Université Paris-Saclay – Masters BEE, BIP, BS

Biostatistics

P. Brabant, C. Dillmann, J. Legrand, D. Manicacci, E. Marchadier, S. Ollier, D. Sicard, D. de Vienne

2021-2022

Contents

Note

This is the English version of the biostatistics tutorial, translated by Malcolm Eden. The handout summarises the teachers' course notes for the biostatistics course units in M1 (Biology-Health, BEE, BIP) and M2 (optometry, ergonomics) in the Department of Biology from Université Paris-Saclay. It was written to accompany lessons, tutorials and practical work in these course units. Please send any questions or comments about improving this document to `christine.dillmann@universite-paris-saclay.fr`.

Chapter 1

Introduction: statistics, random variables, samples

1.1 Statistics

Statistics : *all the methods used to collect, describe and analyse observations (or data). These observations generally involve measuring one or more characteristics of a given set of individuals.*

Probability theory is the mathematical study of phenomena characterised by chance and uncertainty; *statistics* involves collecting, processing and interpreting sets of data. Probabilities and statistics make up the sciences investigating random phenomena.

Statistical unit : *the basic element measured.*

Units may be individuals, pairs of glasses (for example, when we need to test the quality of glasses frames), bacteria, etc. Statistical units may also be a group of individuals (e.g. a class of 25 students whose behaviour is being studied or *Arabidopsis thaliana* plants in a dish, etc.)

Population : *a set of individuals about whom we need to collect information*

This could be, for example, all adult men in the France, the pairs of glasses produced on an assembly line over a given period or all the patients in a hospital.
Generally speaking, when the population is very large, it is impossible to measure all the individuals in the population. We can, however, measure a small number of individuals selected at random from the population, considering them as representative of the population as a whole.
For populations of a smaller size, we can measure all individuals, and the sampled population will be described free of error.

Sample : *a subset of individuals in a population selected at random from all the individuals in the population. If they really are selected at random, without bias, the sample is said to be "representative" of the population.*

There are two broad kinds of statistical approaches, *descriptive statistics* and *inferential statistics.*

Descriptive statistics. The aim is to highlight the characteristics of a sample using quantified and graphical data. We might need to describe the distribution of variables, variable by variable. We could also be aiming to bring out the main characteristics of a sample, by summarising information using certain parameters (e.g. an average, the variance) and by determining the main relations between the variables. For example, we could give the average and the variance of the height of individuals in a sample. We could also measure the extent to which the height and weight of individuals are linked.

Inferential statistics. We might want to "infer" the properties of a given population (for example, the average height of the individuals in the population) based on data from a sample. Our aim is to estimate the average and also to establish a "margin of error". We may also want to test out hypotheses. For instance, does average height increase from one generation to the next in a given human population?

1.2 Random variables and probability laws

1.2.1 Random variables

Random variable : *a random variable (r. v.) is any variable with values depending on the outcome of a random phenomenon.*

The random variable is written as X, and an outcome of this random variable is given as x, which is a particular value taken by the variable in a random selection.

Examples of random variables

- We could study the result of throwing (non-loaded) six-sided dice. The random element comes from the way the dice are thrown. The variable is the value shown on the dice face (a whole number between one and six). When a die is thrown, each of the six values has a 1/6 chance of being shown.
- In a sample of butterflies of the *Biston bistularia* species captured in southern England, we could study whether the butterflies are light or dark in colour. The random variable here is the colour, which falls into two categories or modalities, *light* or *dark*. The probability of seeing a light-coloured butterfly is linked to pollution levels and the colour of tree bark.
- In a sample of individuals in the population of Iceland, we could study whether the blood group of individuals is $O+$. The random variable is an individual's blood group, which is divided into two categories, $O+$ or *other*. The probability that an individual is in the $O+$ group depends on genetic factors, especially on the history of the colonisation of Iceland.
- In a sample of ears of corn, we could count the number of grains on each ear of corn. The random element arises from the environment in which the plant is grown and its genotype. The possible values will be whole numbers, with wide variations between different ears of corn.
- We could measure the height of six-year-old girls beginning primary education and sampled in different schools. The random element comes both from the environment and from the genetic differences between the children. Height is a real value with a continuous variation.
- The common frog *Rana temporaria* is the most widespread frog in the *Rana* genus in Europe. It is often found in northern France and Belgium. It lives in all kinds of wetland: woodland (alongside forest paths, ponds), moors, valley pastures, farmland, parks and gardens. It is a poikilotherm species, that is to say, its body temperature adjusts to the surrounding temperature. We could measure the body temperatures of frogs sampled in a pond. The random element results from temperature variations on the surface of the pond, due to the position of areas shaded from the sun. Here, too, the temperatures measured will be real numbers with a continuous variation.

A random variable is characterised by :

1. *The values it can take, which are called the* ***support*** *of the random variable.*
2. *The probability of finding each value in the population or* ***probability law****.*

We can distinguish between several different categories of random variables:

Quantitative variables: numerical elements (size, age, etc.) obtained by measuring individuals. These elements are expressed in numbers when basic arithmetical operations (total, average, etc.) are meaningful. A quantitative random variable may be ***discrete*** (number of descendants of an individual, number of thoracic bristles in drosophila, etc.) or ***continuous*** (weight, size, etc.).

Qualitative variables: non-numerical characteristics. They can be ***nominal*** (such as eye colour) or ***ordinal***, when all categories can be classified (for example: little infected, moderately infected, highly infected). The different values of a qualitative random variable are called ***modalities*** or ***levels***. It should be noted that a qualitative variable may be coded in the form of a numerical value, but if we are dealing with a nominal variable, there is not point in carrying out mathematical operations on its values.

The table below summarises the categories of random variables for the examples given, and the relevant probability law.

Variable	Category	Support	Probability Law
Throwing dice	qualitative or quantitative	1, 2, ..., 6	$\{1/6, 1/6, ...\}$
Butterfly colour	nominal qualitative	light/dark	$\{p_c, 1-p_c\}$
Blood group in Iceland	nominal qualitative	$O+$/other	$\{p_{O+}, 1-p_{O+}\}$
Number of grains/plant	discrete quantitative	positive number	normal distribution
Girls' height	continuous quantitative	positive real	normal distribution
Temperature of frogs	continuous quantitative	positive number	normal distribution

1.2.2 Probability law

1.2.3 Qualitative random variable

For a qualitative random variable, we can draw up a list of all possible categories. Let X be a random variable with categories $\{a_1, a_2, ..., a_J\}$. We can calculate the probability $P(X = a_j)$ that X will have the value a_j $(j = 1, ..., J)$. The probability law of X is defined by all the $P(X = a_j)$.
An outcome x from X is a selection from the population. x will have a value among all the possible values. In other words, X could have the values a_1, or a_2, or a_3, etc., but x can only have a single value. Since the categories are mutually exclusive, we have:

$$P(X = a_i \quad \text{and} \quad X = a_j) = 0; \qquad i \neq j$$

It follows from the above that the probability that X belongs to one or other of the J categories is equivalent to 1.

$$\sum_{j=1}^{J} P\left(X = a_j\right) = 1$$

1.2.4 Discrete quantitative random variable

Discrete quantitative random variables are mainly count variables. For count variables, the support of the random variables are whole values: either all the whole numbers or a finite (e.g. 2,4,7,12) or infinite (e.g. all even numbers) series. We may list these values and call them, as above a_1, a_2, ...,.... We could associate a probability $P(X = a_j)$ with each value a_j. An outcome x from X can only produce one of these support values a_1,a_2,.... We should note that there can be discrete quantitative variables that do not produce whole values (e.g 0.56, 5.4, 3.2). Discrete quantitative random variables may have a finite or infinite support.

Example. Let X be the number of males among the 396 descendants of a Belgian Blue bull whose sperm is available from the breed catalogue. If we consider an individual descendant, there is a one in two chance for it to be male ($p = 0.5$). The total number of the bull's male descendants may be any whole value between 0 and 396. We can calculate the exact probability $P(X = a_j)$ for each whole value a_j in the whole set $\{0, ..., 396\}$ (by using the Binomial distribution $\mathcal{B}(396; 0,5)$; *cf.* chapter 2).
For a discrete quantitative random variable with a finite support, we also have:

$$\sum_{j=1}^{J} P(X = a_j) = 1.$$

We should note that if the support is infinite, we will have an infinite total (we replace J by ∞).
In a case in which the values a_1, a_2, ..., a_J are listed in order, we have:

$$P(X \leq a_k) = P(X = a_1 \text{ or } X = a_2 \text{ or } X = a_k) = \sum_{j=1}^{k} P(X = a_j).$$

To demonstrate this, we use the following property: if A and B are two disjoint events, then $P(A \quad \text{or} \quad B) = P(A) + P(B)$. For example, if $A : X = 0$ and $B : X = 1$, we have:

$$P(X \leq 1) = P(X = 0 \quad \text{or} \quad X = 1) = P(X = 0) + P(X = 1).$$

1.2.4.1 Continuous quantitative random variable

Continuous quantitative random variables are variables with values in $\mathbb{R}$, or, more often in biology, in an interval included in $\mathbb{R}^+$ (biometric measurements, concentrations, etc.). The probability of a random variable X (e.g. the rate of glucose in the blood) being precisely the value $x = 0.3846$ mg/L) is almost zero. However, we can calculate the probability $F(x)$ of X being smaller than a certain x value:

$$F(x) = P(X \leq x),$$

which is called the ***cumulative distribution function*** of X. The cumulative distribution function is used to calculate the probability that X will be found in an interval between x and $x + dx$: $P(X \in]x, x + dx]) = F(x + dx) - F(x)$.
The ***probability density function***, shown as $f(x)$, is the derivative of F:

$$f(x) = lim_{dx \to 0} \frac{(F(x + dx) - F(x))}{dx}.$$

Reminder: the integral function (written as $\int$) associates the area below the $f(x)$ curve with each interval $]a; b]$ for the x values in the interval $]a; b]$. It is written as $F_{a,b}(x) = \int_a^b f(x)dx$. If we derive $F(X)$, we again find the function $f(x)$.
Thus, the cumulative distribution function of the random variable X can be defined as the area below the curve f between $-\infty$ and x (figure 1.1). This is the probability that X will be smaller than x:

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$

It should be noted that the total area below the density is equivalent to 1 (this is the sum of probabilities):

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

For the normal probability laws, the cumulative distribution functions $F(x)$ are set out in table form.

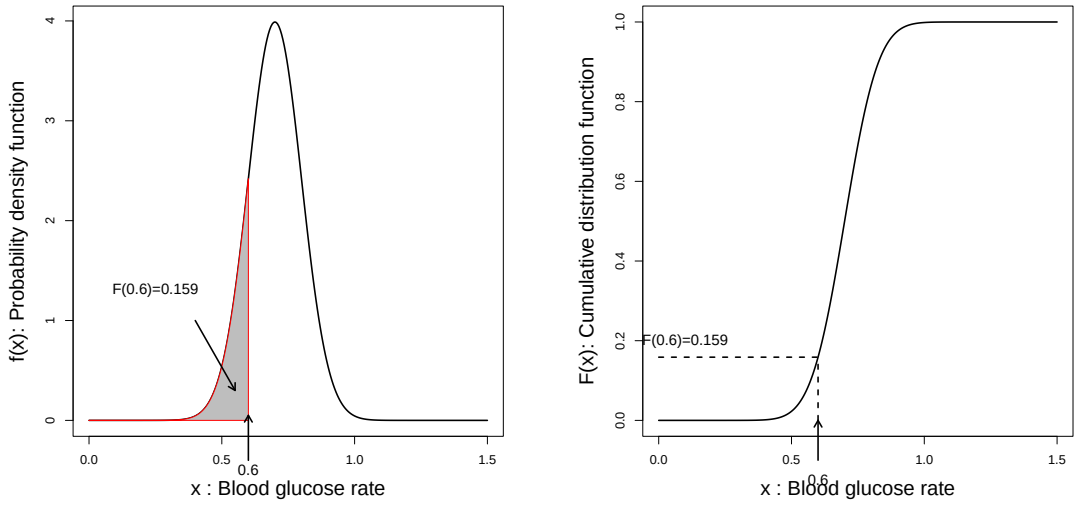


Figure 1.1: **Probability law of a random variable.** In this example, X is the random variable associated with the measurement of blood sugar on an empty stomach. $F(0.6)$ is the probability of this value being smaller than 0.6. To the left is the probability density function and to the right the cumulative distribution function.

A more intuitive way of understanding the link between the cumulative distribution function $F(x)$ and the probability density function $f(x)$ is to consider an n-sample $(X_1, ..., X_n)$ of the random variable X,

with n being very large. We can make a chart of the x_i values, grouping them into J classes of Δx size. The j class corresponds to the x_i values, which belong to the interval $]x_{min} + (j-1)\Delta x; x_{min} + j\Delta x]$. We can define the probability $P(X \in]x; x + \Delta x])$ and estimate it, as is the case for a discrete variable, by the frequency of the class $P_{\text{obs}}(]x; x + \Delta x])$ observed in the sample. If the number of observations is very large, we can then define a very large number of classes of a very small size, dx, and draw up a continuous function that "enfolds" the empirical distribution of X: this is the probability density function of X.

Quantile. For each value x, we can calculate the cumulative distribution function $F(x)$. We might also want to look at the value of x, so that $F(x)$ will have a certain value. We call the α quantile, written as q_α, the value such that

$$F(q_\alpha) = \alpha.$$

It statistical tests, the quantiles $q_{0,025}$, $q_{0,05}$, $q_{0,95}$ and $q_{0,975}$ are often used.
The **median** of a distribution is by definition the quantile at 50%: half of the population has a value below the median.

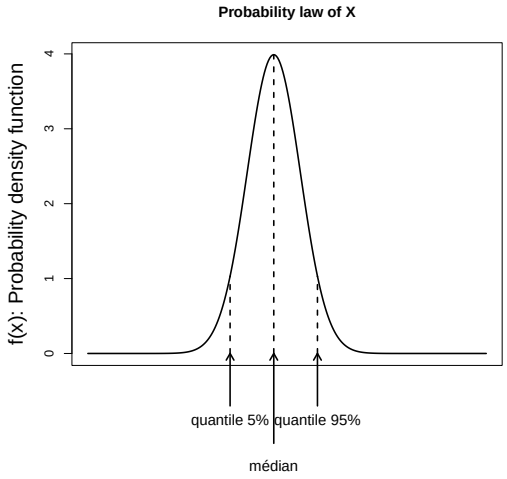


Figure 1.2: **Quantiles in a normal distribution.**

1.2.5 Expected value and variance

The two most widely used parameters to characterise the probability law of a random variable are expected value and variance.

Expected value (centering parameter): this is the value that we would expect to find, on average, if we repeated the same random experience infinitely. It is written as $E(X)$ and is expressed as the "expected value of X". It corresponds to a weighted average of the values that this variable can take:

- *For discrete random variables,* $E(X) = \mu_X = \sum_j (P(X = a_j) \cdot a_j)$
- *For continuous random variables,* $E(X) = \mu_X = \int_{-\infty}^{+\infty} (xf(x)dx)$.

We should note that if we study two random variables, X and Y, the expected value of the sum of the two random variables is equal to the sum of the expected values.

$$E(X + Y) = E(X) + E(Y)$$

Centering variable: a variable is centred when we remove the expected value from each of the values of X. $X_{\text{centered}} = X - E(X)$. Consequently, we can show that $E(X_{\text{centred}}) = 0$.

Variance (dispersion parameter): in statistics and in probability theory, variance is a measurement used to characterise the dispersion of a variable. It shows how the statistical series or the random variable is scattered around its average or its expected value. A variance of zero shows that all the values are identical. A small variance is a sign that the values are close to each other, whereas a high variance is the sign that they are far apart. We define variance as a weighted sum of squared deviations from the mean.

- *For discrete r.v.,* $V(X) = \sigma_X^2 = \sum_j \left[P(X = a_j) \cdot (a_j - E(X))^2\right]$
- *For continuous r.v.,* $V(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} \left[(x - E(X))^2 \cdot f(x)dx\right]$

We can notice that the variance is expressed as the expected value of $(X - E(X))^2$:

$$\begin{aligned} V(X) &= \sigma_X^2 \\ &= E\left[(X - E(X))^2\right] \\ &= E(X^2) - [E(X)]^2 . \end{aligned} \tag{1.1}$$

In this last form, which is useful for making calculations, we can see that the variance is the difference between the expected value of the square of the variable and the square of the expected value of the variable.
It is also important to notice and to remember that the variance of a variable X multiplied by a constant k is:

$$V(kX) = k^2 V(X)$$

and that the variance of a variable to which a constant k is added is:

$$V(X + k) = V(X).$$

Example: estimating the number of boys (X) in a family with two children. If the probability of having a boy is $1/2$, then the law of probability of X is written as:

a_j	$P(X = a_j)$
0	1/4
1	1/2
2	1/4

The expected value of X is equal to $E(X) = (0/4 + 1/2 + 2/4) = 1$, and the variance $\sigma^2 = \frac{1}{4}(0-1)^2 + \frac{1}{2}(1-1)^2 + \frac{1}{4}(2-1)^2 = 1/2$

Standard deviation: when looking at a continuous probability distribution, what we can visualise is not the variance directly (which is expressed in the square of the unit of measurement). The parameter that is expressed in the unit of measurement is the square root of the variance, which is called the **standard deviation**:

$$\sigma = \sqrt{\sigma^2} = \sqrt{E\left[(X - E(X))^2\right]}.$$

Standardised variable: A variable divided by its standard deviation is called a standardised variable. As a result, the variance of a standardised variable is equivalent to one.

1.2.6 Pair of random variables

A pair of random variables : (X, Y) *is a pair of random variables whenever two different measurements* X *and* Y *are observed in the same statistical individual during a probability experiment.*

When we carry out different measurements (e.g. height and weight at birth) on the same statistical individual, the results may depend on each other. For example, it is obvious that, for biological reasons (allometry), we expect there to be a positive relation between height and weight in human beings at birth. So, two random variables measured in the same individual are not necessarily independent of each other.

Notion of covariance : *We define the covariance between two quantitative random variables* X *and* Y *as* $cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$.

Notion of independence : *We consider that there is independence between two random variables* X *and* Y *when, irrespective of the value taken by* X*, the law of* Y *does not change, and vice versa. From this definition, the result, for discrete random variables, is:* $P(X = a \quad \text{and} \quad Y = b) = P(X = a) \cdot P(Y = b)$ *and for continuous random variables* $f(x, y) = f(x) \cdot f(y)$.

Variance of a sum. For any pair of random variables X and Y, the variance of $X + Y$ is written:

$$V(X + Y) = V(X) + V(Y) + 2cov(X, Y),$$

If X and Y are independent, then $V(X + Y) = V(X) + V(Y)$.
Please note, independence implies zero covariance. However, the opposite is not always the case.
The covariance of *standardised* variables X and Y is:

$$\rho_{X,Y} = cov\left(\frac{X}{\sqrt{V(X)}}, \frac{Y}{\sqrt{V(Y)}}\right) = \frac{cov(X, Y)}{\sqrt{[V(X)V(Y)]}}. \tag{1.2}$$

This is the Pearson **correlation coefficient**, a measurement of covariance going beyond the units in which X and Y are expressed.
It is important to note that the covariance of a variable with itself is simply its variance: $cov(X, X) = V(X)$. We can deduce the boundaries of ρ_{XY}. If X and Y are strictly linked ($X = Y$), we have

$$\rho_{XY} = \frac{cov(X, X)}{\sqrt{V(X)V(X)}} = 1.$$

If X and Y are strictly anti-correlated ($X = -Y$), we have

$$\rho_{XY} = \frac{cov(X, -X)}{\sqrt{V(X)V(X)}} = -1.$$

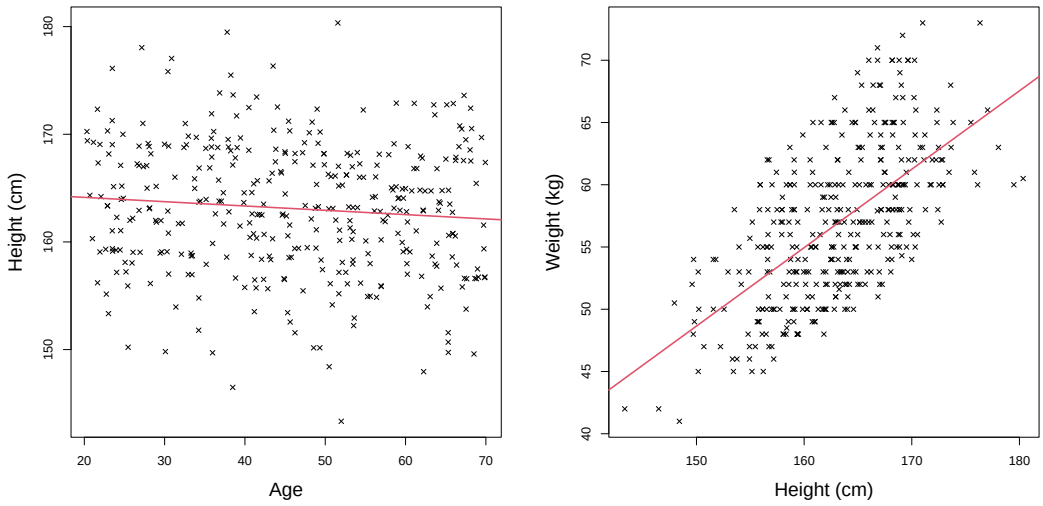


Figure 1.3: **Pairs of quantitative random variables.** We measure the age, height and weight of a sample of 350 adult women aged between 20 and 70. Each dot represents a woman. On the left: the relation between age and height. There is a slight negative dependence, the correlation is -0.04. On the right: the relation between height and weight in the same individuals shows a strong positive link, with a coefficient of correlation of 0.63.

