AN INTRODUCTION TO

BAYESIAN ANALYSIS AND MCMC

REDUCED VERSION

A. 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

CHAPTER 1

DO WE UNDERSTAND CLASSICAL STATISTICS?

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.

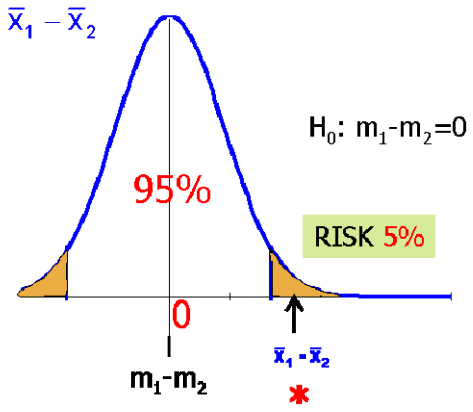


Figure 1.1. Distribution of repeated samples if H_0 is true. When our actual difference between sample averages lies in the shadow area we reject H_0 and say that the difference is "significant". This is often represented by a star.

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, 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 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).

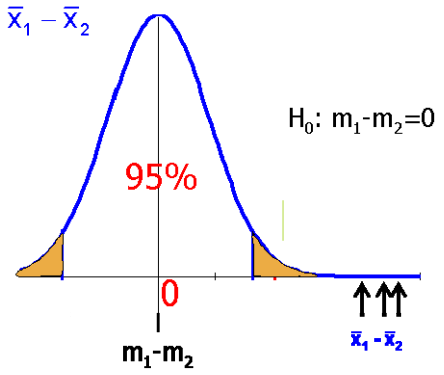


Figure 1.2. An error level of 5% of being wrong when rejecting the null hypothesis was accepted, but along his career, a researcher discovered that his data showed much higher evidence about the null hypothesis being wrong

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 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).

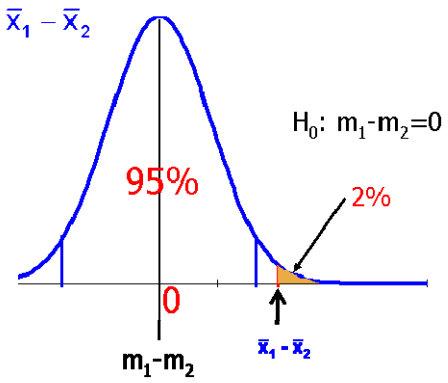


Figure 1.3. A *P-value* of 2% gives the probability of finding the current sample or a higher value if the null hypothesis holds. However this does not mean that the difference between treatments is "significant at a 2%".

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 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.

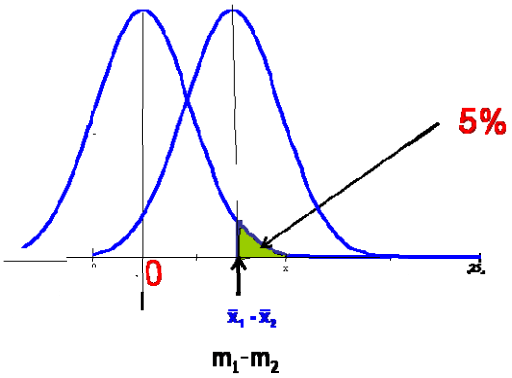


Figure 1.4. Distribution of the samples when the true value is the same as the actual sample. The current sample gives us a P- value of 5%. If the true value is the same as our sample, notice that when repeating the experiment, half of the times we will obtain "non significant " values (P>0.05).

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 ([1]). 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 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 $|\overline{x}_1 - \overline{x}_2| \geq 100$ 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

[1] 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 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 ([2]), 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 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.

The true value should be about the middle of the confidence interval, it is more proable that the true value is in the middle than in one of the sides: This is not true. Figure 5 shows that we do not know which one is our confidence interval. If it is one of the last ones of the figure, the true value will be closer to the left side of the interval. This is why it is important to consider the sides of the intervals, just because the true value may be there.

