Basic Statistical Analysis in Life and Environmental Sciences

Chapter 2 - Statistical Models and Inference

Rodrigo Labouriau [1]

2022

[1]Applied Statistics Laboratory, Department of Mathematics, Aarhus University.

Contents

Chapter 2

Statistical Models and Inference - Draft

R.Labouriau [1]

Draft - Please do not circulate. [2]

2.1 Parametric statistical models

In this section we illustrate and then define the important notion of statistical model. Emphasis will be given in parametric models, but we will describe also non-parametric models. We start with four simple examples.

Example 1 (Binary trials) *In the example of the binomial experiment we introduced the random variable Y representing the number of "successes" observed in two independent binary trials with probability of success p. The distribution of the discrete variable Y has probability function given by*

$$f_Y(y) = \begin{cases} (1-p)^2, & \text{if } y = 0, \\ 2p(1-p), & \text{if } y = 1, \\ p^2, & \text{if } y = 2\,. \end{cases} \tag{2.1}$$

[1] Applied Statistics Laboratory, Department of Mathematics, Aarhus University.

[2] Last revised: February, 2022.

That is, each value of p determines a distribution (using the formula (2.1) above), which has a clear interpretation in terms of the experiment that generated Y. The class of all such distributions is what we call a parametric statistical model *and p is a* parameter *indexing this model. That is, we deduced that the distribution of Y is contained in a certain class of distributions (the binomial distributions with two trials), using the very basic properties of the experiment. These properties allowed us only to determine a family of distributions containing the distribution of Y, not* **the** *distribution of Y. Now, to completely determine the probability law of Y we must find which element of this class of distributions is the distribution of Y (or is the best candidate in the class to represent the distribution of Y). This is equivalent to say that we want to determine, or to estimate, the unknown value of the parameter p that indexes the parametric model. This estimation is usually based on observations of executions of the experiment.*

Example 2 (Counts of alpha-particles) *In a classic experiment Rutherford and Geiger (1910) counted the number of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds. The first observed counts were:* 2, 1, 3, 5, 3, 5, 3, 4, ... *The results of* $10,097$ *counts performed by Rutherford and Geiger are displayed in Table 2.1 and Figure 2.1.*

Table 2.1: Frequency of 10,097 counts of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds.

Counts:	0	1	2	3	4	5	6	7
Frequency:	57	203	383	525	532	408	273	139
Counts:	8	9	10	11	12	13	14	+ 15
Frequency:	45	27	10	4	0	1	1	0

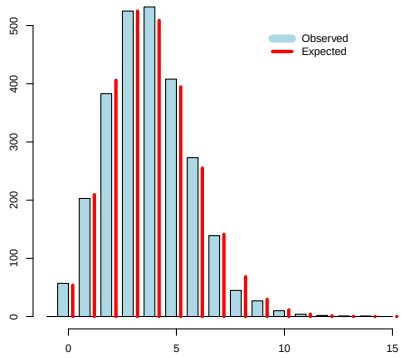

Figure 2.1: Counts of alpha-particles and the expected number of counts under a Poisson distribution with $\lambda = 3.87$.

This data has been classically modelled using a Poisson distribution. Indeed, it is possible to prove that if this experiment fulfil three rather basic assumptions, then the number of counts should be Poisson distributed. The basic assumptions are: 1) The time of arrival of a particle in the counter is homogeneously distributed in the observation interval; 2) The number of particles that arrive in two disjoint intervals are independent; 3) The particles do not arrive at the same time (i.e., the probability of two or more particles arrive in the counter in a short interval divided by the probability that only one particle arrives in this interval tends to zero as the length of the interval approaches zero). That is, under these assumptions it can be shown that[3] *the probability of observing, say k particles (for $k = 0, 1, 2, 3, \ldots$) is given by the formula*

$$\frac{e^{-\lambda}\lambda^k}{k!} \quad \textit{for a } \lambda > 0 \ . \tag{2.2}$$

Here $k! = k \cdot (k-1) \cdot \ldots \cdot 1$ and $0! = 1$. Note that the formula (2.2) above depends on the constant λ which is only assumed to be positive up to now.

[3]We come back to this point, for the moment let us just believe in this result.

Changing the value of λ, changes the probability law described in (2.2). In fact, without specifying the value of λ, equation (2.2) is only saying that the observed number of particles is Poisson distributed. That is, we just specified a collection (or family) of possible distributions for the number of particles; this collection of distributions is called a statistical model. Since there is a parameter that label uniquely the distributions in the statistical model, namely λ, we say that the statistical model in question is a parametric model and we refer to λ as the parameter of the model. Just as in example 1, the probability law describing the observed number of particles in the present experiment is specified in two steps: First we determine a collection of possible probability distributions for the observed results (i.e., we determine the statistical model, in this case we deduced it from simple basic assumptions of the data), and secondly, we (will) determine which of those possible distributions well describe the results (a process that is called estimation and will be discussed below).

Example 3 (Weights of batches of seeds) *In this example we consider the data of seed weights of Dolichos biflorus (a bean) obtained with an automatic weighting device. The data consisted of the individual weights of the seeds of* 50 *batches, each batch containing* 50 *seeds, all in all we have the weights of* $2,500$ *seeds. The histograms and the normal Q-Q plots displayed in Figure 2.2 suggest that the individual weights of the seeds are clearly not normally distributed; however, it is reasonable to assume the total weight of the batch to be normally distributed. This is in agreement with the central limit theorem (see chapter 1). To see that, note that the weight of the batch is the sum of the weight of* 50 *individual weights of seeds, and if we assume that the* 50 *weights of seeds in each batch are independent, identically distributed (i.e., the batches are homogeneous) and have a finite variance, then we might evoque the central limit theorem and conclude that the total weight of the batches are approximately normal distributed (here we assume also that number of observations in each batch, i.e.,* 50*, is large enough to ensure that the approximation to the normal distribution is good enough). It is then rea-*

sonable to model the 50 *total weights the batches as (approximately) normally distributed. Formally, we might represent the data by* 50 *random independent and identically distributed random variables, say* $X_1, X_2, \ldots, X_{50}$*, with,*

$$X_1 \sim N(\mu, \sigma^2) \ . \tag{2.3}$$

Here μ *and* σ^2 *are not specified (yet);* μ *might be any real number (although it would not be reasonable to use a negative value for* μ*) and* σ^2 *might be any positive number. Again, changing the values of* μ *and* σ^2 *changes the probability law that is used to describe the data; therefore, equation (2.3) defines a collection of probability distributions, which we say that is a parametric (statistical) model. Here the model is said to be parametric because there is a finite number of parameters, namely* μ *and* σ^2*, that index all the distributions in the model.*

Example 4 (*Weights of individual seeds) [4] *As discussed in the example 3 above, it is not reasonable to use the normal distribution to describe the individual weights of the seeds (see Figure 2.2). Working with a normal model for the total weight of the batches might suffice for most of the practical purposes, but there are other alternatives. One of those alternatives is to assume that the distribution of the individual weights of the seeds is a (unspecified) continuous symmetric distribution (i.e., a continuous distribution with symmetric probability density). The collection of all continuous symmetric distributions would be then the statistical model. Note that this model is much larger than the normal distribution model used in example 3. Indeed, the normal distributions are symmetric (with symmetry around the expectation,* μ *in the notation of the formula (2.3)), and there are other symmetric distributions. This model cannot be indexed by a finite number of parameters (but the prove of this statement is tricky); therefore we say that this model is non-parametric. This type of model is behind a range of non-parametric tests as the Wilcoxon test and Kruscal-Walis test.*

[4] Optional reading.

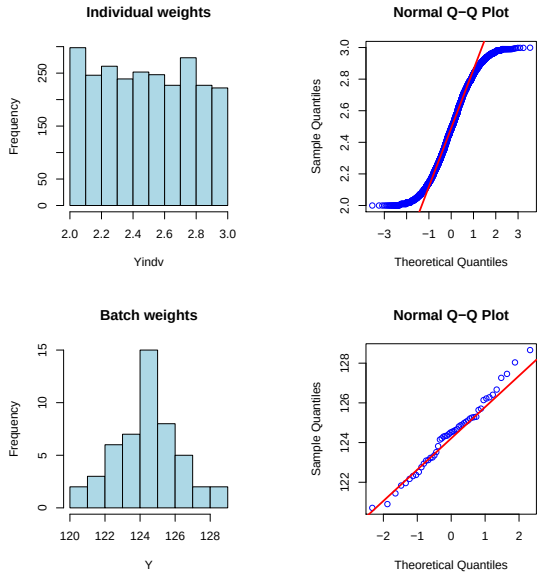

Figure 2.2: Histogram and normal Q-Q-plot (*i.e.*, theoretical quantiles for a normal distribution plotted against the observed quantiles, the points should lie in a strait line, if the data is normally distributed) for the individual weights of the seeds (above, $2,500$ seeds) and for the batch total weights (below, 50 batches).