1.3 Notion of n-sample

When a given population is very large, we cannot measure all individuals. Instead, we use a subset of the population to make inferences concerning information about the population.

***n*-sample**: a subset of n individuals selected at random and independently from the reference population. We call X_i the random variable associated with the selection of the individual i ($i = 1, ..., n$), and x_i the value observed in the individual i. The way the sample is constituted (independent random selection) means we can make the hypothesis that the random variables X_i are independent and part of the same law. If the selection really is made at random, without bias, then the sample will be considered as "representative".

x_i observations can be used to make hypotheses concerning the shared law of the X_i, that is the law of X.

Example: Using a sample of 350 women in the graph (1.3), we can ask whether the observed variables are independent. Based on this sample, we can also estimate the average and the variance of the observed variables. This will be studied in chapter 1.5.

1.4 Descriptive statistics

We can study the outcomes $(x_1, ..., x_i, ..., x_n)$ of the random variables X_i from an n-sample of the reference population.
We can produce a visual representation of the x_i by grouping them by class of values. We can also use the measurements to describe the distribution of the x_i (for example, average or variance). We call these measurements *summary statistics*.

1.4.1 Qualitative variable

If we study a qualitative variable, then the x_i can only have a finite number of categories or modalities $(a_1, ..., a_j, ..., a_J)$.

Size : *size observed in each category*

Relative frequency : *size of each category compared with the total number of individuals* n.

We can show the relative sizes or frequencies in a bar chart.

1.4.2 Quantitative variable

Discrete random variable. If we study a discrete r.v., we can calculate the n_j number of each value a_j in the sample or its relative frequency $\frac{n_j}{n}$ and create a *bar chart*.
For example, the table below is from an INSEE survey showing the number of children per woman, for women born between 1961 and 1965, as well as the frequency of each possible case (zero, one, two, three or more children):

Number of children per woman	0	1	2	3	> 3
Frequency (%)	13.5	18.2	38.9	20.2	9.2

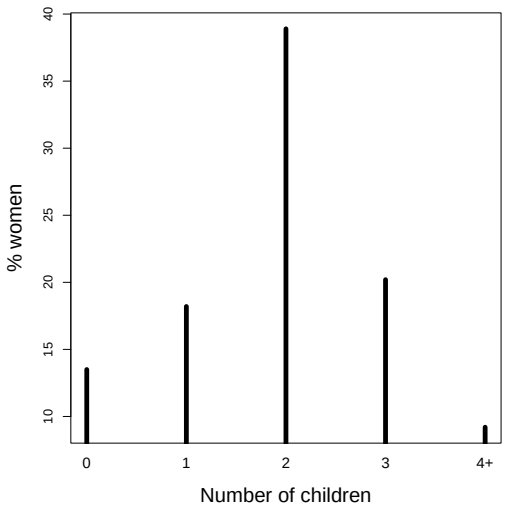


Figure 1.4: **Bar chart.** Number of children per woman.

Continuous random variable. If we study a continuous r.v., the x_i are all different, but we can group them together by class, by defining the adjacent intervals of the values. We then count the numbers observed in each interval. The **histogram** or empirical distribution shows the size of each class. Similarly, each class can be represented by its frequency. In a histogram, on the y axis, we see either the size, frequency or value, such that the area of the bar is equal to the frequency. The histogram may vary in visual terms, depending on the boundaries of the classes chosen.

Mode of an empirical distribution : *this is the class including the most individuals in the sample. The mode may vary according to the boundaries of the classes chosen.*

It is worth noting that some distributions may have 2 or more peaks, and in this case we use the terms bimodal or multimodal distribution. This is the case, in particular, when individuals from the sample belong to different populations (figure 1.5).

Quantiles. If we group the x_i values in ascending order, we can easily calculate the proportion of observations below a certain value a, $F_{\text{obs}}(a)$.
The ***median*** is the value a such that $F_{\text{obs}}(a) = 0.5$, that is, half of the x_i have a value below the median and half of the x_i have a higher value.
The quantile α is the value a such that $F_{\text{obs}}(a) = \alpha$. We often use quantiles at 5% and 95%, resulting in an interval containing 90% of the sample. The quantiles at 25% and 75% result in an interval containing half of the individuals in the sample.

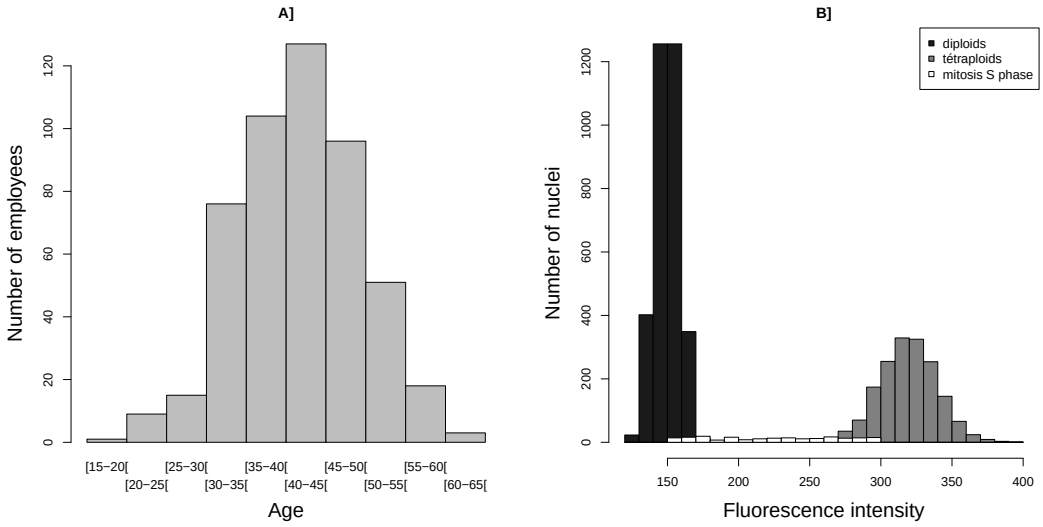


Figure 1.5: **Histograms.** Continuous data are grouped into classes. A] Pyramid of ages among bank employees. The mode corresponds to the class aged 40 to 45. B] The fluorescence intensity of the cell nuclei in corn leaves measured with flow cytometry. We can distinguish three populations of nuclei, which differ in the degree of fluorescence, pointing to different levels of ploidy. In black, diploid cell nuclei. In grey, the tetraploid cell nuclei. In white, cell nuclei in phase S of mitosis.

An alternative to the histogram is the box-and-whisker plot or boxplot (1.6), which uses the quantiles of the empirical distribution. With this representation we can show the distributions of several different samples in the same graph.

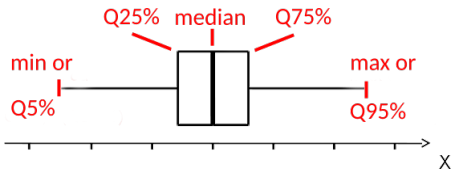


Figure 1.6: **Box-and-whisker plot or boxplot.** A common representation of the empirical distribution of a sample. The box corresponds to the interval defined by the quantiles 25% and 75%. The limit of the horizontal bars has a variable meaning depending on the software used and may represent either the minimal/maximal values observed in the sample or values dependent on quartiles. In the latter case, the observations of the sample beyond these limits are shown as dots or crosses. The vertical bar shows the median of the empirical distribution.

Mean and variance. We can use the analogy with expected values and the variance of a random variable to calculate the mean and the variance of a sample.
The mean of the sample is an indicator of position:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

It is often close to the mode, that is, to the most common value.
The variance of the sample gives an indication of the scope of variations around the mean. It is an indicator of dispersion:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

For a pair of random variables, covariance gives an idea of the linear relation between two variables:

$$s_{xy_n} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right].$$

1.5 Inferential statistics

We will now move from the sample to the population. There is a quantitative random variation, and we will use a sample to form an idea of or to *estimate* the descriptors of the distribution of the random variable, such as the average or variance, for example.
Each x_i is an outcome of a random variable X_i. The X_i values are independent and have the same law, characterised by their expected value $E(X_i) = \mu_X$ and variance $V(X_i) = \sigma_X^2$.

1.5.1 Estimation of the population mean

We can define the following random variable:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which is a calculation of the mean of the selection. To do so we use the properties of the sum of expected values of several different random variables:

$$E(X+Y) = E(X) + E(Y)$$

So it is easy to find that $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu_X$, since each X_i has the same expected value μ_X.

Unbiased estimator : *an unbiased estimator is a random variable with an expected value that is exactly equal to the quantity we want to estimate.*

So we can say that $\overline{X}$ is an unbiased estimator of μ_X, which is written as $\widehat{\mu}_X$. By using the outcomes x_i from the X_i in a sample, we can thus provide the empirical mean of the sample as an estimation of the population mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The estimator of μ_X is $\overline{X}$, it is a random variable. The estimation of μ_X is $\overline{x}$, it is an outcome of the random variable $\overline{X}$.

Estimator accuracy. The fact that an estimator is unbiased gives no indication of its accuracy. To gauge its accuracy, we can ask how the estimator varies around its mean, that is, we calculate the estimator's variance.
For the variance of the estimator $\overline{X}$ of the mean, if we recall that X_i values are independent and have the same law, then we have:

$$\begin{aligned} V(\overline{X}) &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma_X^2 \\ &= \frac{\sigma_X^2}{n}. \end{aligned}$$

The bigger the size of the sample, the lower the variance of the estimator. It should be noted that accuracy is measured through standard deviation and so decreases by $1/\sqrt{n}$ with n. To increase accuracy by a factor of 10, we need to increase the size of the sample by a factor of 100.

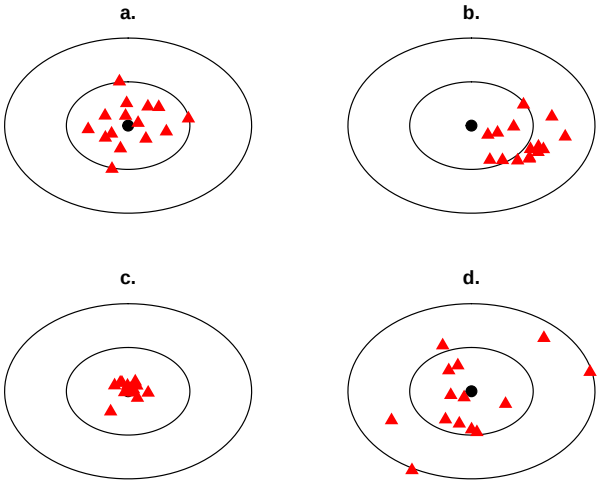


Figure 1.7: **Properties of an estimator.** The black dot is the parameter to be estimated in the population. Each red triangle is an estimation from an n-sample. **a.** Unbiased estimator. The red dots are on average close to the parameter. **b.** Biased estimator. The estimation average differs from the value of the parameter. **c.** Unbiased and accurate estimator. The red dots are centered on the parameter and have low variance. **d.** Unbiased but inaccurate estimator. The estimator has a large amount of variance.

1.5.2 Estimation of the variance

By using the property of the sum of the expected values, we have

$$E\left(\sum_{i=1}^{n}(X_i-\mu_X)^2\right)=n\sigma_X^2$$

If μ_X is unknown, then we cannot use it to estimate the variance, so we will replace μ_X with its estimator $\overline{X}$, which will modify the expected value. So we can write:

$$\begin{aligned}
E\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) &= E\left(\sum_{i=1}^{n}\left[(X_i-\mu_X)-(\overline{X}-\mu_X)\right]^2\right)\\
&= E\left(\sum_{i=1}^{n}(X_i-\mu_X)^2-n(\overline{X}-\mu_X)^2\right)\\
&= n\sigma_X^2-\sigma_X^2\\
&= (n-1)\sigma_X^2. \qquad (1.3)
\end{aligned}$$

So, as an unbiased estimator of the variance, we can use:

$$S_{X_{n-1}}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,$$

which is also written as $\widehat{\sigma}_X^2$.
By using the x_i outcomes of the X_i in a sample, we can estimate the variance of the population, which is written as s_X^2:

$$s_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Corrected empirical variance : *s_X^2 is called the corrected empirical variance of the sample. It is an estimation of the variance of the population.*

Corrected empirical covariance : *similarly, s_{XY} is called the empirical covariance of the sample. It is an unbiased estimation of covariance in the population.*

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}((x_i - \overline{x})(y_i - \overline{y}))$$

1.6 Summary

We want to know the parameters of the law of distribution of a random variable or a pair of random variables in a given population. We study an n-sample of individuals from the population. It is possible to calculate a certain number of **summary statistics** in this sample. We can also use the sample to make inferences about the parameters of the population. We use the values measured in the sample to calculate **estimations** of the parameters of the population:

	Population	**Estimator**	**Estimation**
	r.v. X and Y	r.v. $X_1, ..., X_n$, independent and with the same law	observations $x_1, ..., x_n$
	.	r.v. $Y_1, ..., Y_n$, independent and with the same law	observations $y_1, ..., y_n$
Mean	$E(X) = \mu_X$	$\overline{X}$	$\overline{x}$
Variance	$V(X) = \sigma_X^2$	$S^2_{X_{n-1}}$	s_X^2
Covariance	$Cov(X,Y) = \sigma_{XY}$	S_{XY}	s_{XY}
Correlation	ρ_{XY}	$\hat{\rho}_{XY}$	r_{XY}

with:

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
s_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})] \\
r_{XY} &= \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}.
\end{aligned}$$

Chapter 2

Laws and tests

2.1 Normal distributions

2.1.1 Case of discrete quantitative variables

For a discrete variable X with a support $a_1, ..., a_J$, the distribution is given by the probability that the variable will have the value a_j, $P(X = a_j)$, for any value of j.
The sum of all the $P(X = a_j)$ is equal to 1. We can always represent the normal distribution in a bar chart (figure 2.1).

2.1.1.1 Bernoulli distribution

We consider a variable with two possible outcomes (for example, heads or tails, wearing glasses/not wearing them, man/woman, segregation of dominant monogenic traits in F2-generation). We can always give the codes 0 and 1 to the two outcomes. For example, we can write $X = 1$ for tails and $X = 0$ for heads.
The law of the variable is given by $P(X = 1) = p$ and $P(X = 0) = 1 - p$. We then consider that X follows a Bernoulli distribution, written as $\mathcal{B}(p)$. Its expected value is $E(X) = p$ and its variance is $V(X) = p(1 - p)$ (see the equation (1.1)).

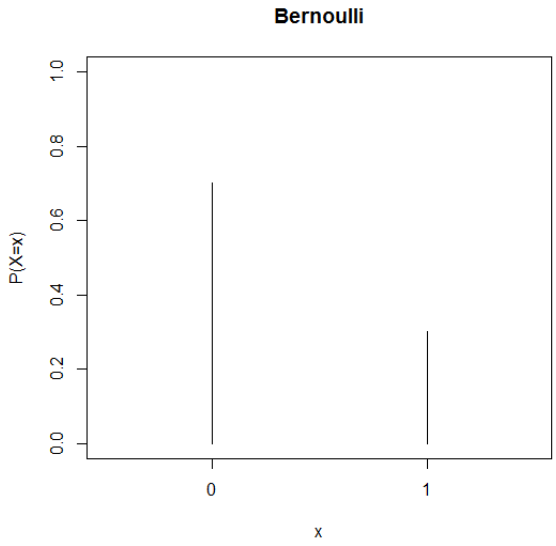


Figure 2.1: **Bar chart** showing the probability law of a Bernoulli distribution variable $p = 1/4$.

2.1.1.2 Binomial distribution

Imagine that we repeat the Bernoulli process n times. We then have n independent Bernoulli variables X_i for $i = 1$ at n, an n-sample. We count the number of times the result is tails, that is, we produce a new variable Y equal to the sum of the X_i values.

$$Y = \sum_{i=1}^{n} X_i$$

The r.v. Y follows a binomial distribution of parameters n, the number of times the coin is tossed, and p, the probability of the coin showing tails. The binomial distribution is written as: $\mathcal{B}(n; p)$.

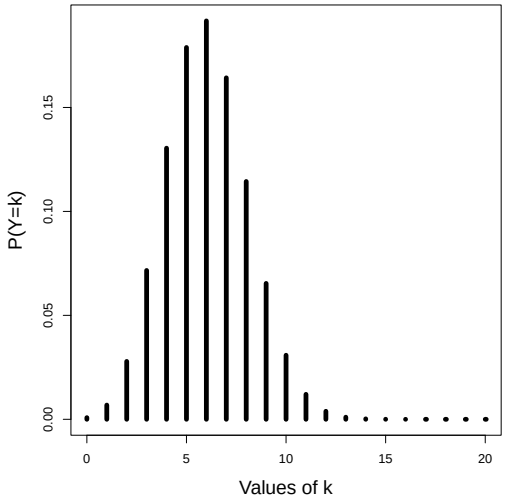


Figure 2.2: **Probability law of a Binomial variable.** $Y \sim \mathcal{B}(20; 0, 3)$.

The Y law is given by $P(Y = k)$ for $k = 0, ..., n$.

$$P(Y = k) = C_n^k p^k (1 - p)^{n-k},$$

where C_n^k is the combination of k among n.
The expected values of Y is

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np$$

and its variance is

$$V(Y) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = np(1 - p).$$

All the count variables from independent events follow Binomial laws. A nominal characteristic of an individual can always be described as a Bernoulli variable (whether the characteristic is present or not). When we look at the number of individuals having the characteristic in an n-sample, we are dealing with a Binomial distribution with the support $\{0, 1, ..., n\}$.

Estimation of *p*. Often we do not know the probability of success p, that is, the probability for an individual in the randomly chosen population to have the characteristic. We try to estimate it using observations in the sample.
From the law of Y we can deduce the law of $Z = \frac{Y}{n}$, which describes the proportion of success for n independent draws. Z takes the values $\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}\}$.

$$P\left(Z = \frac{k}{n}\right) = P(Y = k) = C_n^k p^k (1 - p)^{n-k}.$$

We can calculate the expected value of Z:

$$E(Z) = E\left(\frac{Y}{n}\right) = \frac{E(Y)}{n} = p$$

and its variance:

$$V(Z) = V\left(\frac{Y}{n}\right) = \frac{V(Y)}{n^2} = \frac{np(1 - p)}{n^2} = \frac{p(1 - p)}{n}.$$

We can note that Z is an unbiased estimator of p, and that the variance of the estimator decreases with the size of the sample n. We can thus propose the estimation:

$$\widehat{p} = \frac{y}{n}$$

where y is the number of individuals observed to have the characteristic in the sample.

Example: we want to find out the probability p of men aged over 40 suffering from back pain. We carry out a survey on 20 men aged over 40. Seven say they suffer from back pain. We can thus estimate the probability by

$$\widehat{p} = 7/20 = 0.35$$

.

2.1.1.3 Poisson distribution

Let there be a random variable Y that can take discrete values 0, 1, 2, 3, ..., $+\infty$ (in theory). If Y follows a Poisson distribution, we have

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

and

$$E(Y) = V(Y) = \lambda$$

For a Poisson distribution, the mean and the variance are the same.

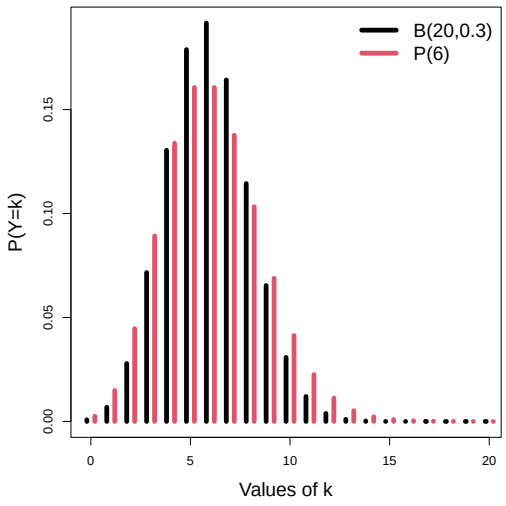


Figure 2.3: **Probability laws for a Binomial variable and a Poisson variable with the same expected value $E(Y) = np = \lambda = 6$.** For the Poisson distribution (in red), the probability of extreme events (to the right or to the left) is higher, leading to a bigger variance.

The Poisson distribution is sometimes also called the law of rare events because if we have a random variable X following a binomial law of parameters that is very large, n, and very small, p ($n \geq 50$, $p \leq 0.1$ and $np < 15$), we can approximate this Binomial law to a Poisson distribution with the parameter np. For example, the number of accidents at a crossroads on a given day. If a large number of cars pass through it every day and the probability of an accident involving a car each time is low, then the number of accidents will follow a Poisson distribution.
In this handout, when we study a random variable X, if we know its distribution, we show it with

the sign $\sim$ and if it is an approximate distribution we will show with the sign $\approx$. For example, if $X \sim \mathcal{B}(n;p)$ with n and p fulfilling the conditions listed above, $X \approx \mathcal{P}(np)$.
If we compare two random variables with the same mean, one following a Binomial law (mean np), and the other a Poisson distribution (mean $\lambda = np$), we can see that the Binomial random variable will have a smaller variance because $np(1-p) < np$ (see figure 2.3).
When p becomes very small and n is big, we have $np(1-p) \approx np$, which justifies approximating the Binomial distribution by a Poisson distribution.