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

[2] 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).

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

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).

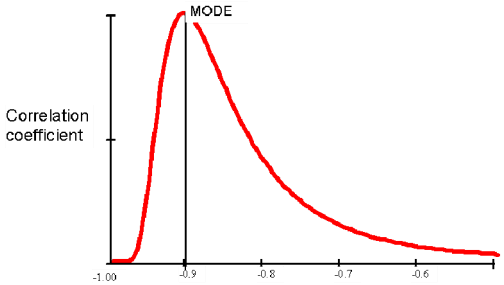


Figure 4. Sampling distribution of a correlation coefficient. Repeating the experiment, the samples are not distributed symmetrically around the true value**.**

.

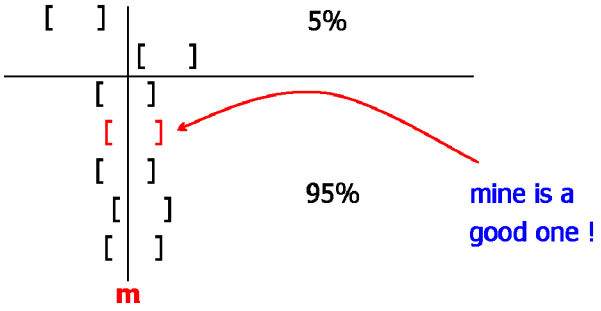
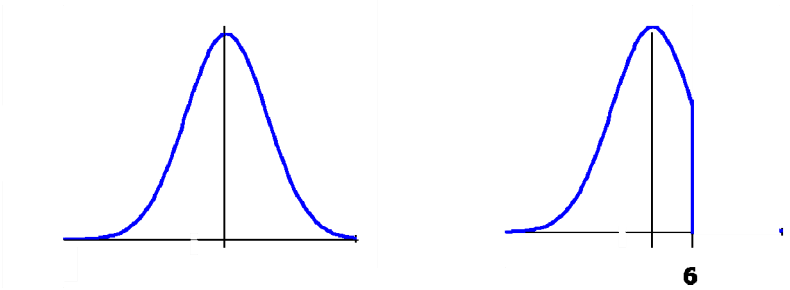


Figure 5. Repeating the experiment many times, a 95% of the intervals will contain the true value m. We do not know whether our interval is one of these, but we assume that it is. We hope not to be wrong many times along our career



[4.1, 4.5, 5, 5.5, 5.9]

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([3])

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

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

[3] 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 these definitions.

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

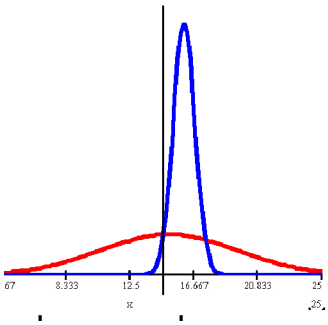


Figure 7. A biased estimator (blue) is not distributed around the true value but has lower risk than an unbiased estimator (red) that is distributed around the true value with a much higher variance.

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

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

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

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.

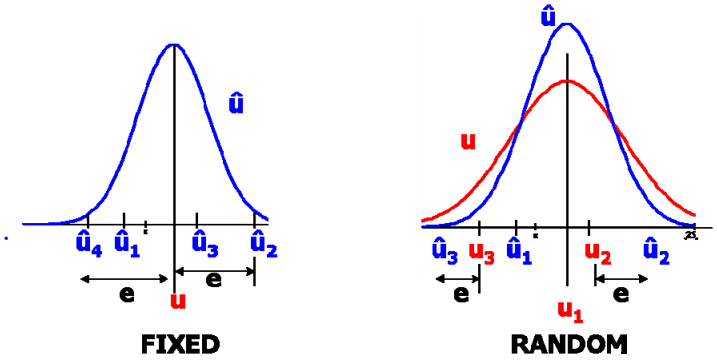


Figure 8. Distribution of the effects and their estimates when repeating the experiment an infinite number of times. When the effects are fixed the true value is constant, but when the effect is random it changes its value in each repetition. In red, the distribution of the true values; in blue, the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

The variances of the errors are also different

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

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

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.