In general, the probability law of the random quantity in study, say Y, is unknown and we determine it in two steps: First, we choose a class of distributions that are good candidates for being the distribution of Y (or for well approximate the probability law of Y). This class of distributions is called the *statistical model* . If the statistical model can be indexed by a finite number of parameters (as in the examples 1, 2 and 3) we say that we have a *parametric statistical model* or simply a *parametric model*. [5]The next step is to determine the best candidate in the parametric model for representing the probability law of Y, on the basis of observations of the experiment. This general procedure is called *parametric point estimation* or simply *point estimation* . The theory of point estimation is rather developed in modern statistics and there are many general techniques for deriving good point estimates. We will concentrate here in one classical general technique called maximum likelihood estimation. This technique produces good estimators in many cases (but, not always!) and is by far the most popular estimation method.

2.2 Likelihood function and maximum likelihood estimation

We will introduce here the basic notions of likelihood function, maximum likelihood estimation and other related notions. These is the kernel of the classic techniques for estimation. We will introduce these notion using the examples 1 and 2 but please keep in mind that these techniques are quite general and can be easily extended for much more complex models.

Example 5 (Binary trials - continued) *Let us discuss a bit more the simplest example we studied, namely, the binary trial where we observe the results*

[5]In non-parametric statistics the statistical models are very large families of distributions. They are so large that it is not possible to index them with a finite number of parameters. One example is the class of symmetric continuous distributions. This family can not be indexed by a finite number of parameters (but this is not easy to prove!).

of a trial with two possible outcomes: success or failure (e.g., tossing a coin and observing whether the result is a tail). Suppose that the probability of success in the is trial is the number p *contained in the interval* $(0,1)$ *(e.g.,* $p = 1/2$*, if the coin in the previous example is fair). Now, consider the situation where we perform this binary trial four times independently and register the four result. We use four random variables to represent this situation, say* X_1*,* X_2*.* X_3 *and* X_4*, which will denote the result of the first, the second, the third and the fourth binary trial, respectively. We use the convention that* X_1 *takes the values* 0 *or* 1 *if we observe a failure or a success in the first trial, respectively (i.e.,* X_1 *is the number of successes in the first trial). An analogous convention is used for the other three random variables. Now, these four random variables are independent (because the four trials are assumed to be independent) and identically distributed (since we assume that the probability of success is constant, namely equal to* p*). These two assumptions will be used strongly in the calculations below.*

For example, suppose that we observe $X_1 = 0$*,* $X_2 = 1$*.* $X_3 = 0$ *and* $X_4 = 0$ *(i.e., we observe the results* $0, 1, 0$ *and* 0*). The probability of obtaining this result is*

$$\begin{aligned} P(X_1{=}0, X_2{=}1, X_3{=}0, X_4{=}0) &= \textit{(by independency)} \\ &= P(X_1{=}0)P(X_2{=}1)P(X_3{=}0)P(X_4{=}0) \\ &= (1-p).p.(1-p)(1-p) \\ &= p(1-p)^3. \end{aligned} \tag{2.4}$$

The probability calculated above depends on the parameter p*. Figure 2.3 shows how the probability of observing the results* $0, 1, 0$ *and* 0 *changes when the value of* p *varies. Analogous calculations yield the same probability for the results* $X_1 = 1$*,* $X_2 = 0$*,* $X_3 = 0$ *and* $X_4 = 0$ *or* $X_1 = 0$*,* $X_2 = 0$*,* $X_3 = 1$ *and* $X_4 = 0$ *or even* $X_1 = 0$*,* $X_2 = 0$*,* $X_3 = 0$ *and* $X_4 = 1$ *(the order of the factors in the second equality of the expression (2.4) are just permuted for the different results). That is, in each these cases the probability of observing the result is just* $p(1-p)^3$*.*

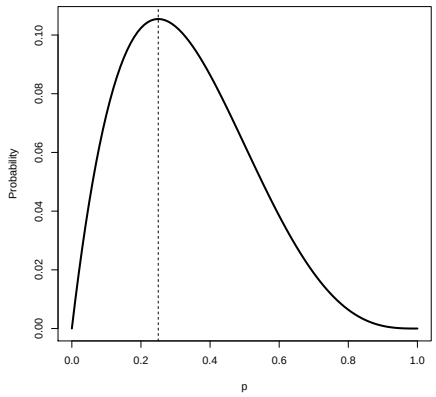

Figure 2.3: Probability of the results $X_1 = 0$, $X_2 = 1$, $X_3 = 0$ and $X_4 = 0$ as a function of the parameter p (where p is the probability of success in one binary trial). Note that this function has a maximum at $p = 1/4$.

The probability of observing the result $X_1 = 1$, $X_2 = 1$, $X_3 = 0$ *and* $X_4 = 0$ *or* $X_1 = 0$ *is*

$$\begin{aligned} P(X_1{=}1, X_2{=}1, X_3{=}0, X_4{=}0) &= P(X_1{=}1)P(X_2{=}1)P(X_3{=}0)P(X_4{=}0) \\ &= p.p.(1-p)(1-p) \\ &= p^2(1-p)^2. \end{aligned} \tag{2.5}$$

Figure 2.4 shows how the probability of observing the results $1, 1, 0$ *and* 0 *changes when we change the value of* p*. Analogous calculations yield the same probability for all the results for which exactly two success are observed (and therefore two failures are also observed). Continuing in this way we can calculate the probability of any possible result of the four binary trials. To do so, denote by* k *the number of successes observed in the four binary trials. Clearly,* k *can only take the values* 0*,* 1*,* 2*,* 3 *or* 4*.*

Using a procedure similar to the calculations presented in (2.4) and (2.5) it is not difficult to see that the probability of observing any result that yields

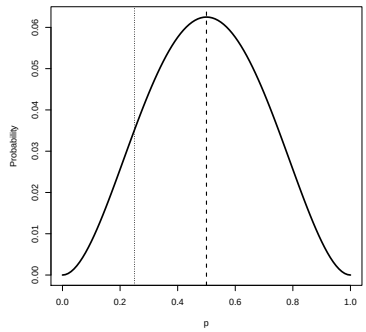

Figure 2.4: Probability of the results $X_1 = 1$, $X_2 = 1$, $X_3 = 0$ and $X_4 = 0$ as a function of the parameter p. Note that this function has a maximum at $p = 1/2$, marked as an interrupted line; for comparison, a dotted line marks the maximum of the probability of observing $X_1 = 0$, $X_2 = 1$, $X_3 = 0$ and $X_4 = 0$ viewed as a function of p.

exactly k, for $k = 0, 1, 2, 3$ or 4, is

$$P(\text{"observing a result with } k \text{ successes"}) = p^k(1-p)^{4-k}. \tag{2.6}$$

Note that the probabilities given in (2.6) are always a function of the parameter p (see Figure 2.5). In this way, the probability of any possible result of the experiment "perform four independent binary trials with probability p of success in each trial and register the number of successes" can always be expressed as a function of p. Now, once we have performed the experiment and observed a certain result we can think on the probability function for the sample as a fixed function of the parameter p. This function is called the likelihood function for the parameter p based on the observations. I stress that the likelihood function depends on the observations in the sense that if we had observed another result (i.e., another value of k), then the function would be different. It is reasonable to estimate the parameter by choosing the value of the parameter that maximizes the likelihood function, i.e., estimate the parameter by the value that is associated to the distribution that attributes

the larger probability to the observed sample (among the distributions in the statistical models). This estimator is called the maximum likelihood estimate of p based on the observations. In the particular example we consider the likelihood function takes the form, for $k = 0, 1, 2, 3$ and 4,

$$L(p) = p^k(1-p)^{4-k}. \tag{2.7}$$

The likelihood function has the a maximum at, 0, $1/4$, $2/4$, $3/4$ and 1 when k is equal to 0, 1, 2, 3 and 4, respectively. That is, when we observe a result that yields k successes, the value of p that maximises the likelihood function is $\hat{p} = k/4$ and we estimate the parameter p by this value (see Figure 2.5).

* Some details on the maximisation of the likelihood function [6]

Direct maximisation of the likelihood function is feasible, however, in many situations it is easier to maximize the logarithm of the likelihood function, called the log-likelihood function. (since the maximum of a positive function coincides with the maximum of its logarithm). [7] The many products always present in the likelihood function (due to the independence of the observations) are transformed into sums in the log-likelihood function, since the logarithm transforms products in sums. We will denote the log-likelihood by l.