2.1.2 Case of continuous quantitative variables

There are many probability laws enabling us to *model* a continuous quantitative variable. The best known is the normal distribution. However, not all continuous quantitative variables follow a normal distribution. We will see other distributions in this course.
We should recall that a continuous variable X is described by its density function $f(x)$ or by its cumulative distribution function $F(x) = P(X \leq x)$.

2.1.2.1 Uniform distribution

The uniform distribution is characterised by the following property: all the intervals of the same length included in the distribution support have the same probability. As a result, the associated density function is a constant.
The density function of a uniform distribution $\mathcal{U}(a,b)$ whose support is the interval $[a,b]$ is

$$f(x) = \frac{1}{b-a}$$

and the distribution function $F(x)$ is a linear function:

$$F(x) = \frac{x-a}{b-a}.$$

The expected value is $E(X) = (a+b)/2$, and the variance is $V(X) = \frac{(b-a)^2}{12}$.

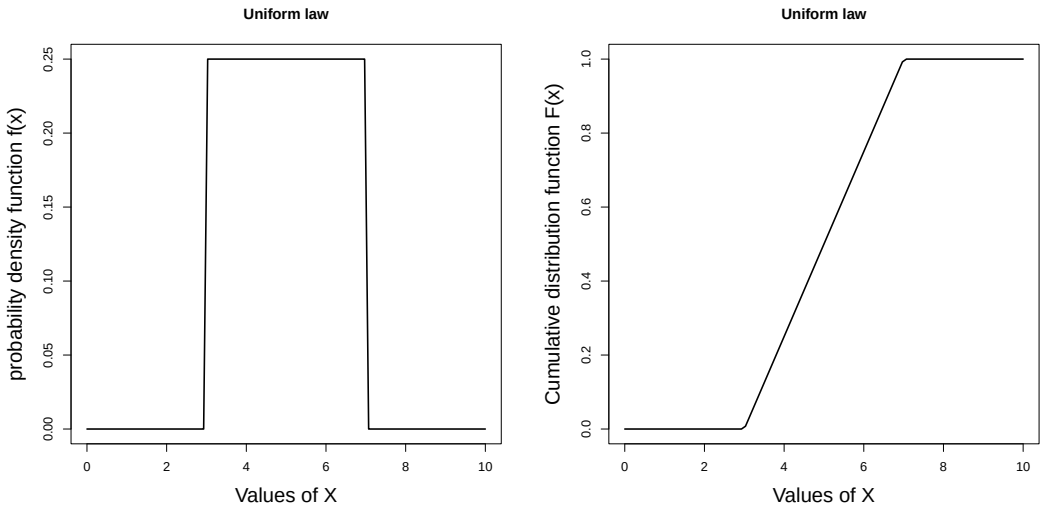


Figure 2.4: **Example of a uniform distribution $\mathcal{U}(3,7)$.** The density function is shown on the left, the distribution function on the right.

2.1.2.2 Gaussian distribution

A variable X that follows a normal distribution (or Gaussian distribution) with a mean of μ and a variance σ^2 has the following density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Standard normal distribution. This distribution corresponds to the distribution of a random variable following the expected value $\mu = 0$ and variance $\sigma^2 = 1$. We can always go from a normal variable $X \sim \mathcal{N}(\mu; \sigma^2)$ to a standard normal distribution variable by means of a transformation. Thus, the standardised variable $X - \mu$ follows the law $\mathcal{N}(0; \sigma^2)$, and the standard normal variable $\frac{X-\mu}{\sigma}$ follows the law $\mathcal{N}(0; 1)$.
As a result, only the standard normal distribution is generally included in statistical software.

Which variables are often modelled as Gaussian random variables?

- Numerous biological variables, such as weight, height, etc.
- A sum of independent random variables following normal distributions follow a normal distribution. So, in particular if X_1, ..., X_n are Gaussian variables independent of the same distribution $\mathcal{N}(\mu; \sigma^2)$, then their mean is a normal (Gaussian) distribution $\mathcal{N}(\mu; \frac{\sigma^2}{n})$.
- Let X_1, X_2, ..., X_n be independent variables in the same distribution, with a mean μ and a variance σ^2. We want to find the mean of X, $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$. Whatever the distribution of the X_i, the $\overline{X}$ distribution *tends* towards a normal distribution when n tends towards infinity. This is the **central limit theorem**. In practice, if n is quite large (we often take $n \geq 30$), we can make the hypothesis that $\overline{X}$ follows a normal distribution asymptotically with a mean μ and variance $\frac{\sigma^2}{n}$. This will be written $\bar{X} \approx \mathcal{N}(\mu; \sigma^2/n)$ and not $\bar{X} \sim \mathcal{N}(\mu; \sigma^2/n)$
- A direct application of the central limit theorem is the approximation of the binomial distribution to a normal distribution. The binomial distribution is a sum of Bernoulli variables with the same distribution and which are independent. So, if Y follows a binomial distribution $\mathcal{B}(n; p)$ and $n > 30$, $np > 5$ and $n(1-p) > 5$, then Y approximately follows $\mathcal{N}(np; np(1-p))$.

There are distributions other than the normal distribution to describe the quantitative variable distribution, such as the exponential distribution of the gamma distribution. For example, the random variables measuring the time before an event takes place often have exponential probability distributions, with a density function given by:

$$f(x) = \lambda e^{-\lambda x}$$

2.2 Test principle

Take a random variable X following a known distribution (for example, a Poisson distribution or a normal distribution) but with unknown parameters. We can estimate these parameters or try to locate them in a range of values or else compare them with a reference or between them. We then speak about **hypothesis testing**.

Examples

- Is the population of mainland France representative of the French population as a whole?
- Does an operation on cataracts affect long-sightedness?
- Are male and female swallows' tails the same size?
- Is the joint angle of workers' arms during installation of waterproof covering on building sites larger than INRS recommendations?

One difficulty is to choose the right test. We need to identify the question asked and the nature of the variables we are studying, in some cases to validate complementary hypotheses *a posteriori*, etc. Here, using a simple example, we will set out an approach that can be applied in all hypothesis testing.

2.2.1 How to carry out a statistical test

Imagine we want to check if some equipment used to measure the focal length of contact lenses is correctly adjusted. The manufacturer says that the accuracy of measurements is such that $\sigma = 0.1$. We have a benchmark with a known focal distance at a value of 16 cm. We carry out $n = 10$ measurements of focal length of the benchmark using the equipment.

2.2.1.1 Model

We will call X the measurement of the focal length of the benchmark. X is a random variable, since, as the manufacturer points out, if we repeat the measurement, we will not have exactly the same result, since this depends on the equipment's accuracy. We have a 10-sample test of independent measurements X_i. If the equipment is correctly adjusted, we can expect each X_i to follow a normal distribution $\mathcal{N}(16; \sigma^2 = 0.01)$. If the equipment is incorrectly adjusted, the expected value of X_i should be different from 16. This is what we want to find out. We will call μ the unknown expected value of X_i and we put forward the following model: X_i follows a normal distribution $\mathcal{N}(\mu; 0, 01)$. This model presumes that, for the time being, we will trust the manufacturer about the equipment's accuracy ($\sigma = 0.1$ is assumed to be known).

2.2.1.2 Hypotheses H0 and H1

In this test, we can make two mutually exclusive hypotheses:

- H0: expectation that the focal distance measured is 16 ($\mu = 16$)
- H1: expectation that the focal distance measured is different from 16 ($\mu \neq 16$)

Please note that the hypotheses always refer to a feature of the probability law of the random variable. We can make no hypotheses about the outcome (x_1, ..., x_n). Even if H0 is true, there is little chance that the mean of the sample $\overline{x}$ will be exactly equal to 16.
The principle of the test is to find a random variable:
- for which we can calculate an outcome based on observations
- for which we know the distribution under H0.
- for which the distribution under H1 (even if we do not know it) is different.
We call this random variable **the test statistic**. H0 is called the **null hypothesis**: this is the hypothesis we want to test, which will be refuted or confirmed. In general, it is a quantified hypothesis, that is to say, we know the value of the parameters and we can thus find the distribution of the test statistic. H1 is called the "alternative hypothesis".

2.2.1.3 Choosing a test statistic

To choose the test statistic we can return to the model. We know that if the X_i values follow a normal distribution $\mathcal{N}(\mu; 0.01)$, then $\overline{X}$ follows a normal distribution $\mathcal{N}(\mu; 0.01/10)$, and $\overline{X}$ is an unbiased estimator of μ. We can calculate an outcome of $\overline{X}$ based on a sample: $\overline{x}$. However, the distribution of $\overline{X}$ is not known, since the mean μ is unknown. So, as a test statistic we can put forward the random variable $Z = \frac{\overline{X}-16}{0.1/\sqrt{10}}$, where the distribution under H0 is known. Since we know that

$$\frac{\overline{X} - \mu}{0.1/\sqrt{10}} \sim \mathcal{N}(0; 1).$$

So, since under H0, $\mu = 16$, we have:

$$Z = \frac{\overline{X} - 16}{0.1/\sqrt{10}} \sim_{H0} \mathcal{N}(0; 1).$$

We can also calculate an outcome of Z in the n-sample:

$$z = \frac{\overline{x} - 16}{0,1/\sqrt{10}}.$$

The random variable Z is therefore a promising candidate as a test statistic for testing H0 vs H1. We call $P_{\text{H0}}(z) = P_{\text{H0}}(Z \leq z)$ the distributive function of the random variable Z under the H0 hypothesis. If H0 is true, then Z really is a random variable that depends only on $\overline{X}$, since the other parameters are constant. We can note the outcome of Z in a sample as z_{obs}.

2.2.1.4 H0 rejection region and choice of risk α

Definition of a rejection region We can compare z_{obs} to the theoretical distribution of Z under H0. If H0 is true, we can expect that most of the outcomes of Z will be close to 0 ($\overline{X}$ close to 16). It is unlikely that Z will be very large as an absolute value. If we find a high value for z_{obs}, as an absolute value, we will decide that there is little chance of observing it under H0, and we will reject H0. **We define the rejection region as an interval:**

$$]-\infty; -z_{\text{threshold}}] \cup [z_{\text{threshold}}; +\infty[.$$

This interval, shown in figure 2.5, contains the test statistic values that have little chance of occurring if hypothesis H0 is true. To find out where the rejection region is situated, we need to study the location of the distribution of the test statistic under H1: to the right of the distribution under H0? To its left? To the right or the left?

Choice of risk α Even if H0 is true, the probability that Z is in the rejection region is not zero. There is thus a certain probability of making an error by rejecting H0, although H0 is correct, which is called the **type I error**, written as α. The principle of tests is that we set α *a priori*, that is *before* carrying out the experiment. We will say we are ready to accept the risk α of making a mistake by rejecting H0 when H0 is in fact correct. By writing as $P_{\text{H0}}()$ the probability of an event under the hypothesis that H0 is correct, we can write:

$$\alpha = P(\text{rejecting H0} \mid \text{H0 true}) = P_{\text{H0}}(\text{rejecting H0})$$

For the test given above, we have: $\alpha = P_{\text{H0}}(Z > z_{\text{threshold}} \quad \text{or} \quad Z < -z_{\text{threshold}})$
In practice, we often choose $\alpha = 5\%$ or $\alpha = 1\%$. In writing, α is the probability shown by the grey areas under the density function of the test statistic under H0 (see Figure 2.5), in the rejection region of H0. So we look for the corresponding $z_{\text{threshold}}$ value. It should be noted that the $-z_{\text{threshold}}$ is the quantile of the order $\alpha/2$ in the Z distribution, and the $z_{\text{threshold}}$ is the quantile in the $(1-\alpha/2)$ order. The rejection region is calculated by looking for the quantiles in the Z distribution. Here, we look for the $z_{\text{threshold}}$ such that:

$$F_{\text{H0}}(z_{\text{threshold}}) = 1 - \alpha/2.$$

The inverse functions of the distribution functions of the normal distributions are available in most statistical and spreadsheet software. In our example, by setting a risk $\alpha = 5\%$, we have $z_{\text{threshold}} = 1.96$ (*function R* : `qnorm(p=0.975)`). The rejection region for the H0 hypothesis is thus here:

$$]-\infty; -1,96] \cup [1,96; +\infty[.$$

If z_{obs} is in the rejection region, we reject H0 and choose H1 at the risk of type I error α. However, if z_{obs} is not in the rejection region, we cannot reject H0. But we cannot automatically choose H0 since we have not checked the risk of type II error in wrongly accepting H0.

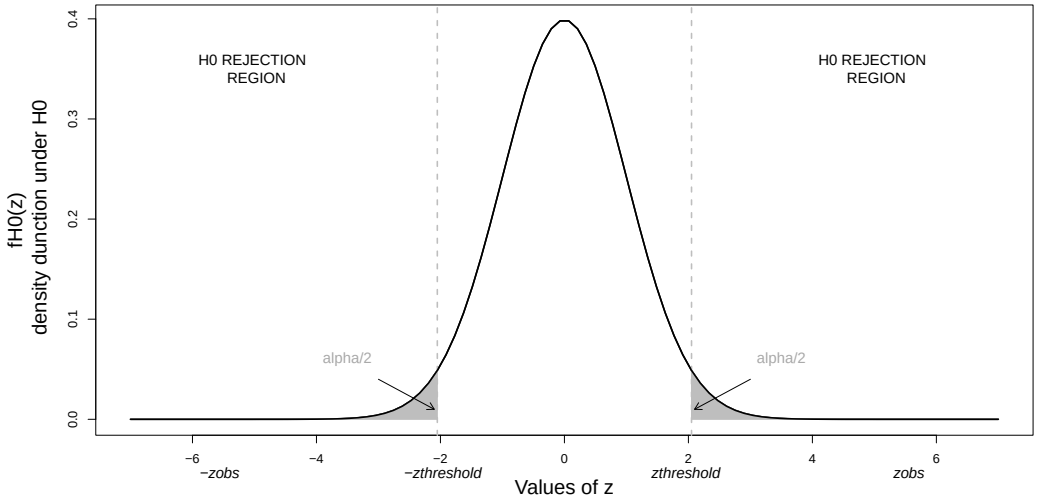


Figure 2.5: **Illustration of the rejection region.** The black curve is the density function of Z under H0. The rejection region of the hypothesis H0 is defined as an interval that is calculated so that the area below the curve in the interval of the rejection region is equal to the chosen threshold α, called *alpha* in the graph.

2.2.1.5 Carrying out the test: calculating the observed value of the statistic

Up to now, we have not used the numerical values of the sample. Carrying out the test involves calculating z_{obs} on the basis of the sample.
The values of the 10 measurements are as follows: 16.03 15.89 16.33 16.44 16.07 16.41 16.12 15.94 16.45 16.08.
We can calculate the mean of the sample: $\overline{x} = 16.176$, and we know $n = 10$ and $\sigma = 0.1$. Thus we have

$$z_{\text{obs}} = \frac{16.176 - 16}{0.1/\sqrt{10}} = 5.56.$$

Bearing in mind the error in measurement stated by the manufacturer, if the machine is well adjusted, it is unlikely we will arrive at "large values" (in absolute value). The principle of the test is to reject H0 beyond a given value threshold by selecting the type I error α. If we choose $\alpha = 0.05$, the threshold is 1.96. Since 5.56 is larger than 1.96, we are in the rejection region, that is, we are rejecting the H0 hypothesis. In this case, we can conclude that the equipment is probably badly adjusted.

2.2.1.6 *p*-value

To finish the test, we calculate the p-value (*pval*) or the observed level.

p-value : *This is the probability that the test statistic is beyond (in the rejection region) z_{obs} under H0. In other words, it is the value of α that we should have chosen so that a limit of the rejection region should be z_{obs}. If this probability is lower than the chosen risk α, we will reject H0. The p-value varies between 0 and 1.*

$$pval = P_{\text{H0}}(|Z| > |z_{\text{obs}}|)$$

Here z_{obs} is positive, so $pval = 2P_{\text{H0}}(Z > z_{\text{obs}}) = 2.70 \cdot 10^{-8}$ (*function R:* `2*(1-pnorm(q=5.56))`). This p-value shows that the probability of finding an empirical mean beyond 16.176 under the hypothesis H0 was extremely low.
In the figure 2.6 we can see the $z_{\text{threshold}}$ and z_{obs} values found in our example. We can also see that **when $pval < \alpha$, then z_{obs} is located in the test's rejection region**.

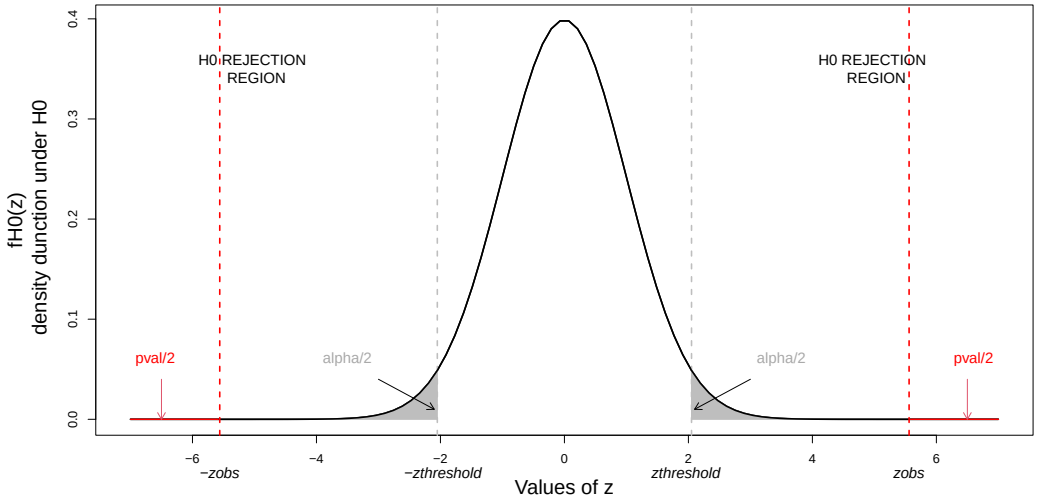


Figure 2.6: **Illustration of the notion of *p*-value.** The black curve is the density function of Z under H0. The p-value is defined as the area below the curve in the interval of values of z such that $|Z| > |z_{\text{obs}}|$.

Carrying out the test. If $z_{\text{obs}} = z_{\text{threshold}}$, then $pval = \alpha$. When z_{obs} moves away from zero, the p-value decreases, becoming smaller and smaller. Outside the rejection region, we always have $pval > \alpha$. So there is another way of carrying out a statistical test:

- Choose the risk α
- Calculate z_{obs}
- Use z_{obs} to calculate the p-value
- If $pval < \alpha$, we reject H0, if $pval > \alpha$, we cannot reject H0.

Today, this is what most statistical software does.

2.2.1.7 Conclusion

The test conclusion comes in two parts. The statistical conclusion consists in deciding whether to accept or reject H0. The biological conclusion requires us to come back to the original question. When we reject H0, we can say that there is a statistically significant difference at the level α.

2.2.2 Type I and II risks, power

When carrying out a statistical test, we do not know the real situation. We do not know if the hypothesis H0 is true or false. So, when we take a decision after the test, we may be mistaken in two different ways:

- We might reject H0 when in fact it is true. We can say there is a *mistaken rejection* of H0, or false discovery, or a *false positive*. The probability of rejecting H0 when it is true is called a **type I risk**, written as α. This risk α, which is also known as level α and sometimes threshold α, is always chosen *a priori*.

$$\alpha = P\,(\text{rejection of H0} \mid \text{H0 true})$$

- We might retain H0 although it is false. It is a false negative. The probability of a false negative is called a **type II risk**, written as β. For H1, we can have a simple hypothesis (for example $\mu = 18$) but, in general, we have a composite hypothesis (in our example $\mu \neq 16$).

2.2.2.1 Notion of power

In the case of a simple hypothesis, we define power by

$$1 - \beta = 1 - P\left(\text{non rejection of H0} \mid \text{H1 true}\right).$$

This is the probability of rejecting H0 if H0 is false. In the case of a composite hypothesis, we define the power function with a value depending on the real value of the parameter.
In general, β cannot be calculated, since we do not know the statistical test law under H1. It depends on μ, which is unknown. In the previous example, the test statistic is calculated as $Z = (\overline{X} - 16)/\frac{0.1}{\sqrt{10}}$. If H0 is false, then $E\left(\overline{X}\right) = \mu \neq 16$. The expected value of Z is thus:

$$E(Z) = \frac{\mu - 16}{0.1/\sqrt{10}}$$

On average, Z will be further away from zero to the extent that μ is different from 16. Here, the power of the test increases when μ is further away from 16.

