrals. The method, called Monte Carlo Markov Chains (MCMC) enabled the integrals to be solved, 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 set of random sample numbers extracted from a probability density distribution, instead of using the
mathematical expression of this distribution. The first use of random samples from a density distribution was proposed by the Guiness brewer William Searly Gosset (1908), called “Student”, in his famous paper in which he presented the t-distribution. 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. Much later, Geman and Geman (1986) applied this method to image an analysis using a particularly efficient type of MCMC that they called “Gibbs sampling” because they were using Gibbs distributions. 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; Sharon McGrayne has written a lively and entertaining history of this procedure (McGrayne, 2011). Today, the association of Bayesian inference and MCMC techniques has produced a dramatic development of the application of Bayesian methods 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 towards some treatment; for example, we are testing whether a selected population for growth rate has a higher growth rate than a control population. What we wish to find is the probability of the selected population being higher than the control one. However, classic statistics does not provide an answer to this question; classical statistics cannot give the probability of a treatment being higher than other treatment, which is rather frustrating. The classical procedure is to start with the hypothesis, called ‘null hypothesis H₀’, that there is no difference between treatments; i.e., that the difference between the means of the selected and control populations \( m_1 - m_2 = 0 \). By repeating the
experiment an infinite number of times, we would obtain an infinite number of samples, and we could calculate the difference $\bar{x}_1 - \bar{x}_2$ for each repetition between the averages of the samples. *If the null hypothesis is true*, these differences will be grouped around zero (Figure 1.1). Notice that although there is no difference between selected and control populations, the difference between our *samples* will never be exactly zero, and by chance, it can be high. Let us consider the 5% of the highest differences (shadow area in Figure 1.1).

**Figure 1.1.** Distribution of the difference between the averages of repeated samples $\bar{x}_1 - \bar{x}_2$, if $H_0$ is true and there is no difference between treatments ($m_1 - m_2 = 0$). 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.

We would actually take only one sample. If our sample lied in the shadow area of figure 1, we could say that:

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

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

Neyman and Pearson (1933) suggested that the “scientific behaviour” should be to take option 2, acting as if the null hypothesis H₀ was wrong. A result of this behaviour would be that ‘in the long run’ we would be right in almost 95% of the cases. Notice that this is not the probability of the treatments being different, but the probability of finding samples higher than a given value. The relationship between both concepts is not obvious, but simulation experiments in simple situations show that the probability of the treatments being different is substantially lower than what its 95% level of rejection seems to provide (Berger and Sellke, 1987; Johnson 2013).

We have stated, before making our experiment, the frequency in which we will say that there are differences between treatments (when actually there are not), using a conventional value of 5%. This is called ‘Type I error’. 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 either say that we do not know whether the two treatments are different (²), or we can accept that both treatments 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 the possibility of being wrong by saying that there is no difference between treatments when actually this difference exists (they called it ‘Type II error’).

Often a ‘P-value’ accompanies the result of the test. P-value is the probability of obtaining a difference between samples equal or higher than the actual difference

² This attitude to scientific progress was later exposed in a less technical way by Karl Popper (1934). 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 (1925, 1935) based his testing hypothesis theory in the same principle. I do not 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).
found when there is no difference between populations (figure 1.2)(3). They were proposed by Fisher as an exploratory analysis to examine how much evidence we can get from our samples (4), but notice that the P-value is not the probability of both treatments being different, but the probability of finding samples of the difference between treatments higher than ours.

![Diagram of a normal distribution with a P-value of 3%]

**Figure 1.2.** A *P-value* of 3% gives the probability of finding the current sample value or a higher value if the null hypothesis holds.

A very low *P-value* gives more evidence of treatments being different than a higher *P-value*, but, as before, it is not clear how much evidence, since they do not measure the probability of the populations being different. In other words, a *P-value* of 2% does not give twice as much evidence as a *P-value* of 4%. Moreover, the result of the test is established with the same Type I error, normally a 5%, independently on the *P-value* we obtain, because we define the Type I error before making the experiment, and the *P-value* is obtained after the experiment is made. It is important to realise that the *P-value* changes if we repeat the experiment. Modern statisticians use *P-values* to express the amount

---

3 Here we use a one-tail test for simplicity.

4 Neyman and Pearson never used *P-values* because they are not needed for accepting or rejecting hypothesis. 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. Often they create more confusion than help in understanding results.
of evidence the sample gives, but there is still a considerable amount of discussion about how is this “amount” measured, and no standard methods have been hitherto implemented (see Sellke et al. 2001, Bayarri and Berger 2004, Johnson 2013 for a discussion).