The log-likelihood functions of the example considered, performing four iid binary trial and counting the number k of successes, are given by, for $k = 0, 1, 2, 3$ and 4,

$$l(p) = \log\left[L(p)\right] = \log\left[p^k(1-p)^{4-k}\right] = k\log(p) + (4-k)\log(1-p)\ .$$

Figure 2.5 displays the graphs of the log-likelihood function for the example in play.

[6] Optional reading.

[7] There are other theoretical reasons for justifying the maximisation of the log-likelihood function, instead of the likelihood function, but we will not go into such mathematical details here.

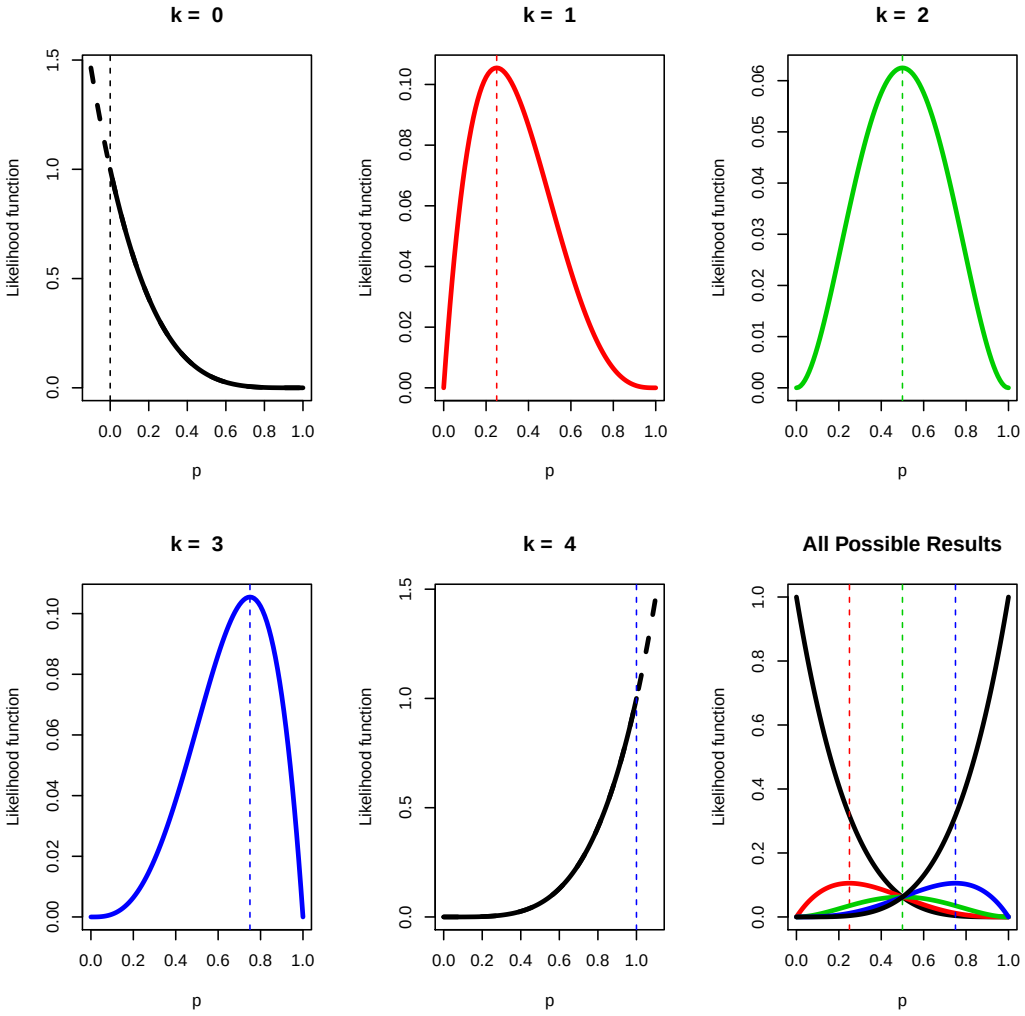

Figure 2.5: Probability of all the possible results of four independent binary trials as a function of the parameter p. Note that these functions have maxima at different points marked with vertical interrupted lines.

Since the derivative of a smooth function [8] *equals to zero at the maximum of the function,* [9] *we can then find the maximum of the log-likelihood functions, by differentiating it with respect to the parameter* p *and equating the derivative to zero. This is an old trick!*

In the case of the experiment "observe the number of successes of four independent binary trials" which we are discussing, likelihood function is given by

$$L(p) = p^k(1-p)^{4-k} \ ,$$

the log-likelihood function is

$$l(p) = \log\left[L(p)\right] = \log\left[p^k(1-p)^{4-k}\right] = k\log(p) + (4-k)\log(1-p) \ .$$

In the case that k *is equal to zero or* 4 *the likelihood (and the log-likelihood) function has maximum at* 0 *and* 1*, since the likelihood function keeps increasing when it reaches the boundary of the interval between* 0 *and* 1 *(see Figure 2.5). We study next the cases where* k *is different than* 0 *and* 1*.*

Differentiating the expression above with respect to p *yields the score function*

$$S(p) = \frac{\partial}{\partial p}\left[k\log(p) + (4-k)\log(1-p)\right] = \frac{k}{p} - \frac{(4-k)}{(1-p)} \ .$$

Equating the score function to zero (to find its maximum) yields the score equation, which is given by

$$\frac{k}{\hat{p}} - \frac{(4-k)}{(1-\hat{p})} = 0 \ , \text{ which is equivalent to } \frac{k}{(4-k)} = \frac{\hat{p}}{(1-\hat{p})} \ . \tag{2.8}$$

It is easy to see that the solution for equation (2.8) is $\hat{p} = k/4$ *(just make a substitution).*

---

[8] You may think on the inclination of the tangent of the graph of the function at the point, if you are not acquainted with derivatives.

[9] Try to make a plot, if you doubt of this result.

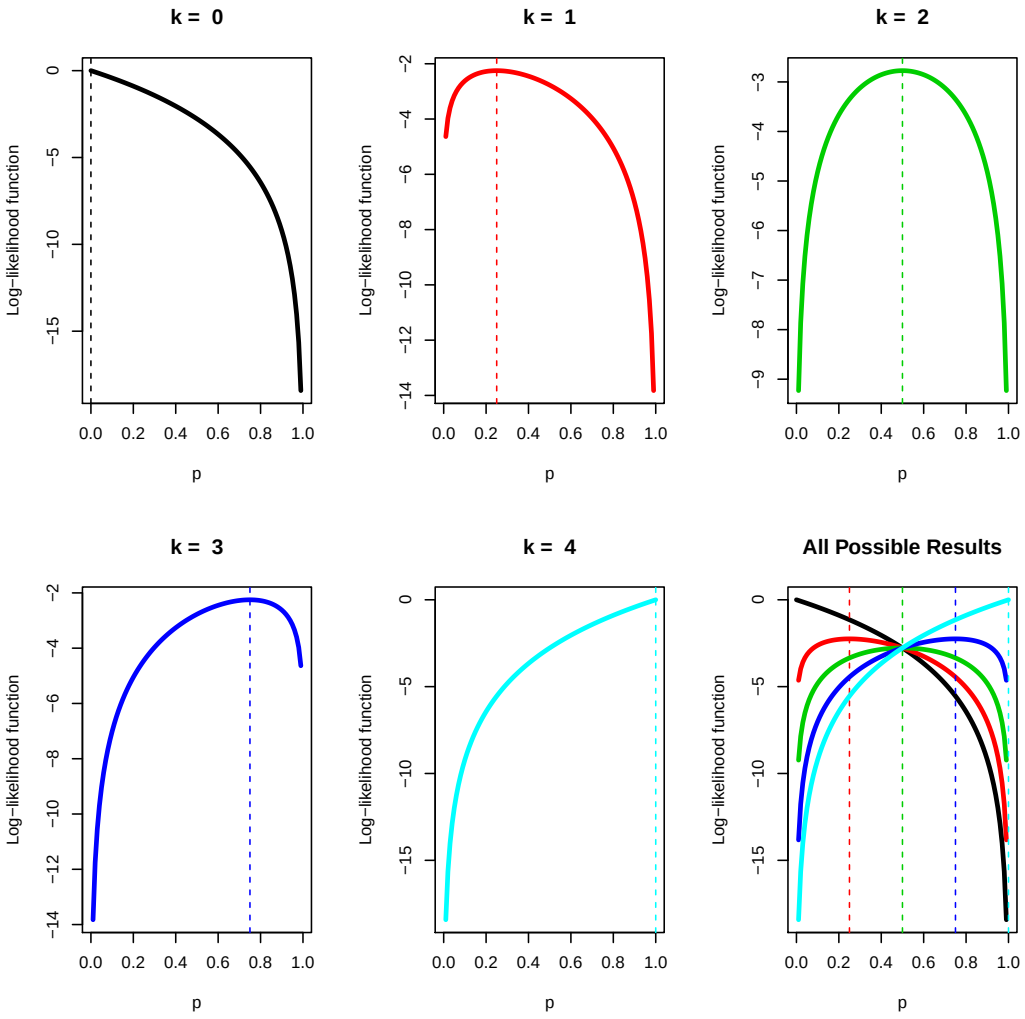

Figure 2.6: The five possible forms of the log-likelihood function (*i.e.*, logarithm of the probability of each of the possible results viewed as a function of the parameter p) for four independent and identically distributed binary trials. Note that these functions have maxima at different points marked with vertical interrupted lines.