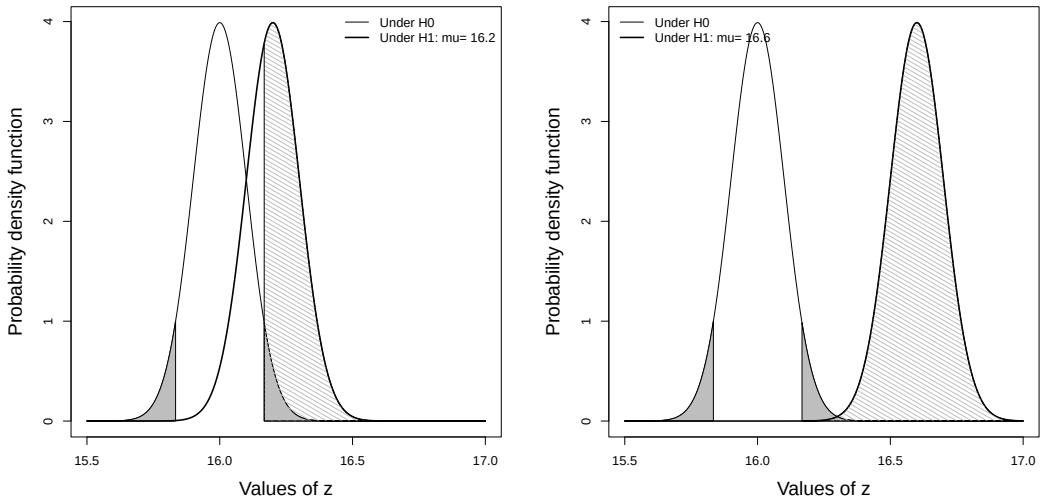


Figure 2.7: **Power of a statistical test.** The fine line is the distribution of X under H0 ($\mu = 16$). The bold line has two examples of the distribution of X for the real value of μ ($\mu = 16.2$ to the left, and $\mu = 16.7$ to the right). For these tests, the alternative hypothesis is H1: $\mu \neq 16$. In each graph, the grey areas show the level of the test. These areas correspond to the area below the density under H0 in the rejection region of H0. The hatched area below the density for the real value of μ is the power of the test. We can see that the power increases when there is a gap between the real value of μ and the supposed mean under H0. We use the term "power function", since the power depends on μ whose value is unknown.

More generally, we can see that the statistics in this test to find the mean can be written:

$$Z = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{\sigma}$$

As noted above, we can expect that Z will fluctuates around zero under H0. If we are under H1, then the expected value of Z is

$$E(Z) = \sqrt{n}\frac{(\mu - \mu_{\text{H0}})}{\sigma} \neq 0.$$

Thus, the power of the test will be stronger to the extent that μ is different from μ_{H0} (figure 2.7). However, for a given value of μ, this mean value will be bigger to the extent that the size of the sample n is large and the variance σ^2 is small. The experimenter cannot check μ. *On the other hand, we can choose to increase the power of the test by increasing the size of the sample or by carrying out*

experiments in the most homogeneous conditions possible, to decrease the variance. A power graph shows the relation between the size of the sample and the power of the test, for the different values set for the other parameters (figure 2.8).

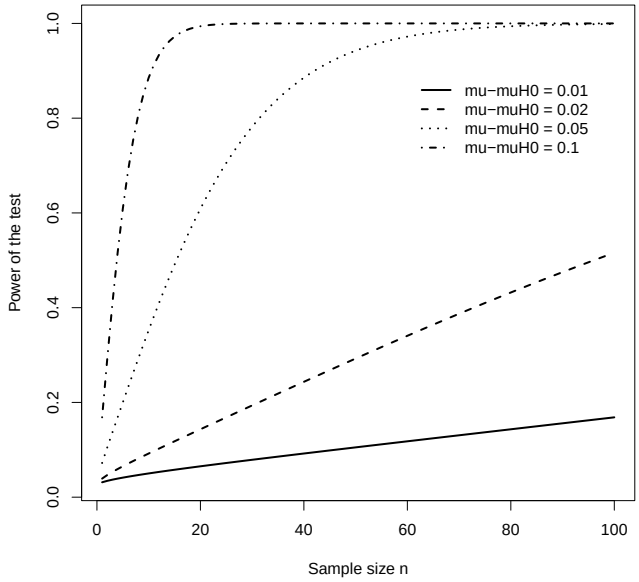


Figure 2.8: **Power curves.** We can represent the power of the test according to the size of the sample for different values of the mean difference $(\mu - \mu_{\text{H0}})$, for $\sigma = 0.1$. For a sample with the size $n = 20$, we are practically sure of finding differences of 0.1 units. We need a sample of size $n = 80$ to have the same chance of detecting differences of an average of 0.05, whereas a difference of 0.01 will only be detected in about 10% of cases.

Experimental planning is a sub-discipline of statistics enabling us to draw up (depending on the question asked and the constraints linked to experiments) experiment plans to maximise the power of the tests.

2.2.2.2 Unilateral test versus bilateral test

In our example, we could have an alternative hypothesis that is not a difference but an unequality:
H0: on average, the focal distance measured is 16 cm
H1: on average, the focal distance measured is higher than 16 cm

There are thus two ways to express the hypothesis H1:

- **Bilateral test:** when we have no *a priori* idea about the value of the parameter, we choose

$$\text{H1} : \mu \neq \mu_{\text{H0}}.$$

  In this case, we will reject H0 both for the positive and negative values of the test statistic.

- **Unilateral test:** when we know *a priori* that the value of the parameter cannot be larger or smaller than the value chosen under H0. We then define

$$\text{H1} : \mu > \mu_{\text{H0}}$$

  in the first case and

$$H1 : \mu < \mu_{H0}$$

in the second. We will reject H0 for the positive and negative values, respectively, of the test statistic.

We should note that whatever the form of H1, the hypothesis H0 is always the same and consists in putting forward one or several numerical values for the parameters of the model. So, in a unilateral test to the right ($H1 : \mu > \mu_{H0}$), it is not possible to test the hypothesis H0 $\mu < \mu_{H0}$ because we cannot quantify the test statistic. So, we choose the most "unfavourable" hypothesis H0, that is, the one that has the least chance of being rejected, namely $\mu = \mu_{H0}$.
Choosing the H1 hypothesis defines the shape of the rejection region of the test. Some tests only have one possible formulation of the H1 hypothesis. (figure 2.9).

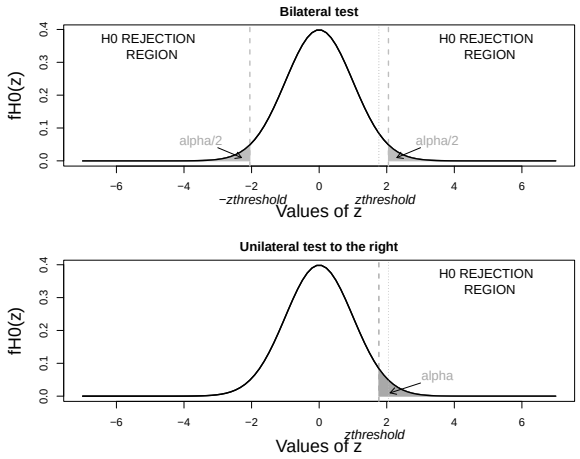


Figure 2.9: **Rejection region in the case of a bilateral test and a unilateral test (to the right).** The black curve is the density function of Z under H0. The grey areas correspondent to the risk α. For each type of test, the dotted lines show the limits of the rejection region. For the unilateral test, the rejection region is only on the right and the zthreshold limit is slightly to the left of the upper limit for the bilateral test.

2.2.2.3 How to carry out a statistical test

To sum up, we have just seen, with the help of an example, the general approach to follow in carrying out any statistical test. This approach can be divided into seven stages:

1. Choice of a model: which variable(s) will be studied? What is its law (their laws)? How can we translate the question in terms of the model's parameters?

2. Formulation of the H0 and H1 hypotheses.

3. Choice of a test statistic and determination of its law under H0.

4. Choice of the type I risk α (also called level) and definition of the rejection region

5. Calculation of the observed value of the statistic

6. Calculation of the p-value

7. Statistical (rejection or non-rejection of H0) and biological (answer to the question asked) conclusion.

The only stages requiring anything beyond a basic knowledge in statistics are stages 1 and 3.
The modelling stage consists in determining the parameters influencing the distribution law of the observations. In this course we will see a number of standard cases that you could use in most cases you will come across. More generally, the modelling stage consists in describing how the experimental data are produced. If you know how to simulate an experiment, then you will be able to model it.
The stage of choosing a test statistic is a field of research in itself. We need to find **a random variable that takes into account the experimental data and that can be calculated only on the basis of these data, for which the law under H0 is known**. The choice of a test statistic depends closely on the model. Here again, during the course we will see the appropriate test statistics for the most common models.

2.3 Multiple tests

2.3.1 The problem linked to carrying out multiple tests

It often happens that from a single sample, we carry out the same kind of statistical test, with the same H0 hypothesis, on a large number of different random variables on the same individuals or we make several tests on a single variable observed over a large number of samples.

Examples

- We carry out a genome scan on a sample of a population of inbred lines to determine whether there is segregation distortion in the population. For a large number of them ($K = 1500$) we have the positions in the genome (for example, SNPs) of the observed frequency of two parental alleles in the sample. For each of the K positions, we will test the H0 hypothesis that there is no segregation distortion, that is, that the frequency of the parent 1's allele is 0.5. We conduct 1500 statistical tests based on the same sample of lines.
- INSEE's health survey collects biometric variables (height, weight, gender), as well as social indicators (salary, socio-professional category, occupation). The list of occupations includes $n = 125$ categories of different occupations. We want to know if it is possible to rank the occupations by the average height of the people working in them. To do so, we will compare the average height in each occupation, that is, we will make $K = n(n-1)/2$ comparisons 2 to 2.

When we carry out K statistical tests at the level α, we will expect to reject H0 wrongly (false positive) with a probability α for each test.

Number of false positives: In the case of K tests at the level α, V is the number of wrongly rejected cases of the H0 hypothesis. V is a random variable. If H0 is true for all tests, then $V \sim \mathcal{B}(K;\alpha)$. If K is very large, we can have quite a high number of false positives. For example, if we carry out 100 tests at the threshold 5%, we will expect on average $100 \times 0.05 = 5$ false positives (if H0 is true for all the tests).

2.3.2 Methods for controlling the global α risk or the false positive rate

2.3.2.1 Bonferroni correction

We can notice that if we decrease the threshold of the test, the number of false positives will necessarily decrease. So, if we carry out 100 tests at the 5‰ threshold and not 5%, and the H0 hypothesis is true for all tests, then we will expect on average $100 \times 0.005 = 0.5$, that is, zero or a single false positive.

Bonferroni correction: : *the Bonferroni correction in the case of multiple K tests consists in carrying out each test at the threshold α/K.*

We should note that when the threshold of a test decreases, it becomes more difficult to reject the H0 hypothesis, even if it is false. To this extent, the Bonferroni correction is highly conservative.

2.3.2.2 *False Discovery Rate*

The reasoning behind the Bonferroni correction is that the H0 hypothesis is true for all tests. In reality, the H0 hypothesis may be false for some tests. We thus need to consider four possible situations:

- **False positive:** V is the number of cases where H0 is rejected when it is true,
- **False negative:** T is the number of cases where H0 is retained when it is false,
- **True positive:** S is the number of cases where H0 is rejected when it is false,
- **True negative:** U is the number of cases where H0 is retained when it is true,

with $K = V+T+S+U$. The rejection rate is $\frac{V+S}{K}$ and the rejected hypotheses include false positives and true positives.

False Discovery Rate : $FDR = V/(V+S)$, *is the proportion of cases where H0 is wrongly rejected among the case where H0 is rejected.*

Probability law of the p-value of a test under H0 We should recall that a **test statistic** is a random variable associated with an n-sample, which is calculated from X_i variables in the n-sample. For example, we have seen that, if X is a Gaussian random variable with expected value μ and variance σ^2, then the statistic $Z = \frac{\overline{X}-\mu_{\text{H0}}}{\sigma/\sqrt{n}}$ follows a law $\mathcal{N}(0;1)$ under H0.
A statistical test consists in defining the H0 and H1 hypotheses such that the test statistic:

- can be calculated in figures from an outcome of the n-sample,
- follows a law of known parameters under the H0 hypothesis.

In the example we have been studying (compliance test on the mean with known variance), we know μ_{H0} and σ, and we can calculate

$$z_{\text{obs}} = \frac{\overline{x} - \mu_{\text{H0}}}{\sigma/\sqrt{n}}$$

from the sample. We can also calculate the p-value associated with the sample:

$$p_{\text{obs}} = P_{\text{H0}}(|Z| > |z_{\text{obs}}|).$$

If we carry out another experiment, we will find a different value for z_{obs} and p_{obs}. So, the *p-value associated with a statistical test is a random variable*, which we can call P.
The probability law of P is easy to calculate. The support of P is the interval $[0,1]$ since P is a probability. Moreover, if we know the statistical test law, we can calculate, for any value of z, the probability

$$p = P_{\text{H0}}(|Z| > |z|).$$

We can deduce the distribution function of P by calculating

$$F(p) = P(P \leq p) = P_{\text{H0}}(|Z| > |z|) = p,$$

and deduce from it the density function:

$$f(p) = 1$$

by using $F(p) = \int_0^p f(p)dp$.
Thus, the probability law of the p-value of a statistical test is a uniform law $\mathcal{U}(0,1)$.
By using the law of P, we again come across the notion of statistical risk. If we run a test at the level α, we retain H0 if $p > \alpha$, and we reject H0 if $p \leq \alpha$. So, the probability of making a bad decision if H0 is true (false positive) is $P_{\text{H0}}(P \leq \alpha) = F(\alpha) = \alpha$.

The FDR can be calculated *a posteriori* from the distribution of p-values from the K tests, while taking into account that the expected distribution of p-values under H0 is a uniform distribution. On the other hand, if H1 is true, we can expect an excess of very low values from the p-value. The observed

distribution of p-values thus results from a combination of several distributions: the distribution of tests under H0 (uniform) and the distributions of tests under H1 (depending on the value of H1 but with an excess of low values). α_{FDR} is the level at which each test should be carried out to guarantee, overall, a given FDR value. Most statistics software can today calculate the α_{FDR} from the observed distribution of p-values from the K tests carried out.

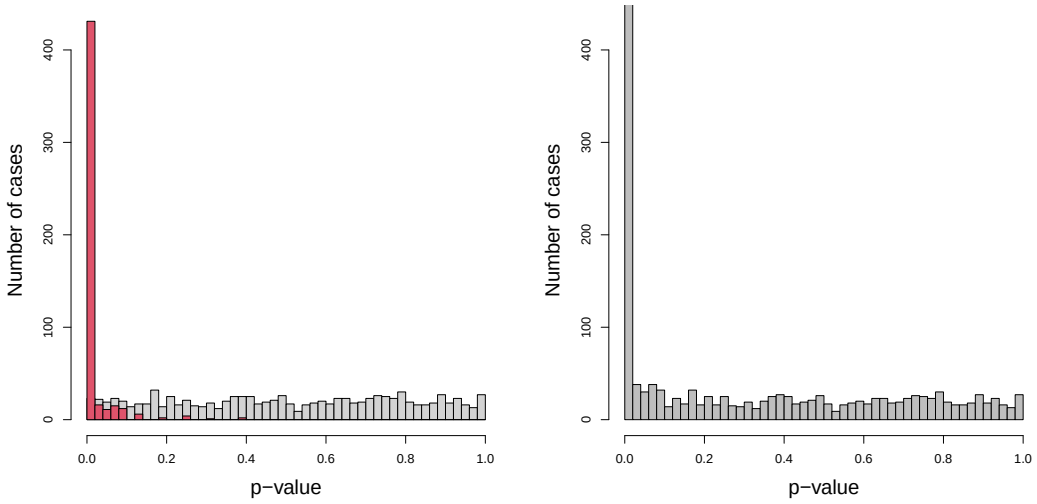


Figure 2.10: **Principle for calculating the *FDR*.** To illustrate the calculation principle, we will again use the example of the test given in chapter 2.2.1. We carry out this test for 1500 machines. For each machine we make $n = 30$ measurements. We carry out a bilateral test with H0: $\mu = 16$ as against H1: $\mu \neq 16$. We will simulate a set of data, by considering that 1000 machines are correctly adjusted, that is, under H0 ($\mu = 16$) and 500 machines are not correctly adjusted, that is, under H1($\mu \neq 16$, we take a mean between 16.01 and 16.2). **Left:** distributions of p-values for the machines under H0 (in white) and under H1 (in red). **Right:** distribution of p-values for all the machines tested. At the level 5%, the calculated FDR is 11.5%, that is to say, 11.5% of the machines for which we rejected H0 are false positives. At the level 1%, we have an FDR of 3.4%. If we want to apply the Bonferroni correction, we need to apply a level of 0.00003 for an overall level of 5% at each test. In this case, we find no false positives. This level is too "stringent", since the calculation made for the Bonferroni correction assumes that all the machines are well adjusted, which is not the case.

Chapter 3

Tests of conformity

When we study the distribution of a random variable in a population, we sometimes want to know if **one of the parameters of the probability distribution of the random variable is equal to a particular value**. The test statistic is then generally called a **test of conformity**, and changes depending on the type of variable studied and the question asked.
Here are a few examples:

- We want to check whether a piece of equipment used to measure the focal length of contact lenses is correctly adjusted. The manufacturer says that the accuracy of measurements is such that $\sigma = 0.1$. We have a benchmark with a known focal distance at a value of 16 cm. We carry out $n = 10$ measurements of focal length of the benchmark using the equipment. In this case, we are studying a quantitative r.v., where we know the variance but not the mean. In some conditions, we can apply a normality (Gaussian) test as described in the previous chapter.

- Wild type drosophila have long wings. We are studying a pure mutant strain, called the *miniature* strain, with individuals with small wings. Crossing flies from the *miniature* strain and wild type flies produces F1 descendants of the wild phenotype, irrespective of the direction of the crossing. If the character is controlled by a locus where the dominant allele is *long wings*, we will expect, when crossing the F1 flies, that 3/4 of the individuals will have the wild phenotype and there will be 1/4 of *miniature* individuals among the descendants. This is a qualitative random variable. In some conditions, we can apply the $\mathcal{X}^2$ test of conformity (chi-square test).

- We want to know the effect of an injection of insulin on the blood sugar level of diabetic patients. In a sample of patients, we measure the difference in blood sugar levels before and two hours after injections. If the treatment has no effect, then on average the difference will be zero. In some cases, we could use the Student's t-test.

The choice of the test statistic depends on the nature of the random variable studied.

3.1 Discrete random variables: conformity to a known distribution

3.1.1 The $\mathcal{X}^2$ (chi-square) test of conformity

3.1.1.1 Example

We perform a genotype test on a sample of 300 recombinant inbred lines of wheat, derived from a cross between two pure lines for a set of micro-satellite markers. The recombinant inbred lines are from ten or so self-fertilised generations from the initial hybridisation, which makes them practically homozygotes. We would therefore expect a 1 : 1 segregation in the descendants for each marker. A deviation from the 1 : 1 segregation is called a meiotic distortion, which may result from bias during meiosis or from an effect of selection (the allele in one of the parents could have a selective advantage over the allele in the other parent). We call a_1 the female parent's allele in the initial hybridisation, and a_2 the male parent's allele in the initial hybridisation. The collected data are given in a count table:

Genotype	a_1a_1	a_2a_2	Total
Number of lines	n_1	n_2	n

1. **Model.** We call X the random variable associated with a micro-satellite locus of a recombinant population line. Since there are only two parental lines in the beginning, the support of X is $\{a_1, a_2\}$, where a_1 and a_2 are the alleles carried by each of the two parents. The probability distribution of X is a discrete distribution, described by:

$$P(X = a_1) = p \quad ; \quad P(X = a_2) = q = 1 - p.$$

In the absence of meiotic distortion, we expect $p = 0.5$. If there is meiotic distortion, we expect $p \neq 0.5$, but the value of p depends on the type of mechanism at work and it is not known. We can consider an n-sample $\{X_1, ..., X_n\}$ corresponding to a random draw of n recombinant lines.

If we call Y_{a_1} the random variable measuring the number of times that $X_i = a_1$ in the sample, we know that this random variable follows a Binomial distribution $\mathcal{B}(n; p)$, and that $E(Y_{a_1}) = np$. We can say that the *expected* value of the class a_1 is $m_1 = np$. With the same reasoning, we can show that the expected value of the class a_2 is $m_2 = nq$.

We have an outcome from an n-sample $\{x_1, ..., x_n\}$ of X.

2. **Hypotheses H0 and H1.** We choose a quantifiable H0 hypothesis.

   H0 : $p = 0.5$
   H1 : $p \neq 0.5$

3. **Choice of a test statistic.** The test statistic is the $\mathcal{X}^2$ (pronounced chi-squared) statistic, which depends on the difference between the expected values under H0 and the observed values. In this example, There are only two modalities. The test statistic is a sum of two terms:

$$Z = \sum_{j=1}^{2} \frac{(n_j - m_j)^2}{m_j}. \tag{3.1}$$

   Under the H0 hypothesis, we can expect to see small values for Z, that is, that the observed numbers should be similar to the expected values. Under H1, we expect to find bigger differences between the observed and the expected values, and so the values of Z will tend to be bigger than under H0. We should note that, according to the equation 3.1, Z is always positive.

   Under H0, the distribution of Z is a $\mathcal{X}^2$ distribution, where the parameter is the number of degrees of freedom. **In the case of a test of conformity, the number of degrees of freedom is the number of classes of the qualitative variable studied minus 1**. In this example, there are two classes (a_1 and a_2), and therefore a single degree of freedom:

$$Z \sim_{\text{H0}} \mathcal{X}^2_1.$$

   Please note, the $\mathcal{X}^2$ distribution for the statistic Z is an asymptotic distribution. In other words, the Z distribution under H0 is close to a $\mathcal{X}^2$ distribution when the *expected* values of the classes are large enough. We consider this approximation as valid for the expected values above or equal to 5. We can therefore apply this test only if the expected values are above or equal to 5.

4. **Rejection region.** Depending on the expression of Z (3.1 equation), any deviation between the observed values and the expected values will increase the value of Z. So we can reject H0 for the values of Z that are too large. The rejection region take the form:

$$[z_{threshold}; +\infty[$$

   We want to find the value of $z_{threshold}$ such that $P_{H0}(Z > z_{threshold}) = \alpha$ so $z_{threshold}$ is the quantile of order $1 - \alpha$ in the $\mathcal{X}^2_1$ distribution.