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 can 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. For example, you may accept an error level of a 5% and find along your career that your data was always distributed far away from the limit of the rejection (figure 1.3).

![Figure 1.3. 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 a much higher evidence about the null hypothesis being wrong](image)

P-value is the probability of not having differences between treatments. This is not true. The P-value gives the probability of finding the current sample value
or a higher value, but we are not interested in how probable it is to find our sample value, but in how probable our hypothesis is, and classic statistic does not have an answer for this question. A conscious statistician knows what a P-value means, but the problem is that P-values suggest more evidence to the average researcher than they actually have. Johnson (2013) has showed that, when testing common null hypothesis, under rather general conditions, about one quarter of significant findings are false. This means that people interpreting P-values as the probability of the null hypothesis being true are wrong and fare away about the real evidence of this. Johnson (2013) recommends the use of P-values of 0.005 as a new threshold for significance.

**P-value is a measure of “significance”:** This is not true. 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.

Notice that if a P-value is small enough for establishing significant differences, if we repeat the experiment, the new P-value will not necessarily be that small. For example, consider that the true value is placed at the 5% level (Figure 1.4). Then take a sample, and the different between treatments obtained is also placed at the 5% level (this is not a rare supposition, since the samples should be distributed near the true value) and establish that the difference between treatments is significant.

However, when repeating the experiment, half of the samples will give significant differences between treatments (P<0.05) and half of them will not (P>0.05). Thus, a P-value of 5% gives the impression of having more evidence than what we actually have (5). 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 not.

---

5 Notice that this argument is independent of the power of the test. It applies whatever this power is.
Significant difference means that a difference exists. This is not always true. We may be wrong once every twenty times on average, if the error level is 5%. The problem is that when measuring many traits, we may detect a false significant difference once every twenty traits (7). The same problem arises when we are estimating many effects. It is not infrequent to see pathetic efforts of some authors trying to justify 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 purely by chance.

N.S. (non-significant difference) means that there is no difference between treatments. This is usually false. First, in agriculture and biology, treatments are generally different because they are not going to be exactly equal. Two pig breeds can differ in growth rate in less than a gram, but this is obviously irrelevant. Secondly, in well-designed experiments, N.S. appears when the difference between treatments is irrelevant, but this only happens for the trait for which the

---

6 This example is shown for one tail test, but the same applies to two tails tests.

7 Once each twenty traits as a maximum if the traits are uncorrelated. If they are correlated, the frequency of detecting false significances is different.
experiment was designed, thus other traits can have relevant differences between treatments but we obtain N.S. from our specific 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.

**Even if the differences are N.S., we can still observe a ‘tendency’.** This statement is nonsense. If we do not find significant differences between treatments A and B this means that now A is higher than B, but after repeating the experiment, B can be higher than A. It is rather unfortunate that referees admit expressions like this, even in competent scientific journals. Moreover, it often happens that the N.S. differences are high, nothing that can be described as a ‘tendency’; N.S. describes my state of ignorance, not the size of the effect.

**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 a difference be in order to consider it as relevant, should be defined before making the experiment. A relevant value is a quantity 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 relevant. The problem then is that experiments are designed for one trait but many other traits are often measured, thus significant differences will not be linked to relevant differences for these other traits.