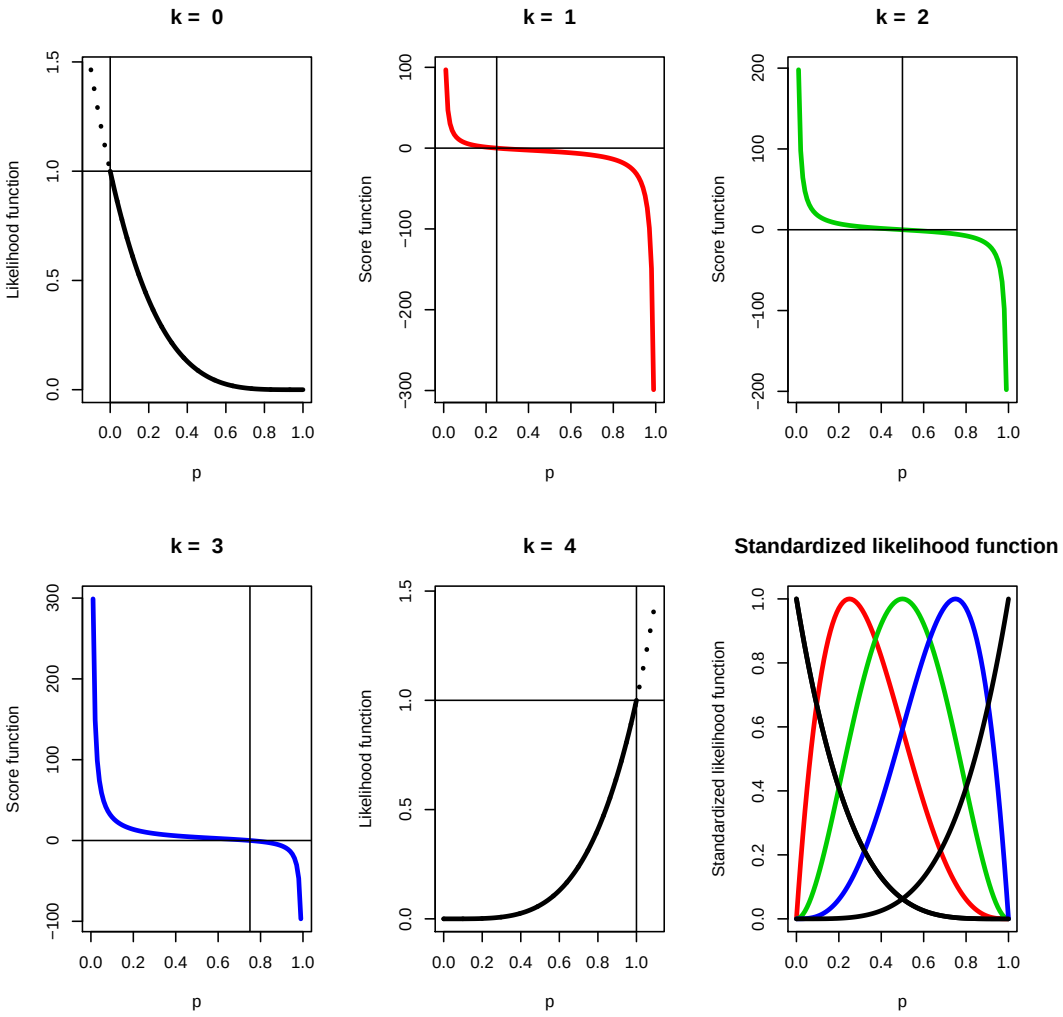

Figure 2.7: The five possible forms of the score function (*i.e.*, the derivative of the log-likelihood function) for four independent and identically distrubuted binary trials. Note that these functions have zeroes at different points marked with vertical interrupted lines (corresponding to the maxima of the log-likelihood functions).

Example 6 (Counts of alpha-particles - Continued) *Here we show how to calculate the maximum likelihood estimate in the simple Poisson model discussed in example 2. There we described the number of alpha particles arriving to a counter as being Poisson distributed. In the experiment,* $10,097$ *counts for alpha particles detected in time intervals of 72 seconds were registered and we argued that these coins were reasonably described as being realisations of a Poisson distribution with a fixed, but unknown, intensity parameter (i.e., expected value)* λ*. That is, the results of this large experiment could be described by* $10,097$ *independent (and identically distributed) random variables, say* $Y_1, Y_2, \ldots, Y_{10,097}$*, each of them being Poisson distributed. Once we have performed the experiment (i.e., counting* $10,097$ *times the alpha particles) we will have* $10,097$ *values, say* $y_1, y_2, \ldots, y_{10,097}$*. These values are not random quantities since they are known after we have performed the experiment (and had been kept in a dataset); therefore we use small letters to denote them (in contrast with the capital letters used to denote random variables). In the statistical terminology the observed values* $y_1, y_2, \ldots, y_{10,097}$ *are said to be a* sample*. We might now calculate the probability of the observed result by*

$$\begin{aligned} P\left(Y_1 = y_1, \ldots Y_{10,097} = y_{10,097}\right) &= \text{(by independence and (2.2))} \\ &= \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \cdot \cdots \cdot \frac{e^{-\lambda}\lambda^{y_{10,097}}}{y_{10,097}!} \,. \end{aligned} \tag{2.9}$$

Note that the the probability calculated in the expression (2.9) above depends on the unknown parameter λ*. This probability viewed as a function of* λ *(and considering the* **observed** *values* $y_1, y_2, \ldots, y_{10,097}$ *as fixed) is called the* likelihood function λ based on the sample $y_1, y_2, \ldots, y_{10,097}$ *or simply the* likelihood function*, which is usually denoted by* L*. That is,*

$$L(\lambda) = \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \cdot \cdots \cdot \frac{e^{-\lambda}\lambda^{y_{10,097}}}{y_{10,097}!} \,. \tag{2.10}$$

The value of the parameter λ *that maximizes the likelihood function (if any) is called the* maximum likelihood estimate of λ*, and is usually denoted by* $\hat{\lambda}$*.*

Note that the likelihood function in (2.10) depends on the sample (i.e., the observed values), although the sample is considered fixed (since it is known after we have performed the experiment).

Once established the likelihood function (using the results of the experiment), the calculation of the maximum likelihood is just a matter of finding a maximum of a function. It is convenient to maximize the logarithm of the likelihood function instead of the direct maximisation of the likelihood function. This can be seen by observing that the right hand of the expression (2.10) is a product of $10,097$ *terms, which is difficult to work with (e.g., try to calculate the derivative of this product!). We define then the* log-likelihood function *by*

$$\begin{aligned} l(\lambda) &= \log\left(L(\lambda)\right) \\ &= 10,097\lambda + \log(\lambda)\sum_{i=1}^{10,097} y_i - \sum_{i=1}^{10,097}\log(y_i!)\ . \end{aligned} \tag{2.11}$$

Now, to find the maximum of the log-likelihood function, we just differentiate it with respect to λ *and equate it to zero. The derivative of the log-likelihood function (with respect to* λ*)* [10] *is called the* score function *and is given by*

$$S(\lambda) = \frac{\partial}{\partial\lambda} l(\lambda) = 10,097 + \frac{\sum_{i=1}^{10,097} y_i}{\lambda}\ . \tag{2.12}$$

Figure 2.8 displays the log-likelihood function and the score function for the Rutherford Geiger experiment. Now equating the score function given in (2.12) yields the following equation

$$10,097 + \frac{\sum_{i=1}^{10,097} y_i}{\hat{\lambda}} = 0\ ,$$

with has solution

$$\hat{\lambda} = \frac{\sum_{i=1}^{10,097} y_i}{10,097}\ .$$

A similar argument yields that in general the maximum likelihood estimate for the intensity parameter of a sample of a Poisson distribution is nothing

[10] That is the function that gives the inclination of the tangent of the graph of the function l at each point.

but the total sum of the counts divided by the number of times the experiment was performed, i.e., the sample mean.