5. **Outcome of test.** The values of $z_{\text{threshold}}$ are tabulated for any level α. The value of z_{obs} is calculated using the 3.1 equation in which the expected values of the classes are calculated under H0, and the p-value is calculated using the cumulative distribution function:

$$pval = 1 - F_{\mathcal{X}^2_1}(z_{\text{obs}}).$$

   R function: `chisq.test(x, p=prob)`, where `x` is the vector of n_i and `prob` the vector of probabilities under H0.

Validation of conditions of application. We check, *a posteriori*, that all the expected values m_j are higher or equal to five. It should be noted that if one or more classes have an expected value below five, it is possible to run the test again, combining classes and reducing the number of modalities and degrees of freedom.

3.1.1.2 General case

We study an n-sample of a random variable X with the support $\{a_1, ..., a_J\}$. The probability distribution is known under H0:

$$P_{\text{H0}}(X = a_j) = p_j$$

Under H0, the expected values of the class j can be calculated:

$$m_j = np_j$$

and the test statistic

$$Z = \sum_{j=1}^{J} \frac{(n_j - m_j)^2}{m_j}$$

approximately follows a $\mathcal{X}^2_{J-1}$ distribution under the H0 hypothesis, provided that the expected values are large enough ($m_j \geq 5$).
For a test of conformity, the degrees of freedom are the number of classes minus 1. The observed values n_j in each class are linked together by the relation $\sum_{j=1}^{J} n_j = n$. The same is true for expected values. So, the test statistic is a sum of J terms, but the J^{th} term of the sum is deduced from the other terms.

3.1.2 Fisher's exact test

When the conditions of the application of the $\mathcal{X}^2$ test are not met (with at least one class having an expected value < 5), and we do not want to group classes, we can carry out a Fisher's exact test.
The principle of this test consists in simulating data under the H0 hypothesis and comparing the observed statistic z_{obs} with the Z distribution simulated under H0. In the case of a test of conformity, we simulate T draws of n-samples in the X probability distribution under H0. For each draw t, we calculate the value z_t of Z. We then estimate the p-value of the test by the proportion of cases where $z_t > z_{\text{obs}}$:

$$pval = \frac{\text{Number of cases where } z_t > z_{\text{obs}}}{T}$$

The higher the number of draws, the closer the simulated distribution under H0 will be to the *true* Z distribution under H0. In practice, we can take $T = 2000$.

3.2 Continuous random variables: tests on the mean

We study an n-sample $X_1, ..., X_n$ of a continuous random variable X. We will look at two situations:

1. **Gaussian** The random variables X_i are independent, normal (Gaussian) with the same distribution $\mathcal{N}(\mu; \sigma^2)$. In this case, the empirical mean $\bar{X}$ follows a normal distribution $\bar{X} \sim \mathcal{N}(\mu; \frac{\sigma^2}{n})$

2. **Large samples** The random variables X_i are independent, have the same distribution, with a mean μ and variance σ^2. We make no hypotheses about the distribution of the X_i but the size of the sample is large enough ($n \geq 30$) for the central limit theorem to be applied. In this case $\bar{X} \approx \mathcal{N}(\mu; \frac{\sigma^2}{n})$

We want to know if the mean μ is equal to a known reference value μ_{H0}. Depending on our initial approach, we can choose one of three tests:
H0 : $\mu = \mu_{\text{H0}}$ as against H1 : $\mu \neq \mu_{\text{H0}}$
or
H0 : $\mu = \mu_{\text{H0}}$ as against H1 : $\mu < \mu_{\text{H0}}$
or
H0 : $\mu = \mu_{\text{H0}}$ as against H1 : $\mu > \mu_{\text{H0}}$

Please note, the choice of the H1 hypothesis must not be based on the values of the sample but on what you want to test, and this results from previous studies. Similarly, H0 must be quantifiable, so μ_{H0} must have a numerical value.

3.2.1 Gaussian case, known variance

If the X_i follow a normal distribution $\mathcal{N}(\mu; \sigma^2)$ and we know σ, then the test statistic Z below follows a normal standardised distribution under H0:

$$Z = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{\sigma} \sim_{\text{H0}} \mathcal{N}(0; 1) \tag{3.2}$$

The statistical test is carried out as shown in the example in chapter 2. We then need to check the conditions of application by ensuring in a graph that the observed values have a distribution resembling a Gaussian distribution (see below).

3.2.2 Gaussian case, unknown variance: Student's t-test

If the X_i follow a normal distribution $\mathcal{N}(\mu; \sigma^2)$ and we don't know the value of σ in the population, which is true in most cases, we can develop a test statistic by replacing σ in the equation 3.2 with its estimator S_{n-1}.
In this case, the statistic

$$T = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{S_{n-1}} \sim_{\text{H0}} \mathcal{T}_{n-1} \tag{3.3}$$

follows a Student's t-distribution at $n - 1$ degrees of freedom under hypothesis H0. Here, the lower degree of freedom results from the fact that we have replaced σ with its estimator S_{n-1}. The Student's t-distribution is a symmetric law with a mean of zero. It tends asymptotically towards a normal distribution $\mathcal{N}(0; 1)$ when the number of degrees of freedom increases. In practice, we can use the distribution $\mathcal{N}(0; 1)$ instead of the distribution $\mathcal{T}_{\text{ddl}}$ when $ddl \geq 30$.
The rest of the statistical test takes place in the same way as before. We call t_{obs} the value of the test statistic in the sample.

Rejection region and *p*-value. If H0 is true, then the mean of the $\overline{X}$ sample should be close to the expected mean μ_{H0}, and the test statistic T will be close to zero. If H0 is false, we will expect high absolute values for T. The form of the rejection region depends on the H1 hypothesis.

- H1 : $\mu \neq \mu_{\text{H0}}$: we reject H0 for large values of T, whether positive or negative. The rejection region is made up of two intervals:

  $$]-\infty, -t_{\text{threshold}}] \cup [t_{\text{threshold}}, +\infty[.$$

  By using the symmetrical properties of the Student's t-distribution, the p-value is calculated as follows:

  $$pval = 2(1 - F_{\mathcal{T}_{\text{ddl}}}(|t_{\text{obs}}|)).$$

- H1 : $\mu > \mu_{\text{H0}}$: we reject H0 for large positive values of T. So the rejection region is the interval:

  $$[t_{\text{threshold}}, +\infty[.$$

The p-value is calculated as:

$$pval = 1 - F_{\mathcal{T}_{\text{ddl}}}(t_{\text{obs}}).$$

Please note, in this case we use t_{obs} and not its absolute value. Negative values of t_{obs} correspond to a p-value higher than 0.5 and not to a rejection of H0.

- H1 : $\mu < \mu_{\text{H0}}$: we reject H0 for the large negative values of T. So the rejection region is the interval:

$$]-\infty, -t_{\text{threshold}}].$$

The p-value is calculated as:

$$pval = F_{\mathcal{T}_{\text{ddl}}}(t_{\text{obs}}).$$

Please note, we also use t_{obs} in this case and not its absolute value. The positive values of t_{obs} correspond to a p-value higher than 0.5 and to the non-rejection of H0.

Validation of conditions of application. If the size of the sample is small ($n < 30$), we need to check that the distribution of X is Gaussian. We can use a specific graph to do so. This is the quantile-quantile graph, also called the quantile-quantile plot or Q-Q plot.
We know that if $X \sim \mathcal{N}(\mu; \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0; 1)$, and the quantiles in the X distribution are linked to the quantiles in the $\frac{X-\mu}{\sigma}$ distribution by the relation:

$$Q_\alpha(X) = \mu + \sigma Q_\alpha\left(\frac{X-\mu}{\sigma}\right)$$

So, in a graph we can compare the quantiles observed in the sample $(x_1, ..., x_n)$ with the expected quantiles in a normal distribution $\mathcal{N}(0; 1)$ (figure 3.1). We can check that the dots on the Q-Q plot are in a straight line, which is basically the case here.

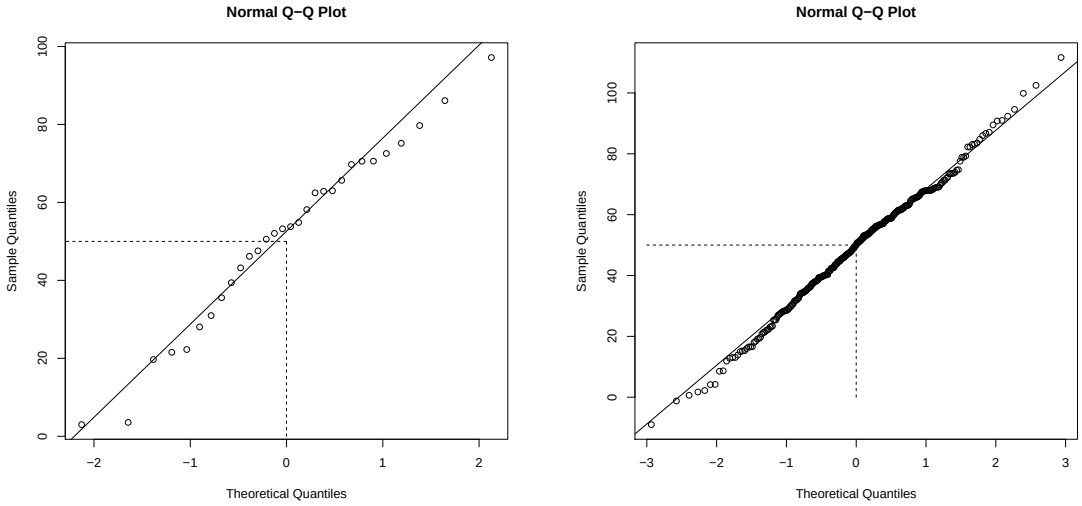


Figure 3.1: **Quantile-quantile plots.** The two graphs were made using simulated data with $X \sim \mathcal{N}(\mu = 50; \sigma^2 = 20^2)$. Two sizes of samples were used, $n = 30$ (left) and $n = 300$ (right). The dotted lines correspond to expected means (0 on the x-axis and 50 on the y-axis). The incline of the straight lines is equal to the expected deviation ($\sigma = 20$).

3.2.3 Non-Gaussian case, large n

Even if X_i does not follow a normal distribution, the central limit theorem guarantees that if the size of the sample is big enough ($n > 30$), then the mean $\overline{X}$ approximately follows a normal distribution

$\mathcal{N}(\mu; \frac{\sigma^2}{n})$. What's more, the law of large numbers guarantees that the estimator S^2 is very close to σ^2. So, whether we know σ^2 or not, the statistic

$$T = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{S_{n-1}} \approx_{\text{H0}} \mathcal{N}(0,1) \tag{3.4}$$

approximately follows a normal distribution $\mathcal{N}(0,1)$. The statistical test is conducted as above, using the normal distribution $\mathcal{N}(0,1)$ for the calculation of the threshold and the p-value. The only application condition is the size of the sample, which must be large enough.

3.2.4 Non-Gaussian case, small n

If we have few observations, we must carry out a non-parametric test (see section 3.5).

3.3 Paired data

It may happen that we want to compare two random variables corresponding to different characteristics of a single individual in a population. For example:

- We want to compare the efficacy of two moisturisers A and B. We apply one moisturiser to patients' right hands and the other to the left hand, and then measure the result after a few hours.
- We try to compare two test methods (blood tests and saliva tests) used to estimate the prevalence rate of the vector of malaria. We use both methods on each patient in a hospital and compare the prevalence rates estimated by each method.
- We want to know if the deer's rear and front legs are the same length. We measure the difference in length between the front and rear legs on a sample of 354 deer living in Fontainebleau Forest to see if on average the length of the front and rear legs is the same.
- To find out whether a medicine for cholesterol is effective, we measure the cholesterol level of 60 patients before and after treatment.

Model. We consider an n-sample of a pair of random variables $\{(X_{1i}, X_{2i}), \quad i = 1, ..., n\}$. We want to know whether X_1 and X_2 have the same mean. We calculate the difference

$$Y_i = X_{1i} - X_{2i},$$

and we are interested in the random variable $\overline{Y} = \frac{\sum_i Y_i}{n}$. If Y follows a normal distribution, then $\overline{Y} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$. If Y does not follow a normal distribution but the size of the sample is large enough ($n \geq 30$), then $\overline{Y} \approx \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$. Here μ is the expected value of Y_i and σ^2 the variance in the population. If the treatment has no effect, we expect that all the Y_i will be distributed around zero, and in particular, that the mean will equal zero: $\mu = 0$. However, we do not necessarily know the variance σ^2, which we can estimate by S^2_{n-1} in the sample.

Hypotheses H0 and H1. We make the following hypotheses:
H0 : $\mu = 0$
H1 : $\mu \neq 0$
or, depending on the question asked,
H1 : $\mu > 0$ or $\mu < 0$

Test statistic: we come back to a test of conformity on the mean. If the variance of Y, σ^2, is unknown, then the test statistic will be of the following kind

$$T = \sqrt{n}\frac{\overline{Y}}{S_{n-1}}.$$

There are two possible cases for the test statistic distribution under H0:

- Y follows a normal distribution. Under H0, $T \sim_{\text{H0}} \mathcal{T}_{n-1}$
- n is large enough. Under H0, $T \sim_{\text{H0}} \mathcal{N}(0,1)$

3.4 Confidence interval on the mean

3.4.1 Definition

Issue. We consider an n-sample of a random variable $X_1, ..., X_n$ which describes a population. We do not know the mean μ or the variance σ^2 of X in the population. We can use the probability distribution of X_i to put forward a probable interval for the unknown mean μ, given the observations.

Confidence interval: *this is an interval containing the true value of the parameter with a certain degree of probability set beforehand. Thus, a confidence interval with a risk α contains the unknown value of the parameter with a probability of $1-\alpha$.*

3.4.2 Calculation of the confidence interval of a mean

Estimator of the mean. We can estimate the unknown mean by the empirical mean of the sample:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

If X follows a normal distribution, then the random variable

$$T = \frac{\overline{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}} \tag{3.5}$$

follows a Student's t-distribution with $n-1$ degrees of freedom, and we can calculate the value t_α such that

$$P_{\text{H0}}\left(-t_\alpha \leq T \leq t_\alpha\right) = 1-\alpha.$$

Using the relation 3.5, we can write:

$$P_{\text{H0}}\left(-t_\alpha \leq \frac{\overline{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}} \leq t_\alpha\right) = 1-\alpha,$$

and we see that it is possible to frame the unknown value of μ with two values depending on $\overline{X}$, S^2_{n-1} and t_α:

$$P_{\text{H0}}\left(-t_\alpha\frac{S_{n-1}}{\sqrt{n}} + \overline{X} \leq \mu \leq t_\alpha\frac{S_{n-1}}{\sqrt{n}} + \overline{X}\right) = 1-\alpha.$$

So, in this example, the confidence interval for μ with risk α is:

$$IC_{1-\alpha} = \left]\overline{X} - t_\alpha\frac{S_{n-1}}{\sqrt{n}}, \overline{X} + t_\alpha\frac{S_{n-1}}{\sqrt{n}}\right[. \tag{3.6}$$

We should recall that S^2_{n-1} is the estimator of the variance σ^2 of the population, and that it is calculated as follows:

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

It may also be useful to remember that when n is large, the value $t_{0.05}$ tends towards 1.96. So we have:

$$IC_{0.95} = \left]\overline{X} - 1.96\frac{S_{n-1}}{\sqrt{n}}, \overline{X} + 1.96\frac{S_{n-1}}{\sqrt{n}}\right[.$$

In practice, the confidence interval is calculated by replacing $\overline{X}$ and S^2_{n-1} with their outcomes $\overline{x}$ and s^2_{n-1} in the sample in question.
If X does not follow a normal distribution, but the sample is fairly large, we can use the same approach, remembering in this case that the T distribution (defined in 3.5) is close to a normal distribution $\mathcal{N}(0.1)$.

3.5 Continuous random variables: non-parametric tests

Non-parametric tests make no hypotheses about the probability distribution of the random variable. Their name comes from the fact that we make no hypotheses about the parameters of the distribution. In practice, we use them when the distribution of the random variable is unknown, and/or the numbers are not large enough to make an approximation using a known distribution. In a given population, we look at a quantitative aspect with which we associate a random variable X. We call $\mathcal{L}$ the probability distribution of X, which is unknown and has a density f. We consider an n-sample taken from the population $\{X_1, X_2, ..., X_n\}$.

3.5.1 Test on the median

Tested hypotheses. We want to find out if the median of the distribution is equal to a given value m (for example $m = 115$). We make the following hypotheses:
H0: the median of X is m, in other words $P(X \leq m) = 0.5$
H1: the median of X is different from m.
We could also use a unilateral formulation for the hypothesis H1, either H1: the median of X is higher than m, or H1: the median of X is lower than m.

Test statistic. Under the H0 hypothesis, we expect half of the values $\{X_1, X_2, ..., X_n\}$ to be lower than m. We can consider the random variables Y_i such that $Y_i = 1$ if $X_i > m$ and otherwise zero. We call Z the number of X_i above m:

$$Z = \sum_{i=1}^{n} Y_i.$$

Under the H0 hypothesis, the Z distribution is a binomial distribution $\mathcal{B}(n; 0.5)$.
NB: To calculate the test statistic, we will not consider the outcomes of X equal to m. If we delete the values, we need to take this into account to calculate n.

Example. The F252 inbred line of maize is, among other things, characterised by the median of the height of the plants, which is $m = 115$ cm. We have a batch of seeds. We want to check whether they really are F252 seeds. We grow a sample of $n = 10$ plants and measure their height after flowering. We call X the random variable "height after flowering". We will test:
H0: the median height of plants from the batch of grains is 115.
H1: the median height of plants from the batch of grains is different from 115.
We find the following values for X and Y:

Plant	x_{obs}=Height (cm)	y_{obs}
1	119.68	1
2	113.22	0
3	108.49	0
4	105.11	0
5	112.52	0
6	120.18	1
7	125.15	1
8	114.13	0
9	136.34	1
10	114.11	0

Under H0, the Z distribution is $\mathcal{B}(10; 0.5)$. Here, we have chosen a bilateral test. The rejection region of the test is symmetrical to 5. It takes the form $[0.5-a] \cup [5+a, 10]$. The value of a is fixed according to the risk α chosen so that $2 * P_{\text{H0}}(Z \leq 5-a) \leq \alpha$.
We find $z_{\text{obs}} = 4$. The p-value is calculated as

$$pval = P_{\text{H0}}(Z \leq 4) + P_{\text{H0}}(Z \geq 6) = 2 * P_{\text{H0}}(Z \leq 4) = 2 \cdot F_{\mathcal{B}(10;0,5)}(4) = 0,754$$

So we cannot reject H0 at the threshold $\alpha = 0.05$.
Function R : `2*pbinom(4,size=10,prob=0.5)`

3.5.2 Wilcoxon rank-sum test

Model. We went to know if the density is symmetrical in relation to a given value a. One of the ways to proceed is to compare the distributions of $X - a$ and $a - X$. To do so, we define two new variables. The first,

$$R_i = \text{rank}|X_i - a|,$$

measures how far X_i deviates from a when we study the ranks. The second, S_i is equal to $+1$ or -1 depending on whether X_i deviates from a in positive ($S_i = +1$) or negative values ($S_i = -1$).
If $X - a$ and $a - X$ have the same distribution, then we expect to find the same number of negative deviations as positive ones. In particular, the sum of negative ranks must be close to the sum of positive ranks.

Hypotheses H0 and H1. H0: $X - a$ and $a - X$ have the same distribution as against H1: $X - a$ and $a - X$ do not have the same distribution.

Test statistic. We call T^+ the sum of the positive ranks (sum of R_i for which $S_i = +1$) and T^- the sum of negative ranks (sum of R_i for which $S_i = -1$). Under the hypothesis H0, the statistic

$$W = \min(T^+, T^-)$$

follows the Wilcoxon distribution, which is tabulated.
We notice that if the ranks are well shared out on either side of a, we expect that $T^+ = T^-$. What's more, the sum of all the ranks $(T^+ + T^-)$ is equal to:

$$1 + 2 + \quad ... \quad + n = \frac{n(n+1)}{2}.$$

So under the H0 hypothesis, we have

$$E(W) = \frac{n(n+1)}{4}.$$

Example. We can again take the previous example, with $a = 115$:

Plant	Height (cm)	$X - a$	R_i	S_i
1	119.68	+4.68	5	+1
2	113.22	–1.78	3	–1
3	108.49	–6.51	7	–1
4	105.11	–9.89	8	–1
5	112.52	–2.48	4	–1
6	120.18	+5.18	6	+1
7	125.15	+10.15	9	+1
8	114.13	–0.87	1	–1
9	136.34	+21.34	10	+1
10	114.11	–0.89	2	–1

So we find $T^+_{\text{obs}} = 5+6+9+10 = 30$ and $T^-_{\text{obs}} = 3+7+8+4+1+2 = 25$, so $w_{\text{obs}} = 25$. The p-value is calculated under R using `2*psignrank(25,10)`. We find $pval = 0.846$. So we cannot reject the H0 hypothesis.

3.6 Summary statement

With a test of conformity we can compare one or more parameters of a probability distribution with known reference values. Depending on the nature of the random variable and the question asked, we will use different statistical tests.

Discrete random variable

For each possible value $\{a_1, ..., a_J\}$ of X we know the probability $P_{\text{H0}}(X_i = a_j) = p_j$. The test statistic is calculated from the deviations between the numbers observed in each class n_j and the expected numbers under H0 $m_j = np_j$. Under H0, this statistic follows a $\mathcal{X}^2$ (chi-squared) distribution:

$$Z = \sum_{j=1}^{J} \frac{(n_j - m_j)^2}{m_j} \approx \mathcal{X}^2_{J-1}$$

The estimation of the Z distribution using the $\mathcal{X}^2$ distribution is only valid if all the m_j are higher than or equal to 5. Otherwise, we cannot run a Fisher's exact test.