**Significant difference means ‘Relevant difference’:** This is often false. In well-designed experiments, a significant difference will appear just when this difference is relevant. Thus, if we consider before performing the experiment that 100 g/d is a relevant difference between two treatments, we will calculate the size of our experiment in order to find a significant difference when the difference from the averages of our samples $|\bar{x}_1 - \bar{x}_2|$ ≥ 100 g/d, and we will not find a significant difference if it is lower than this. The problem arises when we analyse a trait other than the one used for defining the size of the experiment; but also in field data,
where no experimental design has been made, or in poorly designed experiments. In these cases, there is no link between the *relevance* of the difference and its *significance*. In these cases, we can find:

1) **Significant differences that are completely irrelevant**: This case is innocuous; however, if *significance* is confused with *relevance*, the author of the paper will stress this result without reason, since the difference found is irrelevant. *We will always get significant differences if the sample is big enough.* If the sample is high, the mean distribution of the samples, when repeating an experiment many times, will have a lower dispersion. For example, the average of milk production of the daughters of a sire with only two daughters can be 4,000 kg or 15,000 kg, but the average of sires with 100 daughters will be much closer to the mean of the population. Figure 1.5 shows a non-significant difference that is significant for a larger sample. Conversely, if we throw away part of our sample, former significant differences become non-significant. Thus, ‘significance’ itself is of little value, it only indicates whether the sample is big or not.

![Distribution of the samples for two sample sizes.](image)

**Figure 1.5.** Distribution of the samples for two sample sizes. “Non-significant differences for n=10 can be significant for n=100.”

---

8 In simulation studies, it is easy to find significant differences by augmenting the number of repeated simulations.
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 “N.S.” relevant differences would mean that perhaps there is an important effect of the treatment or perhaps not, we do not know but it is important to know it, and a higher sample should be analysed.

3) **Non-significant differences that are irrelevant, but have high errors:** Sometimes the estimated difference between treatments can be, by chance, near zero, but if the standard error of it is high, the true difference may be much higher and relevant. This is dangerous, because a small difference accompanied with a ‘N.S.’ seems to be non-important, but it can be rather relevant indeed. 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 small and “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 (we get 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’ saying ‘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 and confidence intervals 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 experts analyse the aniseed flavour of some meat, and they find significant differences of three points in a scale of ten point; 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, we propose some practical solutions to this problem.

**Tests of hypothesis are always needed in experimental research.** 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 NO. However, this is not actually the question for most biological problems. In fact, we know that the answer to this question is generally YES, because two treatments are not going to be *exactly equal*. The test only adds to our previous information that one of the treatments is higher than the other one. However, in most biological problems, our question is whether these treatments differ in more than a relevant quantity. In order to answer this question we should estimate the difference between treatments accompanied by a measurement of our uncertainty. This is more informative than comparing LS-means only, by showing whether they are significantly different, because as we have seen before, significance is not related to relevance but to sample size. There is a large amount of literature recommending to focus more in confidence intervals or other quantitative measurement of uncertainty than in hypothesis tests (see Lecoutre and Poitevineau, 2014, for a recent review on this controversy).

### 1.3. Standard errors and Confidence intervals

#### 1.3.1. Definitions
If we take an infinite number of samples, the sample mean (or the difference between two sample means) will be distributed around the true value we want to estimate, as shown in Figure 1.1. 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, just one, a large standard error means that we do not know how close we are 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 will be close to the true value.

When the distribution obtained by repeating the experiment is Normal (9), twice the standard error around the true value will contain approximately 95% of the samples. This permits the construction of the so-called Confidence Intervals at 95% by establishing the limits within which 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 1.1, and we have to use our estimate instead of the true value to define the limits of the confidence interval. Our confidence interval is approximately (10) (sample average ± 2 s.e.). A consequence of this is that each time we repeat the experiment we have a new sample and thus a new confidence interval. For example, let’s assume we want to estimate the litter size of a pig breed and we obtain a value of 10 piglets 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. What shall we do? In classical statistics we behave as if our interval would be one of the intervals

---

9 Computer programs (like SAS) ask 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).

10 For small samples, a t-distribution should be used instead of a Normal distribution, and the confidence interval is somewhat larger, but this is not important for this discussion.
containing the true value (see figure 1.7). We hope, as a consequence of our behaviour, to be wrong a maximum of 5% of times along our career.