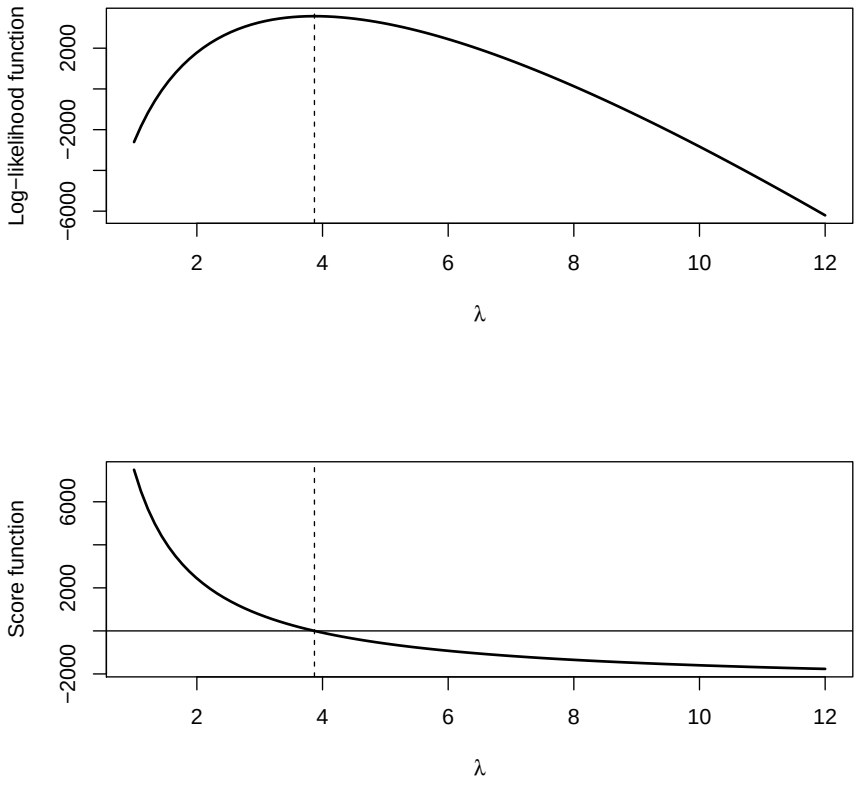

Figure 2.8: The log-likelihood function and the score function (*i.e.*, the derivative of the log-likelihood function) for the Rutherford Geiger experiment.

In general, the estimation can be done by defining a proper likelihood function which express the probability of observing the results of the experiment (or observations) in terms of the parameters in the model. The *maximum likelihood estimate* is defined then by the value that maximises the likelihood function (obtained with the observed data). In the examples considered above it is relatively simple to write the score equation and find explicit solution of it. [11]However, in most of the practical cases this can be a difficult, if not virtually impossible, task. Therefore the maximum likelihood estimate is usually calculated with numerical approximative methods. The most popular is the method of Newton [12] .

In order to verify whether the solution of the score equation is indeed a maximum (and not a minimum or a saddle point) one usually calculate the second derivative (i.e. the derivative of the derivative) of the log-likelihood. If the solution of the score function is a maximum, then the second derivative of the log-likelihood must be negative. The positive quantity given by minus the second derivative of the log-likelihood evaluated at the maximum likelihood is called the *observed information* and denoted by $I(\hat{p})$. That is,

$$I(\hat{p}) \;=\; -\frac{\partial^2}{\partial p^2}\, l(p)\,|_{p=\hat{p}} \quad . \tag{2.13}$$

The observed information measures the curvature of the log-likelihood at its maximum. A low value of the observed information means that the log-likelihood is flat, which in turn implies that the data provide almost no information about the parameters. It can be shown (but we will not do it here) that, under mild regularity conditions, the maximum likelihood estimate is approximately normal distributed with the mean equal to the real value of the parameter and the variance equal to the inverse of the observed information, for sufficiently large samples. We can therefore use the observed information to measure the dispersion of the maximum likelihood estimate

[11]That is why we choose them as the first examples.

[12]Also called Newton-Raphson method, even thought Raphson only made the first FORTRAN program implementing the idea that Newton had *many* years before!

around the true value of the parameter. Figure 2.9 shows the second derivative of the log-likelihood function for the two examples considered above.

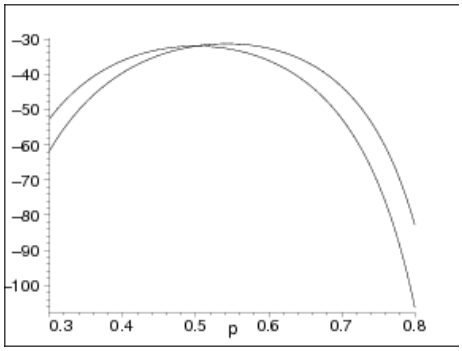

Figure 2.9: Second derivative of the log-likelihood function for the results $y_1 = 1$, $y_2 = 2$, $y_3 = 0$ and $y_4 = 1$ (curve to the left) and $y_1 = 1$, $y_2 = 0$, $y_3 = 2$ and $y_4 = 2$ (curve to the right) as a function of the parameter p.

2.2.1 *Maximum likelihood estimation for exponential dispersion models with fixed scale

[13]

We give now an useful example of the calculations involved in the maximum likelihood estimation. Suppose that we have a sample from a distribution contained in an exponential dispersion model. Let us consider the scale parameter λ as fixed and known and treat the estimation of θ. This will be the case when using the Poisson and the binomial distribution, where the scale parameter assumes indeed only the value 1. Moreover, as we will see, the estimation of the parameter θ will not be affected by the value assumed by the scale parameter.

[13] Optional reading.

The likelihood function for a sample $y_1, \ldots, y_n$ is

$$L(\theta) = \prod_{i=1}^{n} \exp\left[\lambda\left\{y_i\theta - b(\theta) - c(y_i, \lambda)\right\}\right] .$$

The log-likelihood function is given by

$$l(\theta) = \sum_{i=1}^{n} \lambda\left\{y_i\theta - b(\theta) - c(y_i, \lambda)\right\} .$$

The score function is obtained by differentiating the log-likelihood with respect to θ, i.e.

$$S(\theta) = \sum_{i=1}^{n} \lambda\left\{y_i - b'(\theta)\right\} .$$

Equating the score function to zero we obtain

$$\lambda\left\{\sum_{i=1}^{n} y_i\right\} - \lambda n b'(\hat{\theta}) = 0\,.$$

Since λ is different than zero, we can eliminate this parameter from the equation (no matter which value it takes). The last equation is equivalent to

$$b'(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} y_i\,. \tag{2.14}$$

This equation has a simple interpretation. Remember that the expectation of a random variable X following an exponential dispersion model ($X \sim ED(\theta, \lambda)$) is $b'(\theta)$. Therefore, the equation (2.14) says that the maximum likelihood estimator of θ (under an exponential dispersion model) is the value of θ that makes the expectation of the random variable that generated the sample equal to the mean of the sample. This explains the results we obtained for the binomial distribution (with two trials) in terms of a more general principle.

The maximum likelihood estimator of θ is

$$\hat{\theta} = (b')^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} y_i\right\} = u\left\{\frac{1}{n}\sum_{i=1}^{n} y_i\right\} .$$

Moreover, the observed information is given by

$$I(\theta) = n b''(\theta)\,.$$

2.3 Confidence intervals

Providing an estimator of a parameter under a parametric model, i.e. giving a number that points to a distribution in the parametric model that (hopefully) reasonably represents the data, is not always satisfactory. Estimators are functions of the observations, which in turn are realizations of random variables. Therefore *estimators are indeed random quantities* and would take different values if a (even slightly) different sample were observed. As a consequence, when reporting the results of an estimating process it is a good practice to report also the dispersion of the estimator. Another way to circumvent this problem is to supply, when possible, a region that will contain "the true value of the parameters" with high probability (typically 0.95 or 0.99 [14]). Since we are talking about models and not about absolute correct representation of the reality, there are no "true values of the parameters". This is just a way to say that we want to find a region that would contain our good candidate to represent the distribution that generated the data with high probability. From the point of view of prediction, this means that other samples originated from the same mechanism that generated the data would produce an estimator that would fall in the region with high probability. When such a region is an interval, it is called a *confidence interval* with level given by the probability of the region. We illustrate bellow two useful methods of construction of confidence intervals.

2.3.1 Example: Confidence interval for cation exchange capacity in soils

The following example from a study of a large experiment on soil fertility will serve us to illustrate the construction of a confidence interval. The cation exchange capacity (CEC) capacity was measured in 120 soil samples (m equiv/ 100 g soil) and is displayed in Figure 2.10. Here we will use a

[14] There is nothing special with these numbers. They are chosen simply by convention.

simple statistical model to describe the CEC by assuming that the measures come from a normal distribution with variance 1.