Continuous random variable, test on the mean

We want to know if the expected value of X in the population is equal to a known value a.

1. X **follows a normal distribution**,

   - **Known variance.** Gaussian test.

   $$Z = \sqrt{n}\frac{\overline{X} - a}{\sigma} \sim_{H0} \mathcal{N}(0;1)$$

   - **Unknown variance.** Student's t-test.

   $$T = \sqrt{n}\frac{\overline{X} - a}{s_{n-1}} \sim_{H0} \mathcal{T}_{n-1}$$

2. X **does not follow a normal distribution but** $n \geq 30$,

   - Gaussian test

   $$T = \sqrt{n}\frac{\overline{X} - a}{s_{n-1}} \approx_{H0} \mathcal{N}(0,1)$$

3. **Small samples, non-parametric tests**

   - **Test on the median.** Under H0, the number of values of X that are larger than the median a follows a binomial distribution $\mathcal{B}(n; 0.5)$.
   - **Signed-rank test.** Wilcoxon rank-sum test.

   It should be noted that we can always use a non-parametric test, but these tests are less powerful (it is more difficult to reject H0). So if we have a Gaussian variable or a large sample, we prefer to use a parametric test.

Chapter 4

Tests of homogeneity

Tests of homogeneity are used in cases where we have two or more samples, and we want to know if the populations they are based on have shared characteristics, without necessarily knowing the parameters of the distribution of the random variable in these populations.
Here are a few possible cases:

- In the human population, are Swedes on average taller than pygmies?
- Is the conductance of the optical nerve on average lower in a group of individuals suffering from an inflammation of the nervous system than in a group of individuals with no inflammation?
- Is there sexual dimorphism for tail length in the population of French swallows?
- In a population of inbred strains of maize, are the genotypes in a genetic marker linked to a difference in the width of leaves?
- Is the number of births evenly distributed throughout the year in all French departments?

Here again, the choice of the test statistic depends on the nature of the random variable being studied.

4.1 Discrete random variables: $\mathcal{X}^2$ test

4.1.1 General case

This time, we have several n_i-samples. Each one comes from a different population, with the support of a random variable X $\{a_1, ..., a_j, ..., a_J\}$, and we want to know whether this variable follows the same law in each population. We call X_i the random variable associated with a draw in the i^{th} population. After sampling, we can sum up the data in a contingency table where each line represents a sample and each column to one of the possible values of X_i:

Support	a_1	...	a_j	...	a_J	Total
Sample 1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
...	...	...	...	...	...	...
Sample i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...
Sample I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n

where n_{ij} is the number observed in the sample i for the class a_j. The **marginal frequencies** are the sums on the lines or the columns of the table, written as $n_{i.}$ and $n_{.j}$, respectively. The size of the sample i is thus $n_{i.}$, the total number of individuals belonging to class a_j is $n_{.j}$ and the total frequency is n.

1. **Model.** Each sample is taken from a population. So we can associate a probability distribution to each sample:

$$P(X_i = a_j) = p_{ij}$$

   where p_{ij} is the frequency of the class a_j in the population i.

2. **Hypotheses H0 and H1.** We state as hypothesis H0 that the samples all follow the same probability distribution:

$$\text{H0} : P\left(X_i = a_j\right) = p_j$$

that is, that the frequency of each class a_j is the same in each population. We also have $\sum_{j=1}^{J} p_j = 1$. The alternative hypothesis H1 is that at least one sample does not follow this probability distribution:

$$\text{H1} : \quad \exists i, \quad P\left(X_i = a_j\right) = p_{ij} \neq p_j$$

3. **Test statistic.** Under the H0 hypothesis, we can estimate the frequency of a class a_j by using the marginal frequencies:

$$\widehat{p}_j = \frac{n_{.j}}{n}$$

The expected frequencies for the class a_j in the sample i are calculated as

$$m_{ij} = n_{i.}\widehat{p}_j = \frac{n_{i.}n_{.j}}{n}.$$

The test statistic is the $\mathcal{X}^2$ statistic:

$$Z = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{4.1}$$

Under the H0 hypothesis, the test statistic approximately follows a $\mathcal{X}^2$ distribution, providing that the expected frequencies are large enough:

$$Z \sim_{H0} \mathcal{X}^2_{(I-1)(J-1)}$$

The number of **degrees of freedom** (abbreviated as *ddl*) is calculated as follows: it is the number of independent terms of the sum, that is, the number of terms of the sum minus the number of restrictions.

To calculate the expected frequencies, we use all the marginal sums, that is to say:
$\sum_j n_{ij} = n_{i.}$ ($I - 1$ independent sums)
$\sum_i n_{ij} = n_{.j}$ ($J - 1$ independent sums)
$\sum_i n_{i.} = \sum_j n_{.j} = n$ (1 sum).

The total number of terms of the sum is IJ. So we have:

$$ddl = IJ - (I-1) - (J-1) - 1 = (I-1)(J-1).$$

Validation of conditions of application. Just like the $\mathcal{X}^2$ test of conformity, we need to check that the expected frequencies are higher or equal to 5 in each class. If the conditions of application are not met, we can make groupings according to class (if this is relevant biologically) or a Fisher's exact test.

4.1.2 Exact test

For the tests of conformity, we can carry out a Fisher's exact test by simulating the Z distribution under H0. This test will not be shown in detail here.

4.1.3 Examples

Here are a few examples of data that can be analysed using a $\mathcal{X}^2$ test of homogeneity.

- INSEE records the number of births each month in every French department. We want to know if births are distributed through the year in the same way in all departments.
- We are measuring the frequencies of haplotypes by resequencing in a region of the genome in different populations of thistles in fields in the Île-de-France region. We want to know if haplotypical frequencies are the same in all populations (dispersal of seeds), or if they differ from one field to another (dispersal by layering).
- A type of exam that is quite easy to correct is the multiple-choice test (MCQ). For a given question, each answer has the same chance of being chosen if students answer at random, but the right answer has more chances of being chosen if the students have learned their lessons. By analysing the answers to an MCQ test, we can find out whether the students have been working or not.

4.2 Continuous random variables: tests on the mean

We have two independent samples of a random variable X. So we have n_i-samples. We will write as X_{ij} the variable corresponding to the j-th individual of the sample from the population i ($i = 1.2$). Several situations can be considered:

- Each sample is Gaussian and has the same variance:

$$X_{ij} \sim \mathcal{N}(\mu_i; \sigma^2)$$

- Each sample is Gaussian with different variances:

$$X_{ij} \sim \mathcal{N}(\mu_i; \sigma_i^2) \quad \text{and} \quad \sigma_1^2 \neq \sigma_2^2$$

- The size of the sample is large enough ($n_i \geq 30$) for us to apply the central limit theorem: the empirical mean $\overline{X}_i$ approximately follows a normal distribution. We do not make hypotheses about the equality of variances. We have

$$\overline{X}_i \approx \mathcal{N}\left(\mu_i; \frac{\sigma_i^2}{n_i}\right).$$

4.2.1 Comparison of two means, Gaussian case, equal variances

4.2.1.1 Test of hypotheses

Model. We have two independent samples of a random variable X. Each sample is Gaussian and has the same variance σ^2, $X_{ij} \sim \mathcal{N}(\mu_i; \sigma^2)$. The X_{ij} are independent.

Hypotheses H0 and H1. We state the following hypotheses:

$$\text{H0} : \mu_1 = \mu_2$$

as against

$$\text{H1} : \mu_1 \neq \mu_2 \quad \text{(bilateral test)}$$

or

$$\text{H1} : \mu_1 < \mu_2 \quad \text{or} \quad \text{H1} : \mu_1 > \mu_2 \quad \text{(unilateral test)}$$

or

$$\text{H1} : \mu_1 > \mu_2 \quad \text{(unilateral test)}.$$

Test statistic. We can use the test statistic:

$$T = \frac{(\overline{X_1} - \overline{X_2})}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.2}$$

Under the H0 hypothesis, the test statistic follows a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom:

$$T \sim_{\text{H0}} \mathcal{T}_{n_1+n_2-2} \tag{4.3}$$

Since we have

$$\overline{X}_1 \sim \mathcal{N}\left(\mu_1; \frac{\sigma^2}{n_1}\right)$$

and

$$\overline{X}_2 \sim \mathcal{N}\left(\mu_2; \frac{\sigma^2}{n_2}\right).$$

so, as we have two independent samples,

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2; \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

hence

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0;1).$$

Under H0 $\mu_1 = \mu_2$, so we simply have:

$$\frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim_{H0} \mathcal{N}(0;1).$$

Since σ is unknown, we can replace it by its estimator S, hence the equation 4.2. This changes the probability distribution, since for the denominator, we replace a parameter with a value given by a random variable, hence the result 4.3.
To find the expression of S^2, we will use the two estimations of the variance of the population (from each of the samples). We can write $S_1^2 = \frac{SC_1}{n_1-1}$ and $S_2^2 = \frac{SC_2}{n_2-1}$ (SC = sum of squares). So the total SC is $(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2$ and the number of ddl (degrees of freedom) is $n_1 + n_2 - 2$. Hence:

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{4.4}$$

Rejection region and p-value. If H0 is true, then the difference between the two empirical means $\overline{X}_1$ and $\overline{X}_2$ will fluctuate around 0. If H0 is false, the distribution of T will not be centred around 0 and will be located to the left or right of the distribution under H0. The form of the rejection region depends on the H1 hypothesis. For a bilateral test, since the Student's t-distribution is symmetrical, it will take the form:

$$]-\infty, -t_{\text{threshold}}] \cup [t_{\text{threshold}}, +\infty[$$

For a bilateral test, the p-value is calculated by using the symmetrical property of the Student's t-distribution:

$$pval = 2\left(1 - F_{\mathcal{T}_{n_1+n_2-2}}(|t_{\text{obs}}|)\right)$$

Validation of conditions of application. In a graph (Q-Q plot) we check that the two random variables X_1 and X_2 follow a normal distribution. If this is not the case, but the samples are large ($n_i > 30$), we can use the test with the Gaussian approximation of the mean. If it is not the case, and we have a small sample or the samples are too small for checking using graphs, we have to use a non-parametric test.

4.2.1.2 Confidence interval

We can develop a confidence interval for the unknown difference between the two means, $\mu_1 - \mu_2$. The random variable

$$T = \frac{\left((\overline{X_1} - \mu_1) - (\overline{X_2} - \mu_2)\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

also follows a $\mathcal{T}_{n_1+n_2-2}$ distribution. We cannot calculate its value because μ_1 and μ_2 are unknown, but we can calculate the value t_α such that

$$P(-t_\alpha \leq T \leq t_\alpha) = \alpha.$$

Using a little algebra, as we did in the section on the tests of conformity, we find:

$$IC_{1-\alpha}(\mu_1 - \mu_2) =](\overline{X_1} - \overline{X_2}) - t_\alpha s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\overline{X_1} - \overline{X_2}) + t_\alpha s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}[. \tag{4.5}$$

Example. We are measuring the average number of seeds per fruit in 50 female plants and 50 hermaphrodite plants in the Gypsophila genus. The observations are summed up in the table below:

n	$\overline{x}_i$	$s^2_{i_{n_i-1}}$
50	15.64	135.37
50	17.30	149.26

The commune variance can be estimated as (*cf.* equation 4.4) :

$$s^2 = (49 \times 135.37 + 49 \times 149.26)/98 = 142.31.$$

We find the confidence interval for the difference $(\mu_2 - \mu_1)$:

$$IC_{95\%} =] - 6.23, +2.90[.$$

Note: Zero is included in the confidence interval. We can find the test conclusion: there is no significant statistical difference between the two means at the chosen level α.

4.2.2 Comparison of two means, Gaussian case, unequal variances

In most cases, we do not know the mean or the variance of the two samples. We can run an F-test of equality of variances (see below). It may so happen that we cannot make the hypothesis of equality of variances. In this case, we can use the Welch's t-test, but this is an imprecise test.

Model. We have two independent samples of a random variable X. Each sample is Gaussian, $X_{ij} \sim \mathcal{N}(\mu_i; \sigma_i^2)$. We have $\sigma_1^2 \neq \sigma_2^2$. The X_{ij} are independent.

Hypotheses H0 and H1. We state H0: $\mu_1 = \mu_2$ as against H1: $\mu_1 \neq \mu_2$.
We can also use a unilateral test, taking as an alternative hypothesis: H1: $\mu_1 > \mu_2$ or H1: $\mu_1 < \mu_2$.

Test statistic. Under the H0 hypothesis, the test statistic is

$$W = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}}$$

Under H0, this statistic follows a Student's t-distribution asymptotically, but with a different number of degrees of ν, which we call *effective degrees of freedom*, which depends on the estimations of variances of X_1 and X_2:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-2)}}$$

So,

$$W \approx_{H0} \mathcal{T}(\nu)$$

By default, the `t.test` function of R software assumes the inequality of variances and performs a Welch's t-test. Please notes that the distribution under H0 is an *asymptotic* law. In other words, the distribution under H0 is only known when the sizes of the two populations are large.

4.2.3 Comparison of two means, non-Gaussian case, large samples

Model. We assume that we have two independent samples. The mean of the X_{ij} is written μ_i and their variance as σ_i^2. We make no hypothesis about the distribution of the X_{ij} but assume that the samples are large enough ($n_i \geq 30$) for us to make the hypotheses that

$$\overline{X}_i \approx \mathcal{N}\left(\mu_i; \frac{\sigma_i^2}{n_i}\right).$$

.

Hypotheses H0 and H1. We state H0: $\mu_1 = \mu_2$ as against H1: $\mu_1 \neq \mu_2$.
As for the other tests, we can also state an alternative unilateral hypothesis.

Test statistic. We will use the test statistic

$$Z = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S_{1_{n_1-1}}^2}{n_1} + \frac{S_{2_{n_2-1}}^2}{n_2}}}$$

Since we have large samples, we consider that the difference in means follows a Gaussian distribution asymptotically and that the estimators of variances converge towards the true values of the variances. So, under the H0 hypothesis, Z follows a standard normal deviate asymptotically:

$$Z \approx_{H0} \mathcal{N}(0;1)$$

4.2.4 Non-parametric Mann–Whitney U test

If we know nothing about the random variables X_1 and X_2, and the size of the samples is too small to apply the central limit theorem (we are not sure whether $\overline{X}_1$ and $\overline{X}_2$ are Gaussian), we can always carry out a non-parametric test, based on ranks.

Model. The idea of this test is to compare two values chosen at random, one from the first sample, and the other from the second sample. If the random variables follow the same probability distribution, then there is a fifty percent chance that the first value is lower than the second:

$$P(X_1 \leq X_2) = P(X_2 \leq X_1) = 0.5.$$

.
We can rank the $n = n_1 + n_2$ elements in the two samples and then define, for each individual, its rank in the sequence formed. If the probability distributions are the same, the ranks of individuals in the two samples should be comparable. It should be noted that the sum of all the ranks is equal to

$$1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}.$$

Hypotheses H0 and H1. We state:

$$\text{H0} : P(X_1 \leq X_2) = P(X_2 \leq X_1) = 0.5.$$

H0 will always be true if both populations follow the same probability distribution.

We can carry out a bilateral test by stating:

$$\text{H1} : P(X_1 \leq X_2) \neq P(X_2 \leq X_1),$$

or a unilateral test, for example:

$$\text{H1} : P(X_1 \leq X_2) < P(X_2 \leq X_1).$$

In both cases, H1 implies that the probability distributions of both populations are different.

Test statistic. We can calculate the sum $R1$ of the ranks of individuals in the first sample, and the sum $R2$ of the ranks of individuals in the second sample, then calculate, for each sample, the difference with the expected minimal value (if the individuals in the sample are in the smallest ranks from 1 to n_1 or n_2 depending on the sample in question):

$$U1 = R1 - n_1(n_1 + 1)/2$$

$$U2 = R2 - n_2(n_2 + 1)/2$$

Since $R1 + R2 = n(n + 1)/2$, we have $U_1 + U_2 = n_1 n_2$. So, if the ranks of the two samples are comparable (H0 hypothesis), we expect a mean value of $\frac{n_1 n_2}{2}$ for each of the two random variables. Each of these variables follows a distribution that can be given in table form under H0. In practice, the tables show the smallest distribution of the two variables, so the test statistic is: $U = \min(U1, U2)$.

Example. A medical analysis laboratory uses two different pieces of equipment to measure patients' blood sugar. The technician wants to know whether the two pieces of equipment provide the same measurements. To do so, he uses post-prandial blood sugar data (1 hr 30 mins after a meal) obtained in the course of a day by the laboratory. The norm is a value below 1.40 g/L. The data can be shown as follows:

Blood sugar	1.64	1.05	1.58	1.20	1.43	1.35	1.10	1.65	1.25	1.63	1.51	1.43	1.32	1.27
Equipment	1	1	1	1	1	1	2	2	2	2	2	2	2	2
R rank	13	1	11	3	8.5*	7	2	14	4	12	10	8,5*	6	5

* Non-integral value since there are two 8^{ths} at the same rank.

The sizes of the samples are $n_1 = 6$ and $n_2 = 8$. We can use the table to calculate $R1 = 43.5$ and $R2 = 61.5$. So we find

$$U_1 = 43.5 - 6 \times 7/2 = 22.5$$

$$U_2 = 61.5 - 8 \times 9/2 = 25.5$$

The test statistic is equal to $U = 22.5$. The p-value can be calculated using the function `wilcox.test(x1,x2)` in R, where x1 is the vector of values from equipment 1, and x2 the vector of values from equipment 2. We find a p-value of 0.8972, and we cannot reject H0.

4.3 Test of homogeneity on the variance

Model. We have two samples with a random variable X. We call X_i the random variable corresponding to the sample i. We assume that each sample is Gaussian:

$$X_i \sim \mathcal{N}(\mu_i; \sigma_i^2)$$

We want to know whether the variances in the two populations are equal.

Hypotheses H0 and H1. We state the following hypotheses:

$$H0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$$

as against

$$H0 : \sigma_1^2 \neq \sigma_2^2$$

Test statistic. To find a test statistic, we use the following property: if X follows a Gaussian distribution, then

$$\frac{S_{n-1}^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2.$$

This property is true for each of the two samples. So, under the hypothesis H0,

$$F = \frac{S_{1_{n_1-1}}^2}{S_{2_{n_2-1}}^2}$$

is a relation of two random variables, each following a $\mathcal{X}^2$ distribution. The result is an F-distribution $\mathcal{F}_{n_2-1}^{n_1-1}$.

Shape of the rejection region and *p*-value. Unlike the normal distribution and the Student's t-distribution, the F-distribution is not symmetrical. The quantile on the left is not the opposite of the quantile on the right. However, if $F \sim \mathcal{F}_{n_2-1}^{n_1-1}$, then $1/F \sim \mathcal{F}_{n_1-1}^{n_2-1}$. Moreover, the values of F are always positive.
Under H0, we expect S_1^2 and S_2^2 to estimate the same quantity σ^2, and so their relation must be close to 1. Very large or very small values in the relation are less probable under the hypothesis H0. So the form of the rejection region is:

$$[0, f_{\text{inf}}] \cup [f_{\text{sup}}, +\infty[.$$

For a test at the level α, the values of f_{inf} and f_{sup} are the quantiles at $\alpha/2$ and $1-\alpha/2$ of the $\mathcal{F}_{n_2-1}^{n_1-1}$ distribution.
To calculate the p-value, we will calculate the probability under H0 of an area of the form $[0, a] \cup [b, +\infty[$ where a and b are respectively located to the left and right of the median of the $\mathcal{F}_{n_1-1}^{n_2-1}$ distribution, which we will write here as m.

- If $f_{\text{obs}} > m$, then $b = f_{\text{obs}}$ and the p-value is calculated as

$$pval = 2 * P(F \geq f_{\text{obs}}) = 2\left(1 - F_{\mathcal{F}_{n_2-1}^{n_1-1}}(f_{\text{obs}})\right)$$

- If $f_{\text{obs}} < m$, then $a = f_{\text{obs}}$ and the p-value is calculated as

$$pval = 2 * P(F \leq f_{\text{obs}}) = 2F_{\mathcal{F}_{n_2-1}^{n_1-1}}(f_{\text{obs}})$$

Under R, the distribution function of the F-distribution is obtained using the function `pf(fobs,n1-1,n2-2)`.

4.4 Summary statement

With the tests of homogeneity we can compare the parameters of the distribution laws of two random variables to see if they are equal. We will run different tests depending on the nature of the random variable and the question asked.

1. **discrete r.v.:** $\mathcal{X}^2$ tests. The support contains J classes. We can compare I samples. The theoretical sizes m_{ij} are calculated by $\frac{n_{i.}n_{.j}}{n}$ where $n_{i.}$ is the sample size i and $n_{.j}$ is the number of individuals from the class j in the I samples.

$$Z = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \sim_{H0} \mathcal{X}^2_{(I-1)(J-1)}$$

The theoretical sizes must be higher than or equal to 5 in each class.

2. **continuous r.v., test on the mean**

   - **Gaussian samples:** $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ Student's t-test.
     - **Equal variances**
     $$T = \frac{(\overline{X_1} - \overline{X_2})}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim_{H0} \mathcal{T}_{n_1+n_2-2}$$
     - **Unequal variances**
     $$W = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}} \approx_{H0} \mathcal{T}_\nu$$
   - **Non-Gaussian but large samples** ($n_i \geq 30$)
   $$Z = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}} \approx_{H0} \mathcal{N}(0;1)$$
   - **Small samples with no information about the X_1 and X_2 distribution:** Mann–Whitney U test.