1.3.2. Common misinterpretations

The true value lies between ± s.e. of the estimate: We do not know whether this will happen or not. First, the distribution of the samples when repeating the experiment might not be normal. This is common when estimating correlation coefficients, which are close to 1 or to -1, it is nonsense to write a correlation coefficient as 0.95 ± 0.10. Some techniques (for example, bootstrap), taking advantage of the easy computation with modern computers, can show the actual distribution of a sample. A correlation coefficient sampling distribution may be asymmetric, like in figure 1.6. If we take the most frequent value as our estimate (-0.9), the s.e. has little meaning.

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=100%; i.e., with total certainty. We utter that our interval is one of the “good ones” (Figure 1.7). We may be wrong, but if we behave like
this, we hope to be wrong only 5% of the 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 also from the distribution of samples in ideal repetitions of the experiment.

Figure 1.7. Repeating the experiment many times, 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 too many times along our career.

The true value should be in the centre of the CI, not in the borders: We do not know where the true value is. If the CI for differences in litter size is $[0.1, 3.9]$, the true value may be 0.1, 0.2, 2.0, 3.9 or some other intermediate value between the confidence interval limits. In Figure 1.7 we can see that some intervals have the true value near one side and others near the centre. We expect to have more intervals in which the true value is near the centre, but we do not know whether our interval is one of these.

Conceptual repetition leads to paradoxes: Several paradoxes produced by drawing conclusions not only from our sample but also from conceptual repetitions of it have been noticed. The following one can be found in Berger and Wolpert (1982).
Imagine we are measuring 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 does not work if the pH is higher than six. Although we did not find any measure higher than six, repeating the experiment an infinite number of times, we will obtain a truncated distribution of our samples (Figure 1.8 b). This means that we should change our confidence interval, since all possible samples higher than 6 would not be recorded. 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. However, 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 even though none of our measurements lied 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 also from conceptual repetitions of the experiment, 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 \( \hat{u} \)

\[ e = u - \hat{u} \]

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

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

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

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

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). Nevertheless unbiasedness has not always been considered a particularly attractive property of an estimator; Fisher considered

\[ ^{11} \text{All of this is rather arbitrary and other solutions can be used. For example, we may express the error as a percentage of the true value, the loss function may be the absolute value of the error instead of its square and the risk might be the mode instead of the mean of the loss function, but in this chapter we will use the common definitions.} \]
that the property of unbiasedness was irrelevant due to its lack of invariance to transformations (Yates, 1990), as we will see below.

1.4.2. Common misinterpretations

A transformation of an unbiased estimator leads to another unbiased estimator: This is normally not true. In general, a transformation of an unbiased estimator leads to an estimator that is not unbiased anymore. For example, it is frequent to find unbiased estimators for the variance, and use them for estimating the standard deviation by computing their square root. 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).

Unbiased estimators should be always preferred: Not always. In general, the best estimates are the ones with lower risk. As the risk is the sum of the bias plus the variance of the estimator, it may happen that a biased estimator had a lower risk, being a better estimator than an unbiased estimator (figure 1.9).

Figure 1.9. An example of biased estimator (blue) that is not distributed around the true value ‘m’, 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

$$\hat{\sigma}^2 = \frac{1}{k} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

It can be shown that the bias, variance and risk of this estimator, using the quadratic loss function we defined before,

$$\text{BIAS}(\hat{\sigma}^2) = \sigma^2 - \frac{n-1}{k} \sigma^2$$

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

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

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 $k = n+1$. 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 one of the reasons 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 each sow has four parities. We feed each group with a different diet and we are interested in knowing the effect of the diet in the litter size of the sows. The effect of the diet can be considered as a “fixed” effect, because we are interested in finding the food that leads to higher litter sizes, and we consider that if repeating the experiment the effect of each diet on litter size would be the same. Some sows are more prolific than others are, but we are not interested in the prolificacy of a particular sow, thus if the sows have been assigned randomly to each diet, we consider that each sow effect is a “random” effect. Repeating the experiment, we would have different sows with different sow effects, but these effects would not change our inferences about diets because sows are assigned randomly to each diet.

When repeating an experiment an infinite number of times, a fixed effect has always the same values, whereas a random effect changes in each repetition of the experiment. Repeating our experiment, we would always give the same four diets, but the sows will be different; thus, the effect of the food will be always the same but the effect of the sow will randomly change in each repetition. In Figure 1.10, 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.