More precisely, we denote the measurements by $y_1, \ldots, y_{120}$ and assume that they are independent and identically distributed following a normal distribution with variance 1 and unknown expectation μ. In symbols, $y_i \sim N(\mu, 1)$ (for $i = 1, \ldots, 120$). Here μ is an unknown parameter, which we want to estimate.

The model proposed, although simple, gives a reasonable description of the data. Indeed, the density of a normal distribution with expectation 2.87 (*i.e.*, the sample mean) and variance 1 resembles the histogram of the data as can be seen in Figure 2.10. [15] It is not difficult to show that the maximum likelihood estimate for μ under this model is the sample mean. [16]

We want now to calculate an interval, say CI, around the estimate $\hat{\mu} = \bar{y}$ that contains the parameter μ with high probability. We pre-fix this probability for doing the calculations. Lets say that the pre-fixed probability is $\alpha = 0.95$ (*i.e.*, 95 %). [17] The interval Ci is of the form

$$CI = [\bar{y} - a, \bar{y} + a] \ , \tag{2.15}$$

where a is a number, that will calculate, such that the probability that μ is in CI is $\alpha = 0.95$.

We need some ingredients for the calculation of the interval CI. First we calculate the distribution of $\bar{y}$. Since $\bar{y} = 1/n \sum_{i=1}^{n} y_i$ and each of the

[15] The normality of these data can be verified using more appropriate methods using the methods we will study later. Moreover, it is not necessary to introduce the assumption that the variance is 1, but this simplifies matters and is chosen for pedagogical reasons. Anyway the model used here is reasonable.

[16] Exercise: show that the maximum likelihood estimate for μ under this model is the sample mean. Recall that the density of the normal distribution with expectation μ and variance 1 is $\Phi(x; \mu) = 1/\sqrt{2\pi} \exp(-(x - \mu)^2/2)$.

[17] Please note that the value chosen, $\alpha = 0.95$, is arbitrary. We could as well have chosen other values (typical values are $0.90, 0.95, 0.99$). The important point is that this value is chosen *a priori*.

terms of the sum are normally distributed, $\bar{y}$ is also normally distributed. [18] Moreover, the expectation and the variance of $\bar{y}$ are μ and $1/n$, respectively. Therefore, subtracting μ from $\bar{y}$ and dividing by the square root of $1/n$ yield a random variable distributed according to a standard normal distribution, *i.e.*,

$$\frac{\bar{y}-\mu}{\sqrt{1/n}} = \sqrt{n}\,(\bar{y}-\mu) \sim N(0,1)\ .$$

The next ingredient we need for the calculation of CI is the notion of quantile. Let z_α be the number such that the probability that a normal random variable with mean 0 and variance 1 is less or equal than z_α is α. This number can be calculated by integrating the density probability of the standard normal distribution (you do not need to do that) and is called the *αth quantile* of the normal distribution with mean 0 and variance 1. Tables with values of the quantiles of the standard normal distribution are easy to find in the literature. Moreover, most of the statistical software have facilities to calculate the quantiles of the standard normal distribution. Therefore we will reduce the calculations below to calculations using the quantiles of the standard normal distribution.

Now we have at hand all the ingredients to calculate the confidence interval CI. I claim that using $a = z_{\frac{1+\alpha}{2}}/\sqrt{n} = a = z_{0.975}/\sqrt{120} = 0.1789194$ in the interval CI defined as in equation (2.15) will yield a confidence interval with coverage 0.95.

First we calculate the probability of $\bar{y} - a \leq \mu$,

$$P(\bar{y}-a \leq \mu) = P(\bar{y}-\mu \leq a) = P\left(\sqrt{n}(\bar{y}-\mu) \leq z_{\frac{1+\alpha}{2}}\right) = \frac{1+\alpha}{2}\ .$$

Therefore

$$P(\bar{y}-a > \mu) = 1 - \frac{1+\alpha}{2} = \frac{1-\alpha}{2}\ .$$

[18] Here we are using the fact that the sum of independent random normally distributed random variables is normally distributed; and that the product of a constant (e.g. $1/n$) is also normally distributed.

By symmetry of the normal distribution we have also

$$P(\bar{y} + a < \mu) = \frac{1-\alpha}{2} \ .$$

We conclude that

$$P(\mu \text{ is not in } CI) = \frac{1-\alpha}{2} + \frac{1-\alpha}{2} = 1 - \alpha \ ,$$

which implies that the interval CI given by equation (2.15 with $a = 0.1789194$ is a confidence interval with coverage 0.95. [19]

In summary, according to our calculations the data observed indicates that the parameter μ is between 2.686831 and 3.044669 with probability 0.95

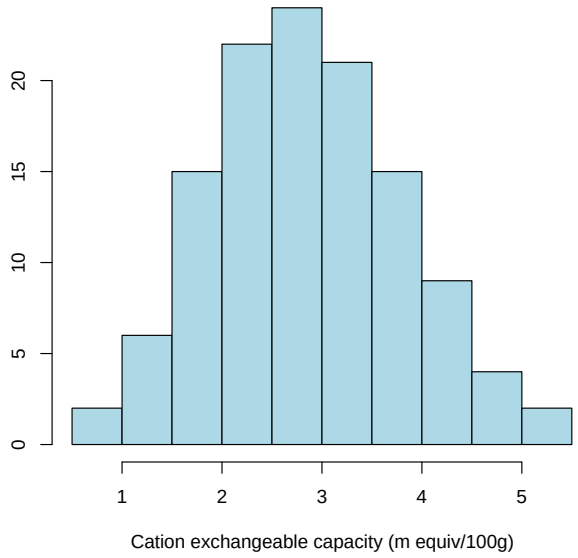

Figure 2.10: Histogram of 120 measurements of the CEC in soil.

[19] Exercise: calculate a confidence interval with coverage 0.99.

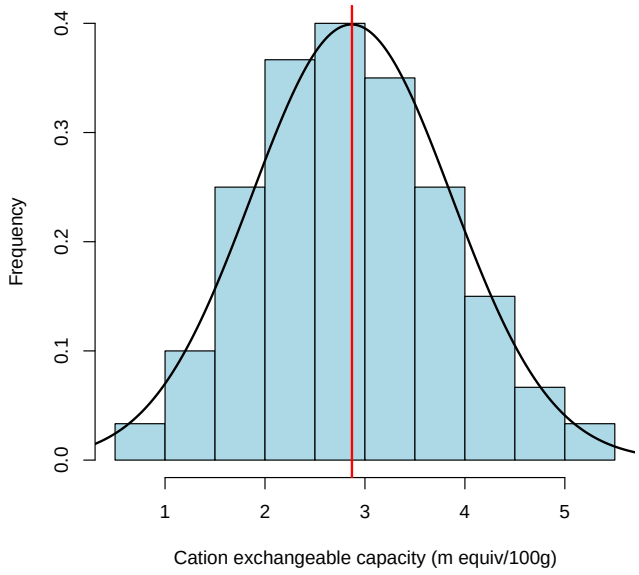

Figure 2.11: Histogram of 120 measurements of the CEC in soil and the probability density of a normal distribution with expectation 2.87 and variance 1. The vertical continuous red line represents the position of the sample mean.

2.3.2 Approximate confidence interval for a binomial experiment

In the example of the binomial experiment with two trials we showed that the maximum likelihood estimator for the parameter indicating the probability of success is the (sample) mean. This has three immediate consequences: First, according to the law of large numbers, the maximum likelihood estimator approximates to the probability of success when the sample size increase. That is, the maximum likelihood estimator approximates to the "true value" of the parameter of interest. Second, the expectation of the maximum likelihood estimator is equal to the probability of success. We say that this estimator is *unbiased*. The third consequence is that the maximum likelihood estimator is approximately normally distributed, for samples with sufficiently large sizes. This is an issue of the classic *central limit theorem*, a classic result of the probability theory.

More precisely, the *central limit theorem* (for sequences of independent and identically distributed random variables) says that: *"If $X_1, X_2, \ldots$ is a sequence of independent and identical distributed random variables with expectation μ and variance σ^2 then*

$$\frac{\sum_{i=1}^{n} X_i - \mathrm{E}\ (\sum_{i=1}^{n} X_i)}{\sqrt{\mathrm{Var}\,(\sum_{i=1}^{n} X_i)}} \tag{2.16}$$

is approximately normal distributed with mean 0 and variance 1, for large values of n. It is not difficult to see that the expression (2.16) is equivalent to

$$\frac{\sqrt{n}\,(\bar{x} - \mu)}{\sigma}, \tag{2.17}$$

where $\bar{x} = 1/n \sum_{i=1}^{n} x_i$.