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 the sow as a "fixed" effect and we could have estimated it, since we had five litters per sow. 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

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 fr ıe second population.

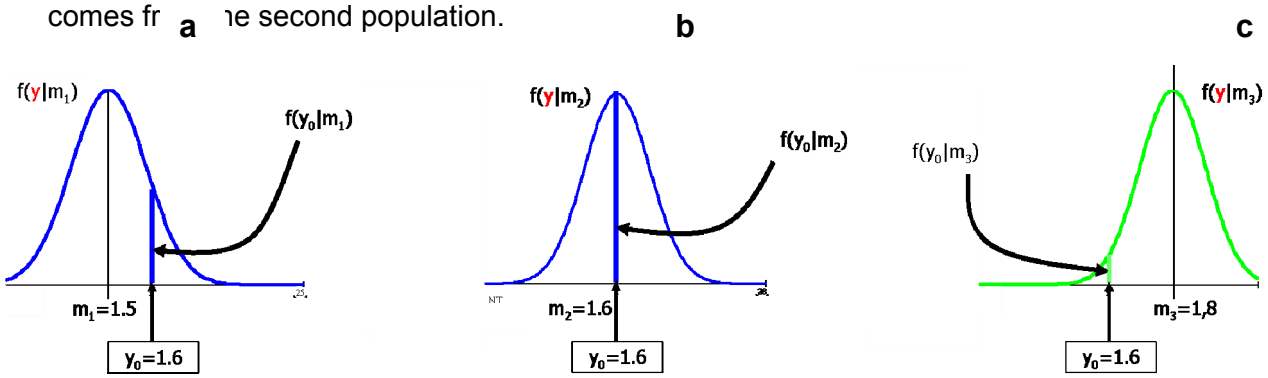


Figure 9. Three likelihoods for the sample y_0= 1.6. **a**: likelihood if the true mean of the population would be 1.5, **b**: likelihood if the true mean of the population would be 1.6. **c**: likelihood if the true mean of the population would be 1.8

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).

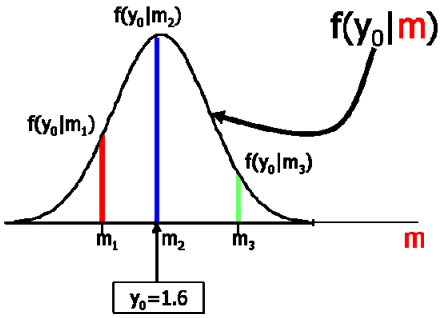


Figure 10. Likelihood curve. It is not a probability because its values come from different probability distributions, bur it is a rational degree of belief. The notation stress that the variable (in red) is m and not y_0 that is a given fixed sample.

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. 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

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

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.

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.

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 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.

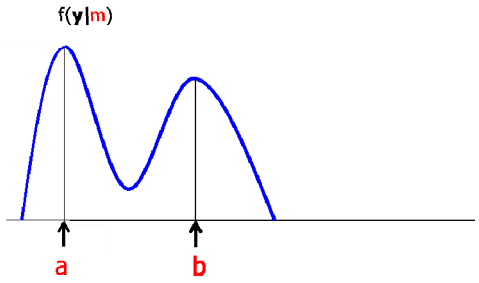


Figure 11. Likelihood curve. Here m can take "likely" the values 'a' or 'b', however the frequentist school will only take the maximum at 'a'

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. As Fisher says, for the case of estimating the true coefficient of correlation ρ from the value r obtained in a sample:

> "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.

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 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, but is trivial for Bayesian statistics when combined with MCMC.