Notice that the errors are the difference between true and estimated values in both cases, but in the case of random effects they are not the distance between the estimate and the mean of the estimates, because the true value changes in each repetition of the experiment (12).

---

12 As random values change in each repetition, instead of “estimating” random values we will say “predicting” genetic values. This is a common nomenclature in genetic studies.
1.5.2. *Shrinkage of random effects estimates*

The estimate (the prediction) of a random effect depends on the amount of data used. Let us take a simple model in which we measure the growth of chicken under several diets, with a different number of chickens per diet.

If we estimate the effect of a diet as a fixed effect, the effect of a diet will be the average of the chicken's growth under this diet

$$\hat{u}_f = \frac{1}{n} \sum y_i$$

However, estimating the diet effect as a random effect, the effect of diet will be

$$\hat{u}_r = \frac{1}{n + \frac{\sigma_e^2}{\sigma_u^2}} \sum y_i$$
where $\sigma^2$ is the variance of the residual effects on growth not explained by the diet; i.e.: $y = u + \varepsilon$ then $\sigma_y^2 = \sigma_u^2 + \sigma_\varepsilon^2$. We will see this different way of estimating fixed or random effects in chapter 7.

Notice that when $n$ is high, both estimates are similar, but when $n$ is small the random estimate suffers a “shrinking” and takes a smaller value than when the effect is considered as fixed. The importance of this shrinking will depend on the number of data used for estimating the random effect. This is well known by geneticists, who evaluate animals considering their genetic values as random; an example is developed in Appendix 1.3.

As we will see later, the researcher is sometimes faced to the dilemma of considering an effect as fixed or random, particularly when correcting for noise effects. If some levels of a noise effect have few data, they will be poorly estimated, which could affect the results, but if the noise is considered random, this correction will be small. This is why it is common to consider random an effect with many levels and few data in each level. The other side of the problem is that the corrections actually applied in these cases are very small, thus the decision of taking an effect as random or fixed is not always clear.

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

In the case of fixed effects, since the true value is constant, $u = E(u)$, 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 much less attractive property (13).

---

13 Henderson (1973) has been criticised for calling the property $E(u) = E(\hat{u})$ “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 $u = E(\hat{u}|u)$, a property that BLUP does not have (see Robinson, 1981).
FIXED: $\text{Bias} = E(e) = E(u-\hat{u}) = E(u) - E(\hat{u}) = u - E(\hat{u})$

RANDOM: $\text{Bias} = E(e) = E(u-\hat{u}) = E(u) - E(\hat{u})$

The variances of the errors are also different (see Appendix 1.2 for a demonstration)

FIXED: $\text{var}(e) = \text{var}(u-\hat{u}) = \text{var}(\hat{u})$

RANDOM: $\text{var}(e) = \text{var}(u-\hat{u}) = \text{var}(u) - \text{var}(\hat{u})$

The best estimators are the ones with the lowest risk; in the case of unbiased estimators, as BIAS=0, the best ones have the lower error variance. For fixed effects, as the true value ‘u’ is a constant, $\text{var}(u) = 0$, thus the variance of the error is the same as the variance of the estimator $\text{var}(\hat{u})$, and the best unbiased estimators are the ones with smallest variance. In the case of random effects, the true values are not constant and the variance of the error is the difference between the variances of the true and estimated values. Thus, in the case of random effects, the best estimator is the one with a variance as close as possible to the variances of the true values, because this minimizes the variance of the error.

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 will be 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.

1.5.4. Common misinterpretations

**An effect is fixed or random due to its nature:** This is not always 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. Conversely, we might have considered the sow as a “fixed” effect and we could have estimated it, since we had five litters per sow (14). Thus, the effects can be taken as fixed or random depending on our interests.

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 predicting genetic values. Using Mendel’s laws, we know the relationships between relatives, thus we can use this prior information if the individual genetic effects are considered random effects. Appendix 1.3 gives an example of this prediction.

Even for random effects to be unbiased is an important property: The property of unbiasedness is not particularly attractive for random effects, since when repeating the experiment the true values change as well and the estimates are not distributed around the true value. We have seen before that, even for fixed effects, unbiasedness may be considered rather unattractive, since usually it is not invariant to transformations.