We can use this result to calculate an approximate confidence interval for the parameter p as follows. Denote by $\bar{Y}$ the mean of the n independent results of the binomial experiment. Recall that the expectation and the variance of each observation are p and $p(1-p)$, respectively. Then, using the

central limit theorem with (2.17) we obtain that, for each $0 < p < 1$,

$$\frac{\sqrt{n}\,(\bar{x} - p)}{\sqrt{p(1-p)}}$$

is approximately normal distributed with mean 0 and variance 1. Replacing the parameter p by the estimate $\bar{x}$ we get the approximation [20]

$$\frac{\sqrt{n}\,(\bar{x} - \mu)}{\sqrt{\bar{x}(1-\bar{x})}} \approx N(0,1) \tag{2.18}$$

We will take advantage of the approximation 2.18 to construct a confidence interval for p. That is, we will calculate an interval such that the probability that the estimate $\hat{p} = \bar{x}$ belongs to this interval is approximately (greater or equal than) a pre-fixed value, say α.

The idea is to use here a procedure similar to the method used for the calculation of the confidence interval for the expected value of the CEC in the last section. The calculations will follow essentially the same steps, but this time they will be approximated and not be exact as there. We want to find an interval around the estimate $\hat{p} = \bar{x}$ such that the probability of the parameter p belong to the interval is approximately α, where α is a pre-fixed "high" value, typically 0.90 or 0.95. The interval can be written as $CI = [\bar{x} - a, \bar{x} + a]$ for a suitable number a. Here a should be a function of the sample (and of α) such that

$$P(p \text{ is in } CI) \approx \alpha\ .$$

I claim that choosing $a = z\left(\frac{1+\alpha}{2}\right)\frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}}$makes CI a confidence interval with coverage α. The following sequence of calculations will prove this claim. First,

$$P(\bar{x} - a \leq p) = P(\bar{x} - p \leq a) = P\left(\frac{\sqrt{n}}{\sqrt{\bar{x}(1-\bar{x})}}(\bar{x} - p) \leq z_{\frac{1+\alpha}{2}}\right) = \frac{1+\alpha}{2}\ .$$

[20] This is not the best approximation we can use here, but it yields a simple procedure suitable for our purposes. We will give another approach that is more efficient.

Therefore,

$$P(\bar{x} - a > p) \approx 1 - \frac{1+\alpha}{2} = \frac{1-\alpha}{2} \ .$$

Analogously,

$$P(\bar{x} + a < p) \approx \frac{1-\alpha}{2} \ .$$

Therefore,

$$P(p \text{ is not in } CI) = \frac{1-\alpha}{2} + \frac{1-\alpha}{2} = 1 - \alpha$$

and we conclude that

$$P(p \text{ is in } CI) = \alpha \ .$$

A better approximate confidence intervals is

$$\left[\left\{S + z^2_{\alpha/2} - z_\alpha\sqrt{[S(n-S)]/n + z^2_{\alpha/4}}\right\}, \left\{S + z^2_{\alpha/2} + z_\alpha\sqrt{[S(n-S)]/n + z^2_{\alpha/4}}\right\}\right] \ ,$$

where $S = n\bar{x}$. The approximation sed in this interval is better than the rough approximation used in (2.18), however the calculations involved a much more complex and definitely not suitable to introduce the idea of confidence intervals.

2.4 Hypotheses Tests

Here we will discuss the basic notion of hypotheses test. We will use a relatively simple example to do that.

Example 7 (The Master Quiz Game) *The so called "Master Quiz Game" is a game where there are three boxes, one of them containing a* **big** *check and the other two boxes are empty. You choose one box, but before you open the box the "Master Quiz" (i.e., a person that knows where the check is and is directing the game) says "*I give you a hint, the check is **not** here*" and he opens one of the remaining boxes, which (he know that) is empty. The Master-Quiz*

*continues: "*Would you like to change and choose the other closed box?*". Now the question is: Question: Is it advantageous to change? This question typically generates controversy, but I would claim that in most of the cases the answer presented is "No, it does't matter to change the boxes". Now, this answer is wrong; indeed there is a great advantage in changing the box, although this is counterintuitive. I will not enter in the details of showing this here (see exercise* **??***), but instead I have performed (together with students in a course) an experiment in which the master Quiz game was played several times, aways accepting the change of the box. The results of* 112 *of those trials were the following:*

```
0 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 0
1 1 1 0 1 1 1 0 0 1 1 0 1 0 1 1 0 1 0 0
1 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0
1 0 1 1 1 1 0 1 1 0 1 1 1 0 1 1 1 1 0 1
0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1
1 1 0 0 1 1 0 1 1 0 1 0
```

Here a "1" and "0" indicates, respectively, that the player received ("success") or not ("failure") the check in of the trials. That is, we observed 77 *successes in* 112 *independent trials. It is natural to assume that the number of successes in this experiment is binomially distributed. More precisely, if* X *is a random variable representing the number of successes, then we say that* $X \sim Bi(112, p)$*. Here* p *is the probability of "success" in one trial (i.e., getting the check), which is a unknown parameter. From the previous discussion, the maximum likelihood estimate of the parameter* p *is* $77/112 = 0.6875$*. The relevant question here is wether we have evidence that the probability of success is different than* $1/2$ *(if not, there would be no advantage in changing the boxes). But, can this result be explained by mere random fluctuation? We will build next a tool that allows to discuss this question: the so called "hypotheses tests" or "statistical hypotheses tests".*

Consider the following two hypotheses: $p = 1/2$ *and* $p \neq 1/2$*. We will denote these two possibilities by* $H_0 : p = 1/2$ *and* $H_A : p \neq 1/2$ *and call them*

the null hypothesis *and the* alternative hypothesis*, respectively. Note that in the way we formulated the two hypotheses there are only two possibilities: whether the null hypothesis H_0 is correct or the alternative H_A is correct; that is, if H_0 is true, then H_A is false, and if H_0 is false, then H_A is true, and there are no other possibilities. We want to decide, on the basis of the available data, which of these two hypotheses is correct. If H_0 would be true, then there would be no advantage in changing boxes.*

The general idea of (statistical) hypotheses tests is to make a rule, based on the available data (the results of the trials), to decide which of the two hypotheses is correct. One possible type rule would be:

"Reject H_0 when the relative frequency of successes is far from $1/2$,"

which is intuitively reasonable. A possibility would be to allow the number of successes to deviates only by one count of the number of successes from half of the number of trials (i.e., 56), which yields the rule:

Rule 1: "Reject H_0 when the number of successes
is smaller than 55 or larger than 57 ".

We call this rule a rejection rule*. If we apply this rule to our data, we would then reject the null hypothesis (since we observed* 77 *successes) and conclude that our data provide evidence that it is not indifferent to change the boxes. At this stage, it is natural to ask: what is the probability that this decision is wrong? To answer this basic question we calculate the probability of the number of successes, X, being smaller than* 55 *or larger than* 57 **when** $p = 1/2$ *(i.e., under the null hypothesis). This probability can be calculated as follows:*

$$\begin{aligned} P(\text{"Reject } H_0\text{"}) &= P(X < 55 \text{ or } X > 57) \\ &= P(X < 55) + P(X > 57) \\ &= P(X = 0) + P(X = 1) + \ldots + P(X = 54) + \\ &\quad + P(X = 58) + P(X = 59) + \ldots + P(X = 112)\,. \end{aligned} \tag{2.19}$$

Here the second equality above comes from the fact that the events $[X < 55]$ *and* $[X > 57]$ *are mutually exclusive. Each of the probabilities of the last equality in (2.19) can be calculated using the distribution function of a binomial distribution with size* 112 *and probability parameter* $1/2$*, that is*

$$P(X = x) = \frac{x!}{112!(112 - x)!}\,(1/2)^x(1 - 1/2)^{112-x}, \tag{2.20}$$

for $x = 0, 1, \ldots, 112$*. Figure 2.12 displays these probabilities. Now, inserting these probabilities in (2.19) yields that the probability of rejecting the null hypothesis when the null hypothesis is in fact correct is approximately* 0.777 *(see the upper left panel in Figure 2.14).*

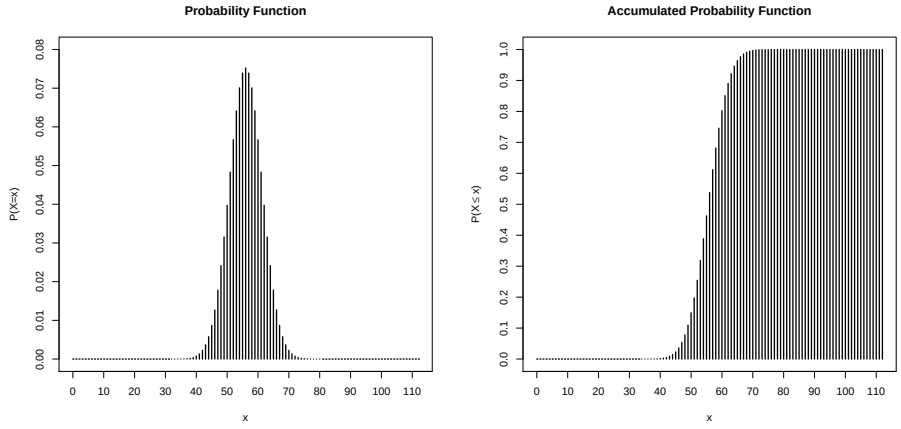

Figure 2.12: The probability function and the accumulated probability function of a binomial distribution of size 112 and probability parameter $p = 1/2$ (*i.e.*, $P(X = x)$ and $P(X \leq x)$, when $X \sim Bi(112, 1/2)$).