3. **Continuous Gaussian r.v., test on the variance:** F-test

$$F = \frac{S^2_{1_{n_1-1}}}{S^2_{2_{n_2-1}}} \sim_{H0} \mathcal{F}^{n_1-1}_{n_2-1}.$$

Chapter 5

Pairs of random variables and statistical dependence

5.1 The $\mathcal{X}^2$ test for independence

Model. In a population, we will look at an n-sample of a pair of qualitative variables:

$$\{(X_1, Y_1), (X_2, Y_2), \quad ..., \quad (X_n, Y_n)\}$$

Each of these variables is defined by its support and by its probability distribution. This means that we can characterise each individual in the population with two variables. For example, in a population of cats, we can characterise each individual by its main fur colour ($a_1 = grey, a_2 = ginger, ..., a_k = white$) and its sex ($b_1 = F, b_2 = M$).

Random variable	Support	Probability distribution
X_i	$(a_1, \quad ... \quad , a_K)$	$P(X_i = a_k) = p_k$ (unknown)
Y_i	$(b_1, \quad ... \quad , b_L)$	$P(Y_i = b_l) = q_l$ (unknown)

We count the number of n_{kl} outcomes in the sample where $X_i = a_k$ and $Y_i = b_l$. The results can be shown in the form of a contingency table:

Support	b_1	...	b_l	...	b_L	Total
a_1	n_{11}	...	n_{1l}	...	n_{1K}	$n_{1.}$
...	...	...	...	...	...	...
a_k	n_{k1}	...	n_{kl}	...	n_{kL}	$n_{k.}$
...	...	...	...	...	...	...
a_K	n_{K1}	...	n_{Kl}	...	n_{KL}	$n_{K.}$
Total	$n_{.1}$	...	$n_{.l}$	...	$n_{.L}$	n

We can estimate $P(X_i = a_k) = p_k$ or $P(Y_i = b_l) = q_k$ using the marginal effects $n_{k.}$ and $n_{.l}$:

$$\widehat{p}_k = \frac{n_{k.}}{n} \tag{5.1}$$

$$\widehat{q}_l = \frac{n_{.l}}{n}. \tag{5.2}$$

We can ask whether X and Y are independent. If this is the case, then

$$P(X_i = a_k \quad \text{and} \quad Y_i = b_l) = P(X_i = a_k) \times P(Y_i = b_l) = p_k q_l$$

Assuming the hypothesis of the independence of the two variables, we can thus calculate the (average) theoretical numbers in each section of the table:

$$m_{kl} = n\widehat{p}_k\widehat{q}_l = \frac{n_{k.}n_{.l}}{n}.$$

H0 and H1 hypotheses. We assume
H0: X and Y are independent: $P(X_i = a_k \quad \text{and} \quad Y_i = b_l) = p_k q_l$, for every k, l
H1: X and Y are not independent: we have k, l such that $P(X_i = a_k \quad \text{and} \quad Y_i = b_l) \neq p_k q_l$
Here, there is only one way of presenting the alternative H1 hypothesis.

Test statistic and distribution under H0. We have calculated the expected numbers in each class under H0. If H0 is true, the numbers will vary around these values. Under H0, the test statistic:

$$Z = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(n_{kl} - m_{kl})^2}{m_{kl}} \approx_{\text{H0}} \mathcal{X}^2_{(K-1)(L-1)} \tag{5.3}$$

follows a law of $\mathcal{X}^2_{(K-1)(L-1)}$ asymptotically. So, if the theoretical numbers are large enough, under H0, Z approximately follows the law $\mathcal{X}^2_{(K-1)(L-1)}$.

Rejection region: Under H1, there will be a wider gap on average between the theoretical numbers and observed numbers than under H0. The distribution of the test statistic under H1 is therefore located to the right of the distribution under H0. So, we will reject H0 for the large values of Z_{obs}. The threshold will be chosen according to the selected α.

Conditions of application: the theoretical numbers must be higher or equal to 5 in each class: $\forall k, \forall l, m_{kl} \geq 5$.

5.2 Correlation

5.2.1 Pearson parametric test

Model. We consider an n-sample of a pair of Gaussian random variables

$$\{(X_1, Y_1), (X_2, Y_2), \quad ..., \quad (X_n, Y_n)\}$$

Gaussian pair : *When we observe two random variables for each individual, we say there is a pair of random variables. A pair of quantitative random variables is a Gaussian pair if each linear combination* $a * X_i + b * Y_i$ *is Gaussian.*

In practice, we will test the two variables for normality (case $a = 0$ and $b = 0$) and check that the relation between the two variables seems linear using a graph.

Comment: If X and Y are independent, then $Cov(X, Y) = 0$, while the opposite is only true in certain cases. For a Gaussian pair, zero covariance between X and Y ($Cov(X_i, Y_i) = 0$) shows the independence of the variables. Thus, to check on the independence of a Gaussian pair, we test the nullity of the correlation coefficient ρ_{XY} which is, we will recall, a measurement of the covariance not computed in units.

Hypotheses H0 and H1. We will assume we have a Gaussian pair. We want to test the independence of two variables. The hypothesis H0 is that X and Y are independent, which we can express by:

$$\text{H0} : \rho_{XY} = 0 \quad \text{(zero correlation)}.$$

The hypothesis H1 is that X and Y are not independent, and we can either choose a bilateral test:

$$\text{H1} : \rho_{XY} \neq 0 \quad \text{(the correlation is not zero)}$$

or a unilateral test:

$$\text{H1} : \rho_{XY} < 0 \quad \text{(the correlation is negative)}$$

or

$$\text{H1} : \rho_{XY} > 0 \quad \text{(the correlation is positive)}.$$

Test statistic and law under H0. We can estimate the correlation using the estimator below, which measures the linear relation between two random variables:

$$r_{XY} = \frac{S_{XY}}{S_{X_{n-1}} S_{Y_{n-1}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Under the hypothesis H0, the test statistic is:

$$Z = \sqrt{n-2}\frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sim_{\text{H0}} \mathcal{T}_{n-2}$$

following a Student's t-distribution at $n-2$ ddl.

Conditions of application. The test is valid for Gaussian pairs of random variables. However, for a large sample ($n \geq 30$), we will have a similar result, providing we do not move away from a linear relation.

5.2.2 Comment on the correlation coefficient

The correlation coefficient of two random variables measures the degree of linear dependence between these two variables (figure 5.1a), of the type

$$Y = a + bX + \epsilon \tag{5.4}$$

where a and b define the relations of dependence, and ϵ is a random variable representing measurement errors. We will see in the linear model chapter how to model these errors.
We can show that b, the gradient on the right, is:

$$b = \frac{cov(X,Y)}{V(X)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

We should note that there is a relation between b and the Pearson correlation coefficient (equation 1.2). Because

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{[V(X)V(Y)]}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2\sigma_Y^2}},$$

so

$$b = \rho_{XY}\sqrt{\frac{V(Y)}{V(X)}} = \rho_{XY}\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}.$$

5.2.3 Necessary precautions

The analysis of the correlations between random variables can lead to mistaken interpretations. The measure of the linear correlation (or the covariance) is not always a good indicator of the dependence between two random variables. An analysis of correlation must always be accompanied by a graphic analysis. The following points **should never be forgotten**:

- **Independence and correlation.** If two random variables are independent, then their covariance is zero. However, the inverse is not always true (figure 5.1b). Two random variables can be linked to each other but still have zero covariance.

- **Non-linear dependence**. Cases of non-linear dependence can lead to an incorrect estimation of the relation of dependence (figure 5.1c).

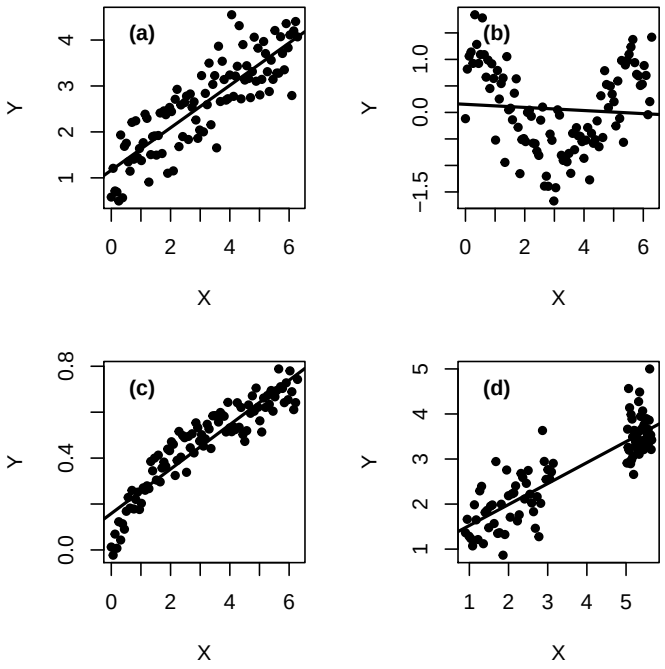


Figure 5.1: **Example of correlations between characters.** We can model a situation where the value of Y depends on X and on a random error that is the same in the four situations. **(a)** Linear model: $Y = 1+0.5X+\epsilon$. **(b)** Non-linear model: $Y = cos(X)+\epsilon$. **(c)** Non-linear model: $Y = \frac{X}{3+X}+\epsilon$. **(d)** Structured population for X: same relation as in **(a)** ($Y = 1+0.5X+\epsilon$), but the values of X are grouped around 2 and 6.

Model	Type	$\widehat{r}_{XY}$	p-value
(a)	linear	0.90	$< 2.2e^{-16}$
(b)	cos	0.04	0; 071
(c)	hyperbolic	0.25	0; 01
(d)	X structured	0.88	$< 2.2e^{-16}$

Table 5.1: **Results of Pearson test for data simulated in the figure 5.1.**

- **Dependence and causality**. Non-zero covariance between X and Y (figure 5.1a) does not necessarily imply a relation of causality between the two variables. There may be a random variable Z that determines both X and Y, resulting in a linear relation between X and Y. The relation of causality is between Z and X and between Z and Y, but not between X and Y.

- **Structured population.** Let us assume that the variable X clearly divides the population into two large classes. Within each class of values of X, the correlation between X and Y is weak or zero. The correlation between Y and X reflects the structured population for X (figure 5.1d) and in this case becomes harder to interpret.

The Table 5.1 shows the result of the Pearson test for sets of data simulated in the figure 5.1 and corresponding to four different cases of dependence between Y and X.

Please note, we must always combine the statistical test with a visual examination to check that the application conditions of the test have been verified.

5.2.4 Spearman's rank correlation coefficient

When the size of the sample is too small (< 30), the random variables are not Gaussian or the relation is not linear, we can replace the parametric test with a non-parametric test by working on the correlations of rank. The aim is to replace the values of observations by their rank and to study the difference in rank D_i. If X and Y are perfectly correlated, then the rank X_i will follow rank Y_i and the difference in ranks will be very small. On the other hand, if there is no relation between X and Y, there will be a random difference in rank.

Spearman's correlation coefficient. Spearman's rank correlation coefficient is given by

$$r_S = 1 - \frac{6 \sum_i D_i^2}{n^2(n-1)}.$$

Example. We want to study the relation between the annual consumption of chocolate (kg a year per person) and the accumulated number of Nobel Prize winners per 10 million inhabitants in all countries in Europe. The idea is that annual chocolate consumption reflects social level. We would expect a positive relation, but we only have values taken from a few countries ($n < 30$). We produce new discrete random variables, corresponding to the ranking of each variable, and we measure d_i, the difference in rank between X_i and Y_i.

Country	Chocolate consumption	Nobel Prizes	X_{rank}	Y_{rank}	d_i
Greece	2.5	2	1	1	0
Poland	3.9	3	2	2	0
Netherlands	4.2	11	3	4	–1
France	6.2	9,8	4	3	+1
United Kingdom	9.8	19	5	5	0
Switzerland	12	32	6	6	0

Table 5.2: **Relation between chocolate consumption and number of Nobel prizes.**

Hypotheses H0 and H1. We assume H0: there is no correlation between X and Y, as opposed to H1: X and Y are correlated.

Test statistic. For small samples, the r_S distribution under hypothesis H0 is tabulated (it does not correspond to a known distribution). For large samples, we have an asymptotic result and we can consider that the transformation of the following r_S approximately follows a Student's t-distribution:

$$r_S \sqrt{\frac{n-2}{1-r_S^2}} \approx \mathcal{T}_{n-2}.$$

5.2.5 Summary statement

The independence tests help us to check the hypothesis of zero covariance between a pair of random variables. To do so, we need to have a measurement for each random variable for every individual in the n-sample.

1. **Pairs of discreet r.v.: $\mathcal{X}^2$ test of independence**
   We have a contingency table giving the numbers observed in each of the classes defined by the combination of two random variables. The theoretical numbers m_{ij} are calculated by using the marginal frequencies

$$Z = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(n_{kl} - m_{kl})^2}{m_{kl}} \sim_{\text{H0}} \mathcal{X}^2_{(K-1)(L-1)}$$

The theoretical numbers must be higher than or equal to five in each class.

2. **Pair of continuous random variables**

- **Pearson's test** For a Gaussian pair:

$$Z = \sqrt{n-2} \frac{r_{XY}}{\sqrt{1 - r^2_{XY}}} \sim_{H0} \mathcal{T}_{n-2}$$

If we have a large sample ($n \geq 30$), there is an asymptotic result:

$$Z = \sqrt{n-2} \frac{r_{XY}}{\sqrt{1 - r^2_{XY}}} \approx_{H0} \mathcal{T}_{n-2}$$

- **Spearman's rank test.** The rank correlation coefficient is calculated as:

$$r_S = 1 - \frac{6 \sum_i D_i^2}{n^2(n-1)},$$

where D_i is the difference in rank between the two random variables for the individual i. The test statistic is

$$r_S \sqrt{\frac{n-2}{1 - r_S^2}} \approx_{H0} \mathcal{T}(n-2).$$

This distribution is only valid for large samples, otherwise the distribution is given in the form of a table.

Chapter 6

The linear model

The linear model consists of a family of models, including the analysis of variance model, the regression model and the analysis of covariance. The overall issue is as follows: we study the relations between a continuous random variable Y and a certain number of descriptive variables $X^{(1)}$, $X^{(2)}$, ..., $X^{(P)}$. We want to know if the mean of Y changes according to the value of the descriptors. The descriptive variables may be qualitative or quantitative. In all cases, we will produce a model of the relations of dependence between Y and X using a linear model. In other words, we assume that we can explain Y as the sum of several factors, including the mean effect of the variables $X^{(p)}$ and a random variable ϵ, which sums up the effects of unknown or non-verifiable factors that can lead to variations in Y.
Here are a few examples of problems that can be dealt with using the general linear model:

- We want to know whether the average production of wheat (hundreds of units per hectare) changes depending on the region in France. We carry out a survey in different regions to find annual yields.

  The variable to be explained, Y, is the yield of wheat in a farm according to the region, written as $(a_1, ..., a_i, ..., a_{13})$, where a_i is the name of the i^{th} region of France (13 regions in all). We conduct samples of n farms in France. We can list the farms from the region i with a_i. And we can produce a model of the yield from the farm j in the region i, written as Y_{ij} by:

  $$Y_{ij} = \mu_i + \epsilon_{ij}$$

  where μ_i is the average in the region a_i and where ϵ_{ij} is the random deviation between this average and the yield from the farm j. We can ask whether the averages per region are different. This type of problem is dealt with in the section *One-way analysis of variance*.

- We want to study the specific diversity of beetles according to soil type (grazing land or otherwise) and geographic region (mountains or plains). The variable Y to be explained is the specific diversity of beetles, which can be described as the cumulative effects of soil type, $X^{(1)}$, the region, $X^{(2)}$, and a random variable summing up the specific features of the sampling site. The model is as follows:

  $$Y_k = \mu + t_{X_k^{(1)}} + r_{X_k^{(2)}} + \epsilon_k$$

  where μ is a general mean. If we indicate each sample by soil type (i) and by region (j), we have:

  $$Y_{ijl} = \mu + t_i + r_j + \epsilon_{ijl}$$

  Here, l shows the replicate for the soil (i) and the region (j). We want to know if there are mean differences between soil types $(t_1 \neq t_2)$ or between regions $(r_1 \neq r_2)$. This type of issue is dealt with in the section *2-way analysis of variance*.

- We want to know if there is a relation in mammals between weight at birth and adult size. The variable Y to be explained is adult size, and the descriptive variable X is weight at birth. We study a sample of mice in captivity, but with parents living in a natural environment in various regions of the world. This time X is a continuous variable. We can produce a model of a linear relation between Y and X with

  $$Y_k = a + bX_k + \epsilon_k$$

  where b is the coefficient of proportionality between Y and X and where ϵ measures all the unverified events (which we will consider as random) that may occur in the course of a life to

change the relation between Y and X. We want to know if the coefficient of proportionality b is different from zero. This type of problem is dealt with in the section *Linear regression*.

- The relation of proportionality between weight at birth and adult size may vary according to the mammal species under consideration. We conduct the same study as before, but now consider three species: mice, human beings and pigs. We have two descriptors, the species (qualitative variable) and weight at birth (X, continuous variable). By indicating the sampled individuals according to their species (i), we can write

$$Y_{ik} = a_i + b_i X_{ik} + \epsilon_{ik}$$

We want to know if there are size differences between the species, and if there are differences between the species for the coefficient of proportionality between weight at birth and adult size. This type of problem is dealt with in the section *Analysis of covariance*.

6.1 One-way ANOVA, mean comparison tests

Issue. We examine several n-samples from the same Gaussian Y random variable, but from I different populations. We name Y_{ik} the random variable for the k^{th} individual of the sample from the population i. We want to know if the mean of this variable depends on the population in which we find it. Please note, with the ANOVA we cannot establish causal links but a statistical link.
We write as n_i the size of the sample from the population i (noted as $N = \sum_{i=1}^{I} n_i$). We presume that the variance of the variable is the same in each of the I populations, but that the means are potentially different, that is, that each Y_{ik} follows a normal distribution $\mathcal{N}(\mu_i; \sigma^2)$.

Linear model. We can rewrite the above statement using a mathematical formula helping us to give the information in shorter form:

$$Y_{ik} = \mu_i + \epsilon_{ik}, \qquad \epsilon_{ik} \quad \text{iid} \quad \mathcal{N}(0; \sigma^2) \tag{6.1}$$

The ϵ_{ik} are called the residuals of the model and correspond to the deviations of statistical individuals when compared with the average in their population. σ^2 is called the residual variance. iid means identically and independently distributed and is a reminder that Y_{ik} are from n-samples, that is, that they are independent and from the same distribution of a single sample.
We can also reformulate the model by defining μ, the general mean and $\alpha_i = \mu_i - \mu$, the gap between the mean for the population i and the general mean.

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}, \qquad \epsilon_{ik} \quad \text{iid} \quad \mathcal{N}(0; \sigma^2) \tag{6.2}$$

Examples

- We want to know if French people's average level of cholesterol depends on the region where they live. μ_i is the average level of cholesterol in the region i, and ϵ_{ik} represents the deviation between μ_i and the value of the individual k, a deviation due to genetic differences or differences in diet that cannot be explained by the influence of the region where they live. As a result, the expected value of ϵ_{ik} is zero.

- We want to compare the weight gained by a breed of cows during the summer grazing period in different fields in order to make recommendations to breeders. μ_i is the average weight gain after a season in field i, and ϵ_{ik} represents the variations between two cows due to genetic differences or individual differences about how and where they graze, which cannot be explained by the average quality of the fields. As a result, the expected valued of ϵ_{ik} is zero.

Estimators for averages. To estimate the mean in each population μ_i, we will use an estimator. To minimise the number of written figures, we generally use the symbol *point* $(\cdot)$ instead of a number to write that we have made an average (please note: in chapter 4, this symbol meant a sum total). Thus:

$$Y_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}$$

is used to estimate the average in the population i. Similarly,

$$Y_{..} = \frac{1}{\sum_i n_i} \sum_i \sum_k Y_{ik}$$

is used to estimate the general average μ. We should note that if I is the number of populations, the relation

$$Y_{..} = \frac{1}{I} \sum_{i=1}^{I} Y_{i.}$$

is only true if all the n_i are equal.
We can note that an estimator of α_i is $\hat{\alpha}_i = Y_{i.} - Y_{..}$.

Estimator of the residual variance. We can use the linear model (6.2) to estimate the residual variance. We have:

$$\epsilon_{ik} = Y_{ik} - \mu_i$$

Since the means μ_i are unknown, we can replace them by the empirical means $Y_{i.}$. The residual variance can then be estimated as the empirical variance of ϵ_{ik}:

$$\widehat{\sigma^2} = \frac{1}{\sum_i (n_i - 1)} \sum_i \sum_k (Y_{ik} - Y_{i.})^2 = \frac{1}{\sum_i (n_i - 1)} SSR.$$

SSR is the *Sum of Squared Residuals*. The corrective term $\sum_i (n_i - 1) = N - I$ used to estimate the residual variance from the SSR are the associated degrees of freedom, also written as ddl_R. So we have $\widehat{\sigma^2} = \frac{SSR}{ddl_R}$.

Standard error : *a standard error (SE) is the residual standard deviation estimated from several samples:*

$$SE = \sqrt{\frac{SSR}{ddl_R}}. \tag{6.3}$$

We can use this estimation of the standard error to develop a confidence interval for the mean of each population.