BLUP is the best possible predictor of a genetic random value: Animal breeding values are commonly estimated (predicted, as animal breeders call this) by BLUP (Best Linear Unbiased Predictor, Henderson, 1973). The word “Best” is somewhat misleading because it seems that BLUP is the best possible predictor, but we can have biased predictors with lower risk than unbiased ones. The reason for searching predictors only among the unbiased ones is that there are an infinite number of possible biased predictors with the same risk, depending on their bias and their variance. By adding the condition of unbiasedness, we find a single one, called BLUP, which is not the best possible predictor, but the best among the unbiased ones.

---

14 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 Eisenhart’s proposal (Fisher, 1956).
1.6. Likelihood

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, Lambert and Edgeworth as we said in the historical introduction of section 1.1. 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.

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 with mean 1.5 kg, or from other population with a mean of 1.8 kg or from many other possible populations. Figure 1.11 shows the density functions of several possible populations from which this rabbit can come from, 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 of \( f(y_0|m_2) \). It looks very unlikely that a rabbit of 1.6 kg comes from a population with mean 1.8 kg. Therefore, it seems more likely that the rabbit comes from the second population.
**Figure 1.11.** Three likelihoods for the sample $y_0 = 1.6$. **a:** likelihood if the true mean of the population is 1.5, **b:** likelihood if the true mean of the population is 1.6. **c:** likelihood if the true mean of the population is 1.8

All the values $f(y_0|m_1)$, $f(y_0|m_2)$, $f(y_0|m_3)$, … are called “likelihoods” and show how “likely” we would have obtained our sample $y_0$ if the true value of the mean would have been $m_1$, $m_2$, $m_3$, … (figure 1.12).

**Figure 1.12.** Three likelihoods for the sample $y_0 = 1.6$

These likelihoods can be represented for each value of the means. They define a curve with a maximum in $f(y_0|m_2)$ (Figure 1.13). 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.
Figure 1.13. Likelihood curve. It is not a probability because its values come from different probability distributions, but it is a rational degree of belief. The notation stress that the variable (in red) is m and not $y_0$, which is a given fixed sample.

We have a problem of notation here, because here the variable is ‘m’ instead of ‘y’. Speaking about a set of density functions $f(y_0|m_1), f(y_0|m_2), f(y_0|m_3)\ldots$ for a given $y_0$ is the same as speaking about a function $L(m|y_0)$ that is not a density function \(^{15}\). 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:

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

\(^{15}\) Some classic texts of statistics (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. Likelihood can be found in textbooks as $L(m|y), L(y|m), f(y|m)$ and $L(m;y)$.
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.

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, as we have seen, these values belong to different density functions, and the likelihood function defined taking them is not a probability function. Thus, Fisher preferred to use the word likelihood for all the values (16).

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 expressing uncertainty in a similar manner. What Fisher proposed was to use the whole likelihood curve and not only its maximum, a practice rather unusual nowadays (Figure 1.14). Today, frequentist statisticians typically use only the maximum of the curve because it has good properties in repeated sampling. 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. However, all those properties are asymptotic, 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 asymptotic frequentist properties: any reparametrization leads to the same type of estimator (Appendix 1.4). 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 a ML estimator as well. This was a key-property for Fisher.

---

16 Speaking strictly, these quantities are densities of probability. As we will see in 3.3.1, probabilities are areas defined by \( f(y)\Delta y \).
From a practical point of view, the ML estimator is an important tool for the applied researcher. The frequentist school found a list of properties that good estimators should have, but there are no rules about how to find these estimators. 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 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. This allowed Fisher to propose interesting concepts like “amount of information”, that we will see in chapter 10, providing a way for estimating (asymptotically) standard errors. The frequentist school also discovered that the likelihood was useful for constructing hypothesis tests, since the likelihood ratio between the null and the alternative hypothesis has good asymptotical frequentist properties. We will come back to this in chapter 10.