According to the rule 1 above, we should reject the null hypothesis ($H_0 : p = 1/2$*, since we observed* 77 *successes) and conclude that it is not indifferent to change the boxes. On the other hand, the evidence we are presenting against the null hypothesis is very weak, since with I high probability (0.777) we would have rejected the null hypothesis when the null hypothesis is correct.*

In other words, the rule 1 above is too strict, only allows a deviation of one success from what we would expect under the null hypothesis, and therefore in the cases were a slightly larger deviation occurs we would reject the null hypothesis even if the null hypothesis is correct. Therefore, we have to define a new less strict rule. A way to do that is to use the following rejection rule:

Rule 2: "Reject H_0 when the number of successes
is smaller than 54 or larger than 58 ".

That is, we accept deviations of at most two successes from the expected number of successes when the hypothesised probability of success is $1/2$*. An argument analogue to the calculation used in (2.19) yields (see the upper right panel in Figure 2.14)*

$$\begin{aligned} \alpha = P(\text{"Reject } H_0\text{"}) &= P(X < 54 \text{ or } X > 58) \\ &= P(X < 54) + P(X > 58) \\ &= P(X = 0) + P(X = 1) + \ldots + P(X = 53) + \\ &\quad + P(X = 59) + \ldots + P(X = 112) \approx 0.637\,. \end{aligned} \tag{2.21}$$

As expected, replacing the rejection rule by a less strict rule decreased the probability of error type I, but this probability is still rather high.

The idea of hypotheses test is to construct a rejection rule such that the probability of error type I is low, typically, taking a low pre-fixed value (e.g., $\alpha = 0.10$*, or* 0.05*, or* 0.01*)*[21]*. This construction is easy to be done in the test we are discussing because by increasing the amount of allowd deviations from the expected number of successes, we always decrease the probability of error of type I (since we remove positive parcels from a sum analogue to the sum in the last part of (2.20) or (2.21)). To see that, consider the following general type of rejection rule:*

General rule: "Reject H_0 when the number of successes
is smaller than $56 - k$ or larger than $56 + k$ ",

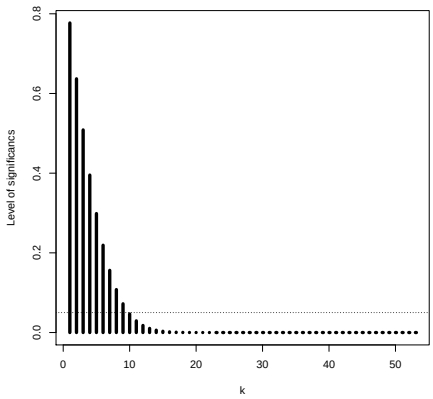

Figure 2.13: Probability of error of type I (α) as a function of the number of admitted deviations from the number of successes expected under the null hypothesis (k). The horizontal interrupted line indicates the probability of 0.05.

where k is an integer number (between 1 and 55) representing the number of deviations of the observed number of successes from the expected number of successes that we admit before we reject the null hypothesis. Figure 2.13 displays the value of the probabilities of error type I (α) for each value of k. Figure 2.14 displays the probabilities rejection rules and the probability of error type I (α) for some choices of k. If we choose k equal to 10, the probability of error type I becomes approximately 0.05 (in fact 0.047). Since we observed 77 successes, we might say that "we reject the null hypothesis when using a test with significance level of 5%" (i.e., 0.05). Now, since the probability of wrongly rejecting the null hypothesis is low when using the rejection rule defined with $k = 10$, we say that we have a strong evidence against the null hypothesis and we conclude that it does matter to change the boxes in the Master Quiz game. If we increase even more the value of k, the

[21] Please, note that there is nothing special with the numbers 0.10, or 0.05, or 0.01, these numbers are typically used in the literature just by an arbitrary convention.

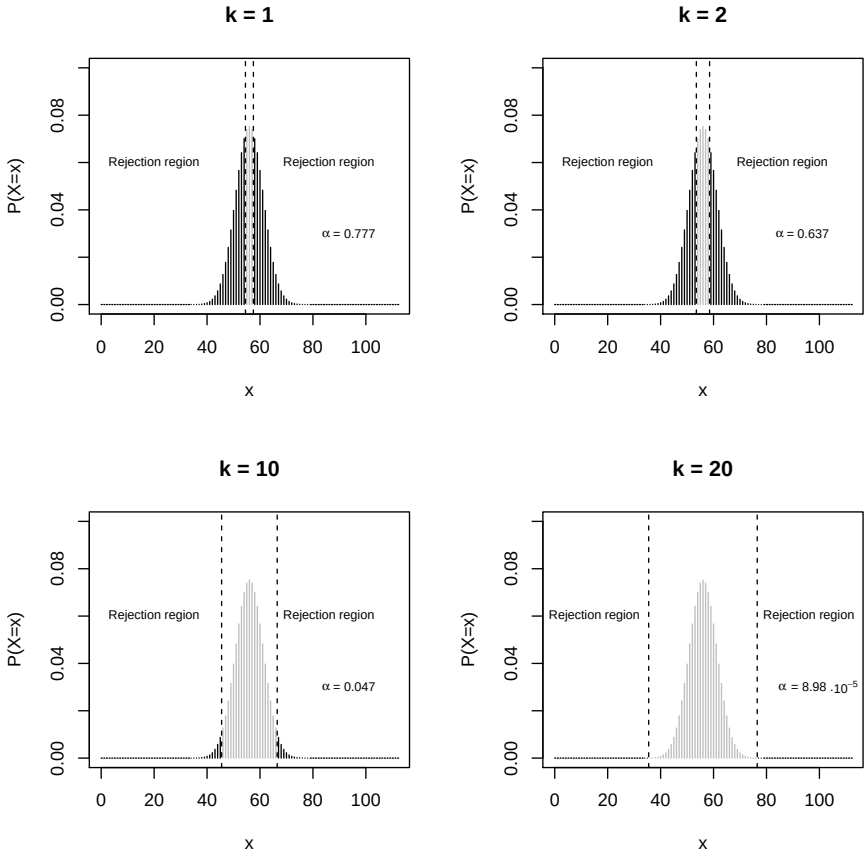

Figure 2.14: The probability function (*i.e.*, $P(X = x)$, when $X \sim Bi(112, 1/2)$) with the rejection regions indicated (in black) for four different rejections rules. The rejection rule obtained using the value $k = 21$ is the smallest rejection region for which the null hypothesis is rejected when the observed number of successes is 77. The probability of rejecting the null hypothesis in this case is $8.98\ .10^{-5}$, which is called the p-value of the test.

probability of committing a type I error decreases even more (see Figure 2.13). In this case, rejecting the null hypothesis would provide an even stronger evidence against the null hypothesis and our conclusion would be more forceful. One could then increase the value of k until we could still just reject the null hypothesis. In the example in question this would correspond to take k equal to 20 (using $k = 21$ would lead to the rule: "reject the null hypothesis when the number of successes is smaller than $56 - 21 = 35$ or larger than $56 + 21 = 77$", and we would not reject the null hypothesis, taking $k = 20$ would yield the rule "reject the null hypothesis when the number of successes is smaller 36 or larger than 77, an we would still reject the null hypothesis). Now, if we use a rejection rule constructed with k equal to 20, then the level of significance (i.e., the probability of the type I error) is $8.98\,10^{-5}$. This probability is called the p-value *for testing the null hypothesis. From the discussion below it is clear that the p-value indicates how strong the evidence that the current result provides against the null hypothesis in the sense that the smaller the p-value is, the stronger is the evidence against the null hypothesis. Since $8.98\,10^{-5}$ is small, we conclude that we have strong evidence that it does matter to change boxes in the Master Quiz game.*