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 own
probability depending on the true value of the distribution from which it comes. For example, if the true value of the population of Figure 9 is the case c (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. 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 (17). As Fisher says:

“We define likelihood as a quantity proportional to the probability that, from a population having that particular value of \( \rho \), a sample having the observed value \( r \), should be obtained”.

Ronald 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. 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 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. We put an example comparing likelihood with probability in chapter 9. There is an obvious risk of confounding likelihood and probability, as people working in QTL should know (18).

---

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

18 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.
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 economic considerations; for example, which difference between treatments justifies doing an investment or preferring one treatment. In classical design of experiments, a “relevant” value of the difference between treatments is defined in order to find a significant difference if this value, or a higher one, is found. 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 enough, we will always find significant differences. We can consider that a relevant difference is a proportion of the variability of the trait. Having 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 pigs from Tribout et al. (2010): dressing percentage has a very small variability; its mean value is 77.6 but its standard deviation only 1.8. If a treatment would increase carcass yield one standard deviation, this would suppose a great increment, although 1.8 is only 2% of the mean. Conversely, if litter size in pigs is 14 (see, for example Lundgren et al., 2010), 2% of the mean is 0.28 piglets, which is rather irrelevant. If we take a list of the important traits in animal production, we will see that the economic relevance appears at a quantity placed between ½ and ⅓ of the standard deviation of the trait for most of them. Therefore, we can postulate a quantity placed between ½ or ¼ of the standard deviation of the trait for all traits in which it is not possible to argue economical or biological reasons. 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. Similar postulates can be proposed using the variance instead of the standard deviation; for example, it is common to consider a relevant effect of a QTL (quantitative trait locus) or a gene from a 10% of the variance, which is approximately 1/3 the standard deviation. These figures should be considered not as thresholds but as rough indications of the importance of the effects, in
order to help the discussion of the results obtained in the experiments. Other values for relevance can be postulated, but if they are not based on economical or biological considerations, it is convenient to refer them to a fraction of the standard deviation of the trait.

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 10% bigger than the other one, or its effect is a 92% of the other one, which is more expressive than finding a difference between treatments of 1.7 points in a scoring for liver flavour. This can be complex in classical statistics, 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 (19), but is trivial for Bayesian statistics when combined with MCMC.

Appendix 1.2

We estimate ‘u’ from the data. If ‘u’ and the data ‘y’ are jointly normally distributed, their relationship is linear

\[ u = by + e = \hat{u} + e \]

and \( \text{cov}(\hat{u},e)=0 \). In this case,

\[ \sigma_u^2 = \sigma_{\hat{u}}^2 + \sigma_e^2 \quad \rightarrow \quad \sigma_e^2 = \sigma_u^2 - \sigma_{\hat{u}}^2 \]

Notice that when the joint distribution is not Normal, \( \text{cov}(\hat{u},e) \) may be different from zero and it should be considered. We assume in our models that the random effects depend on many factors with small effect each one; if so, we can conclude from the “Central limit theorem” that random effects are normally distributed. This

19 For example, the delta method, commonly used in quantitative genetics to estimate s.e. of ratios, does not work well for genetic correlations (Visscher, 1998).
theorem proves that under quite general conditions, the sum of many small independent effects tends to the Normal distribution.

**Appendix 1.3**

Let us estimate the genetic value ‘u’ of a bull as a fixed value or as a random value, using the average value of milk production of its daughters ‘\( \bar{y} \)’.

The estimate of the genetic value of the bull is, as a fixed effect, is

\[
\hat{u}_f = 2 \cdot \bar{y}
\]

For estimating (predicting) the genetic value as a random effect, we will use a linear prediction using the average production of the daughters (offspring)

\[
u = b \cdot \bar{y} + e = \hat{u}_r + e
\]

\[
\hat{u}_r = b \cdot \bar{y}
\]

Where \( b \) is a regression coefficient

\[
b = \frac{\text{cov}(u, \bar{y})}{\text{var}(\bar{y})^2}
\]

We assume in the example that the daughters come from different dams; i.e., they are half sibs, as usual when a bull is evaluated. We also know that the only thing in common that sire and daughters have is that daughters have half of their genetic information from the father. This is also the only thing in common between daughters. We know that the covariance between father and daughter and the covariances between daughters are, respectively:
\[
\text{cov}(u, \bar{y}) = \text{cov} \left( u, \frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n} n \cdot \text{cov}(u, y_i) = \frac{1}{2} \sigma_u^2
\]

\[
\text{cov}(y_i, y_k) = \frac{1}{2} \cdot \frac{1}{2} \cdot \sigma_u^2 = \frac{1}{4} \sigma_u^2
\]

\[
\text{var}(\bar{y}) = \text{var} \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n^2} \left( n \sigma_y^2 + n(n-1) \frac{1}{4} \text{cov}(y_i, y_k) \right) = \frac{1}{n} \left( \sigma_y^2 + (n-1) \frac{1}{4} \sigma_u^2 \right)
\]

\[
b = \frac{\text{cov}(u, \bar{y})}{\text{var}(\bar{y})} = \frac{\frac{1}{2} \sigma_u^2}{\frac{1}{n} \left( \sigma_y^2 + (n-1) \frac{1}{4} \sigma_u^2 \right)} = \frac{n \cdot \frac{1}{2} \sigma_u^2}{\sigma_y^2 + (n-1) \frac{1}{4} \sigma_u^2}
\]

When \( n=1 \), and taking into account that \( \sigma_u^2 \leq \sigma_y^2 \)

\[
b = \frac{1}{2} \cdot \frac{\sigma_u^2}{\sigma_y^2} \leq \frac{1}{2}
\]

And when \( n \to \infty \)

\[
b = 2
\]

This means that there is a “shrinkage” when estimating ‘u’ as a random effect, depending on the number of daughters we have. This is reasonable. If we estimate the value of a bull only by the average of the daughters, a bull with only one daughter having by chance a very high value (say 20,000 kg of milk) is overestimated, but overestimation is much more difficult when the bull has 100 daughters. Thus, bulls with only one daughter can be overestimated, and if we take as a selection criterion only the value of the current average of a bull’s daughters, we will select only overestimated bulls with little information. By evaluating the bull as a random effect, we shrink this value and move it towards the average of the population according to the amount of daughters the bull has; thus, a bull having a daughter of 20,000 kg of milk can see its genetic value estimated as only 12,500 kg. Notice that this is the value of the estimate, which
is different from the accuracy. Of course, a bull with only one daughter will have a lower accuracy than a bull with 100 daughters, but we are now stating that the estimate of its genetic value will be lower.

Appendix 1.4

A transformation of a ML estimator is also a ML estimator of the new parameter. For example, if we have a ML estimator of the variance, \( \hat{\sigma}^2_{\text{ML}} \), the ML of the standard deviation is \( \hat{\sigma}_{\text{ML}} = \sqrt{\hat{\sigma}^2_{\text{ML}}} \). It is intuitively reasonable that when our sample has maximum probability for a ML estimator (20), it also has the true value for a transformation of it.

To find the ML estimator of \( \theta \),

\[
\frac{\partial \hat{f}(y | \theta)}{\partial \theta} = 0
\]

If we have a transformed variable, for example \( g(\theta) \), to find the ML of \( g(\theta) \) we need to solve

\[
\frac{\partial \hat{f}(y | g(\theta))}{\partial g(\theta)} = 0
\]

But

\[
f(y | \theta) = f(y | g(\theta))
\]

because if \( \theta \) is given (‘fixed’), all the transformations are also fixed. Now,

\[20\text{ Remember, this happens when the estimator takes the true value.}\]
\[
\frac{\partial f(y|\theta)}{\partial \theta} = \frac{\partial f(y|g(\theta))}{\partial \theta} \cdot \frac{\partial g(\theta)}{\partial \theta} = 0
\]

As \( \frac{\partial g(\theta)}{\partial \theta} \neq 0 \) because \( \theta \) is not a constant, when \( \frac{\partial f(y|\theta)}{\partial \theta} = 0 \), then

\[
\frac{\partial f(y|g(\theta))}{\partial g(\theta)} = 0,
\]

and the same value for \( \theta \) is obtained solving both equations, therefore \( g(\hat{\theta}_{\text{ML}}) \) is the ML estimator of \( g(\theta) \).