Comparison of means test, ANOVA 1
We try to determine whether the mean is the same in all the populations or if at least one population differs from the others.

- **Model.** The model is the linear model (6.2)
- **Hypotheses H0 and H1.** $H0 : \forall (i,j) \quad \mu_i = \mu_j, \quad H1 : \exists (i,j) \quad \mu_i \neq \mu_j$.
  We can reformulate them: $H0 : \forall i \quad \alpha_i = 0, \quad H1 : \exists i \quad \alpha_i \neq 0$
- **Test statistic.** To write down the test statistic, we start by analysing the variance described below. The empirical variance of Y is calculated from the total sum of squared deviations:

$$SST = \sum_i \sum_k (Y_{ik} - Y_{..})^2$$

that we can analyse by using

$$\begin{aligned} SST &= \sum_i \sum_k ((Y_{ik} - Y_{i.}) + (Y_{i.} - Y_{..}))^2 \\ &= \sum_i \sum_k \left((Y_{ik} - Y_{i.})^2 + (Y_{i.} - Y_{..})^2 + 2(Y_{ik} - Y_{i.})(Y_{i.} - Y_{..})\right) \\ &= \sum_i \sum_k (Y_{i.} - Y_{..})^2 + \sum_i \sum_k (Y_{ik} - Y_{i.})^2 + \sum_i \sum_k 2(Y_{ik} - Y_{i.})(Y_{i.} - Y_{..}) \end{aligned}$$

If we develop the last term, we have:

$$\begin{aligned} \sum_i \sum_k (Y_{ik} - Y_{i.})(Y_{i.} - Y_{..}) &= \sum_i \sum_k \left(Y_{ik}Y_{i.} - Y_{ik}Y_{..} - Y_{i.}^2 + Y_{i.}Y_{..}\right) \\ &= \sum_i Y_{i.} \sum_k Y_{ik} - Y_{..} \sum_i \sum_k Y_{ik} - \sum_i n_i Y_{i.}^2 + \sum_i Y_{i.} n_i Y_{..} \\ &= \sum_i Y_{i.} n_i Y_{i.} - Y_{..} \sum_i n_i Y_{i.} - \sum_i n_i Y_{i.}^2 + \sum_i Y_{i.} n_i Y_{..} \\ &= \sum_i n_i Y_{i.}^2 - N Y_{..}^2 - \sum_i n_i Y_{i.}^2 + N Y_{..}^2 \\ &= 0 \end{aligned}$$

So, we can deduce

$$\begin{aligned} SST &= \sum_i \sum_k (Y_{i.} - Y_{..})^2 + \sum_i \sum_k (Y_{ik} - Y_{i.})^2 \\ &= SSA + SSR \end{aligned}$$

The principle of the test statistic is to compare the size of the SSA and the SSR. It is written as:

$$F = \frac{\frac{SSA}{ddl_A}}{\frac{SSR}{ddl_R}}$$

where ddl_A is the number of degrees of freedom of SSA, that is, the number of independent terms in this sum. We can see that $SSA = \sum_i \sum_k (Y_{i.} - Y_{..})^2 = \sum_i \sum_k \hat{\alpha_i}^2$. There is I term α_i linked by the relation $\sum_i \alpha_i = 0$ so $ddl_A = I - 1$.
Under the hypothesis H0, the test statistic follows an F-distribution $\mathcal{F}_{ddlR}^{ddlA}$.

- **Choice of risk.** We choose the risk α of being mistaken in rejecting H0.
- **Rule of decision.** Under H1, we expect the test statistic to be somewhat bigger under H1 than under H0, so we will reject H0 for "large values" in the test statistic $\mathcal{R} = \{F \in [f_{lim}; \infty[\}$ with f_{lim} the quantile of $1 - \alpha$.
- **Verification of conditions of application.** When writing out the model, we made the hypothesis that the residuals are independent, Gaussian and identically distributed according to the $\mathcal{N}(0, \sigma^2)$ distribution. We will check in a graph that this hypothesis is acceptable. The approach for checking if this is the case is explained in chapter 6.2 2-way analysis of variance.

We often sum up the results of the test in an ANOVA table:

Factor	ddl	SC*	SCM**	F	p-value
A	$I-1$	SSA	$SSA/ddlA$	$F_{A_{\text{obs}}}$	$P_{\text{H0}}(F_A > F_{A_{\text{obs}}})$
R	$N-I$	SSR	$SSR/ddlR$		

* Sum of squares.
* Sum of mean squares.

Tukey's HSD test.
If we reject the hypothesis H0, this means that at least one population differs from the others (on average). In this case, we will try to compare the means 2 by 2. If there is I population, this amounts to carrying out $I(I-1)/2$ tests. To do so, we can use the Tukey-Kramer HSD (*Honestly Significant Difference*) test.

- **Model.** The model is the linear model (6.2)
- **Multiple hypotheses (one per pair** (i,j)**)** H0: $\mu_i = \mu_j$, H1 : $\mu_i \neq \mu_j$.
- **Test statistic for each test**

$$T = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{(\frac{1}{n_i} + \frac{1}{n_j})SE^2}}$$

  We can see that this statistic is similar to the Student's t-test but that the variance is calculated with all the observations and not only the ones from samples from the populations i and j.
- **Choice of risk and rule of decision.** The distribution under H0 is given in table form and takes into account the fact that several tests have been carried out, and so we have artificially increased the value of risk of the first kind.

6.2 2-way analysis of variance

We saw above that the one-way ANOVA is a statistical method used to test whether the mean value of a quantitative variable depends on the variations of a qualitative variable (called a factor). If we bring out the fact that the mean value of the variable depends on the factor, we can then compare the means according to the groups defined by this factor. The groups can be defined by more than one factor. Here we will look at the case where the groups are defined by two factors. We can recall that the ANOVA cannot establish a causal link but only a statistical link.

6.2.1 2-way analysis of variance without interaction, balanced design

Issue. It may happen that the populations we are studying can be put into multiple categories. For example, we may want to know the fattening capacity of several cattle breeds in different fields. We can still describe the data using a linear model:

Model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2). \tag{6.4}$$

This time, the number of unknown means is equal to the sum of the number of levels of each factor (breed and field). We may want to separate the effect of the two factors. For example, we could ask whether there are differences between breeds, on the one hand, and between the fields, on the other. We may want to compare the sizes of the variations between fields and between breeds. In this case, it is simpler to identify each of the factors by rewriting the model (6.4) :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2), \qquad \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0. \tag{6.5}$$

Moving from model (6.4) to model (6.5) brings to light additional terms. μ is the general mean. α_i is the deviation from the mean for individuals in the breed i. β_j is the deviation from the mean for the individuals in field j. The parameters α_i and β_j are defined as deviations from the mean, and so the sum of α_i, as well as the sum of β_j, is zero. The α_i, as well as the β_j, are therefore not independent of each other. The residuals here are always assumed to be independent and to have the same distribution. We write as n_{ij} the number of observations in the group defined by the combination of factors (i,j): for this combination, $k = 1, ..., n_{ij}$.

This model is called an additive model, since we presume that the effects of the breed and of the field are added together. This means that we assume that the deviations between breeds are constant in all the fields, and that the deviations between fields are constant between the breeds.

Estimators of parameters We can estimate a number of parameters:

- $Y_{...}$ is the estimator of the general mean μ.
- $Y_{i..} - Y_{...}$ is the estimator of α_i, the deviation from the general mean of individuals in the breed i.
- $Y_{.j.} - Y_{...}$ is the estimator of β_j, the deviation the general mean of individuals in the field j.
- We can also calculate the estimated residuals: $Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...}$

Variances. The empirical variance of the Y is calculated from the total sum of squared deviations:

$$SST = \sum_i \sum_j \sum_k (Y_{ijk} - Y_{...})^2$$

Analysis of variance: the case of balanced design

For a balanced design, that is, when the number of observations is the same for each combination of factors ($\forall (i,j) \quad n_{ij} = n$), we can analyse SST as follows:

$$\begin{aligned} SST &= \sum_i \sum_j \sum_k ((Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...}) + (Y_{i..} - Y_{...}) + (Y_{.j.} - Y_{...}))^2 \\ &= \sum_i \sum_j \sum_k (Y_{i..} - Y_{...})^2 + \sum_i \sum_j \sum_k (Y_{.j.} - Y_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2 \\ &= SSA + SSB + SSR \end{aligned}$$

SSA depends on the parameters α and ϵ, SSB depends on the parameters β and ϵ, SSR only depends on ϵ.

Degrees of freedom. We call I the number of levels of factor A, J the number of levels of factor B and N the total number of observations. If we express the terms of the squared deviation sums SSA, SSB and SSR according to the parameters and constraints of the model (6.5), we will notice that there are $I-1$ independent terms for SSA ($ddl_A = I-1$), $J-1$ independent terms for SSB ($ddl_B = J-1$), and $N-1-ddlA-ddlB$ independent terms for SSR ($ddl_R = N-I-J+1$).

ANOVA tests
A reminder of the model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2), \qquad \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0.$$

We use the analysis of the sum of the squared deviations to carry out several independent statistical tests.

- **Effect of factor α**
  1. **Model.** The linear model (6.5).
  2. **Hypotheses H0 and H1.** H0: all the α_i are null, there is no factor effect. H1: at least one of the α_i is different from zero.

3. **Choice of a test statistic.** Under the hypothesis H0, the random variable $F = \frac{\frac{SSA}{ddlA}}{\frac{SSR}{ddlR}}$ follows an F-distribution $\mathcal{F}_{ddlR}^{ddlA}$.
4. **Rule of decision.** We reject H0 for the large values of F, that is, if the size of the inter-group variations defined by A ($\frac{SSA}{ddlA}$) is higher than the size of the residual variations $\frac{SSR}{ddlR}$. The limit will be the quantile $1-\alpha$ of the distribution $\mathcal{F}_{ddlR}^{ddlA}$.

- **Effect of factor β**

1. **Model.** The linear model (6.5).
2. **Hypotheses H0 and H1.** H0: all the β_j are null, there is not factor effect. H1: at least one of the β_j is different from zero.
3. **Choice of a test statistic.** Under the hypothesis H0, the random variable $F = \frac{\frac{SSB}{ddlB}}{\frac{SSR}{ddlR}}$ follows an F-distribution $\mathcal{F}_{ddlR}^{ddlB}$.
4. **Rule of decision.** We reject H0 for the large values of F, that is, if the size of the inter-group variations defined by B ($\frac{SSB}{ddlB}$) is higher than the size of the residual variations $\frac{SSR}{ddlR}$. The limit will be the quantile $1-\alpha$ of the distribution $\mathcal{F}_{ddlR}^{ddlB}$.

ANOVA table. The ANOVA table summarises all the statistical tests carried out.

Factor	*ddl*	*SS*	*CM**	*F*	*p*-value
A	$I-1$	SSA	$SSA/ddlA$	$F_{A_{\text{obs}}}$	$P_{\text{H0}}(F_A > F_{A_{\text{obs}}})$
B	$J-1$	SSB	$SSB/ddlB$	$F_{B_{\text{obs}}}$	$P_{\text{H0}}(F_B > F_{B_{\text{obs}}})$
R	$N-I-J+1$	SSR	$SSR/ddlR$		

* Mean Square.

The table enables us to make a direct conclusion about each hypothesis.

Conditions of application. The tests described above apply in the case of a balanced design. To test the effects of factors in the case of an unbalanced design, see 6.2.3. Moreover, we need to check the model's basic hypotheses, namely, that the residuals are independent and have the same distribution. To do so, we use graphs.

- **Graph of residuals.** To check that the residuals are independent, we represent the relation between the predicted values in graphic form:

$$\widehat{Y}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j$$

and the observed residuals:

$$\widehat{\epsilon}_{ijk} = Y_{ijk} - \widehat{Y}_{ij}$$

so, for each value predicted by the estimated mean of the group ij, $\widehat{Y}_{ij}$, we have n_{ij} residuals.

Under the hypothesis of independence, we expect to find no relation between the two measurements.

The figure 6.1 shows the graph obtained with a two-way model without interaction in the example for cows. For each combination of environment and breed factors, we have a predicted value and five different values for the residuals. We see that the deviations between the residual values are of the same size irrespective of the combination.

- **Quantile-quantile plot.** If the hypothesis of the normal character of the residuals proves to be correct, then $\frac{\epsilon_{ijk}}{\sigma} \sim \mathcal{N}(0;1)$. For each statistical individual in a set of data, we can calculate a reduced residual:

$$e_{ijk} = \frac{\widehat{\epsilon}_{ijk}}{\widehat{\sigma}}$$

To see whether these reduced residuals follow a normal distribution, we can calculate the empirical quantiles (or observed quantiles) of their distribution and compare them to the theoretical quantiles of the distribution $\mathcal{N}(0;1)$. We recall that the quantiles of normal distributions are tabulated in statistical analysis software and correspond to the inverse function of the cumulative distribution function. If the residuals do follow a normal distribution, we expect that the dots corresponding to the different pairs of quantiles (observed, theoretical) will be aligned in a straight line $y = x$. If we do not use non-standard residuals, we expect to see a linear relation between the observed quantiles and the theoretical quantiles (figure 6.1).

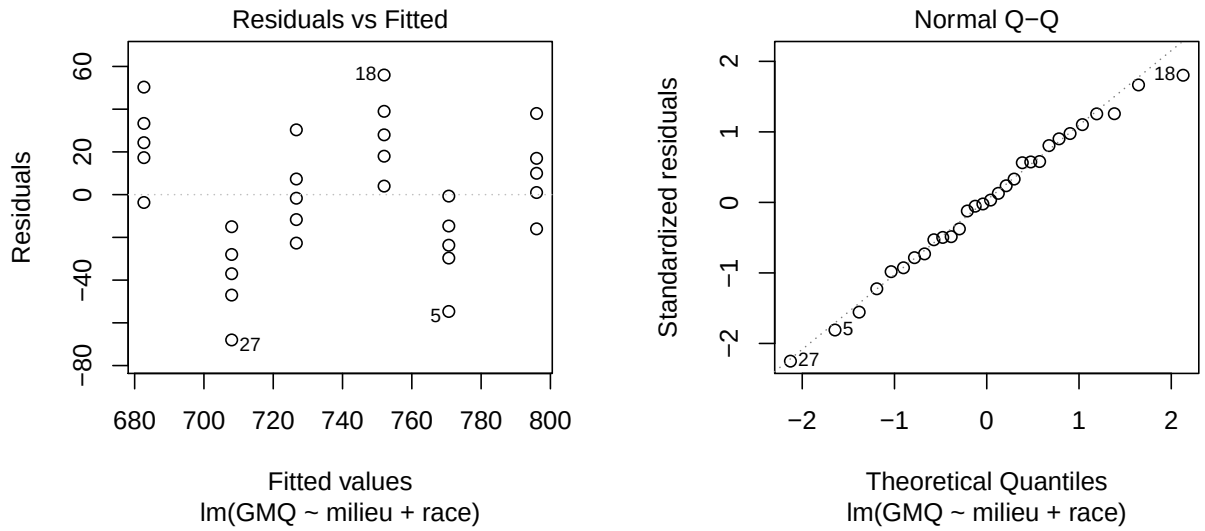


Figure 6.1: **ANOVA : model validation. LEFT : Graph of residuals.** We show the residuals according to the values predicted for the two-way model without interaction. Each dot corresponds to an observation for which we have calculated the value predicted by the model, on the x axis, and the deviation from the prediction, that is, the observed residual, on the y axis. **RIGHT : Quantile-quantile plot.** If the residuals are Gaussian, then the dots on the Q-Q plot should be aligned along a straight line, which is the case in this example.

6.2.2 Two-way analysis of variance with interaction, balanced design

Model. The model (6.5) assumes that the differences between breeds are constant, as are the differences between fields. We may choose to make the model more complex by introducing a term of interaction between the two factors, to take into account the fact that the differences between breeds are not necessarily identical in all fields:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2)$$
$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \theta_{ij} = \sum_j \theta_{ij} = 0.$$

In the example with cattle, the model includes the fact that the effect of the fields may depend on the breed (table below). The effect of interaction is the difference between the expected value predicted by the additive model and the expected valued of each population.

Field	Breed 1	Breed 2	Total
P1	$\mu+\alpha_1+\beta_1+\theta_{11}$	$\mu+\alpha_2+\beta_1+\theta_{21}$	$\mu+\beta_1$
P2	$\mu+\alpha_1+\beta_2+\theta_{12}$	$\mu+\alpha_2+\beta_2+\theta_{22}$	$\mu+\beta_2$
P3	$\mu+\alpha_1+\beta_3+\theta_{13}$	$\mu+\alpha_2+\beta_3+\theta_{23}$	$\mu+\beta_3$
Total	$\mu+\alpha_1$	$\mu+\alpha_2$	μ

Means. We can calculate several means

- $Y_{...}$ is an estimator of the general mean
- $Y_{i..}-Y_{...}$ is an estimator of α_i the deviation between the general mean and the mean of individuals in the breed i.
- $Y_{.j.}-Y_{...}$ is an estimator of β_j the deviation between the general mean and the mean of individuals of the field j.
- $Y_{ij.}-Y_{i..}-Y_{.j.}+Y_{...}$ is an estimator of θ_{ij} the effect of interaction between i and j, that is to say, the deviation from the added effects of the two factors.
- $Y_{ijk}-Y_{ij.}$ are the residuals for an n-sample

Analysis of variance: cases of balanced design.
As before, in the case of a balanced design, we can analyse the sum of the total squared deviations:

$$\begin{aligned}
SST &= \sum_i\sum_j\sum_k (Y_{ijk}-Y_{...})^2 \\
&= \sum_i\sum_j\sum_k [(Y_{i..}-Y_{...})+(Y_{.j.}-Y_{...})+(Y_{ij.}-Y_{i..}-Y_{.j.}+Y_{...})+(Y_{ijk}-Y_{ij.})]^2 \\
&= \sum_i\sum_j\sum_k (Y_{i..}-Y_{...})^2+\sum_i\sum_j\sum_k (Y_{.j.}-Y_{...})^2 \\
&\quad +\sum_i\sum_j\sum_k (Y_{ij.}-Y_{i..}-Y_{.j.}+Y_{...})^2+\sum_i\sum_j\sum_k (Y_{ijk}-Y_{ij.})^2 \\
&= SSA+SSB+SSI+SSR
\end{aligned}$$

where SSI is the sum of squares due to interaction.

Degrees of freedom. As before, we call I the number of levels of factor A, J the number of levels of factor B and N the total number of observations. If we express the terms of the sums of squared deviations SSA, SSB, SSI and SSR according to the parameters and constraints of the model (6.5), we notice that there are $I-1$ independent terms for SSA ($ddl_A = I-1$), $J-1$ independent terms for SSB ($ddl_B = J-1$), $(I-1)(J-1)$ independent terms for SSI ($ddl_I = (I-1)(J-1)$) and $N-1-ddlA-ddlB-ddlI$ independent terms for SSR ($ddl_R = N-IJ$).

ANOVA tests
As before, we use the analysis of the sum of the squared deviations to carry out several independent statistical tests. Along with the effects of the factors A and B, we carry out an additional test for the **effect of interaction**. Below, we will only describe the latter:

1. **Model.** The linear model (6.5)
2. **Hypotheses H0 and H1.** H0: all the θ_{ij} are null, there is no effect of interaction. H1: at least one of the θ_{ij} is different from zero.
3. **Choice of a test statistic.** Under the hypothesis H0, the random variable $F = \frac{\frac{SSI}{ddlI}}{\frac{SSR}{ddlR}}$ follows an F-distribution $\mathcal{F}_{ddlR}^{ddlI}$, with $ddlI = (I-1)(J-1)$.

4. **Rule of decision.** We reject H0 for the large values of F. If there is no effect of interaction (effect θ), $\frac{SSI}{ddlI}$ and $\frac{SSR}{ddlR}$ are two possible estimations for the residual variance. On the other hand, if it exists, $\frac{SSI}{ddlI}$ will be larger than $\frac{SSR}{ddlR}$.

ANOVA table. The ANOVA table sums up all the statistical tests that have been carried out:

Factor	ddl	SS	CM	F	p-value
A	$I-1$	SSA	$SSA/ddlA$	$F_{A_{\text{obs}}}$	$P_{\text{H0}}(F_A > F_{A_{\text{obs}}})$
B	$J-1$	SSB	$SSB/ddlB$	$F_{B_{\text{obs}}}$	$P_{\text{H0}}(F_B > F_{B_{\text{obs}}})$
I	$(I-1)(J-1)$	SSI	$SSI/ddlI$	$F_{I_{\text{obs}}}$	$P_{\text{H0}}(F_I > F_{I_{\text{obs}}})$
R	$N-IJ$	SSR	$SSR/ddlR$		

Using the table, we can make conclusions directly for each hypothesis. Please note: the introduction of additional parameters in the model has reduced the number of degrees of freedom of the residual.

Conditions of application. We need to check the basic hypotheses of the model, namely that the residuals are independent and have the same distribution. To do so, we make use of graphs like the graph of residuals and the quantile-quantile plot (figure 6.1).

Interpretation. In the example in the figure 6.2, the variations in the variable X are explained by two factors, A (two levels, $A1$ and $A2$) and B (two levels, $B1$ and $B2$). We can calculate four means $X_{ij.}$. The expected value of these means is equal to:

$$E(X_{ij.}) = \mu + \alpha_i + \beta_j + \theta_{ij}$$

We can show these means in a graph according to the levels of factor B, by choosing a different colour according to the levels of factor A.

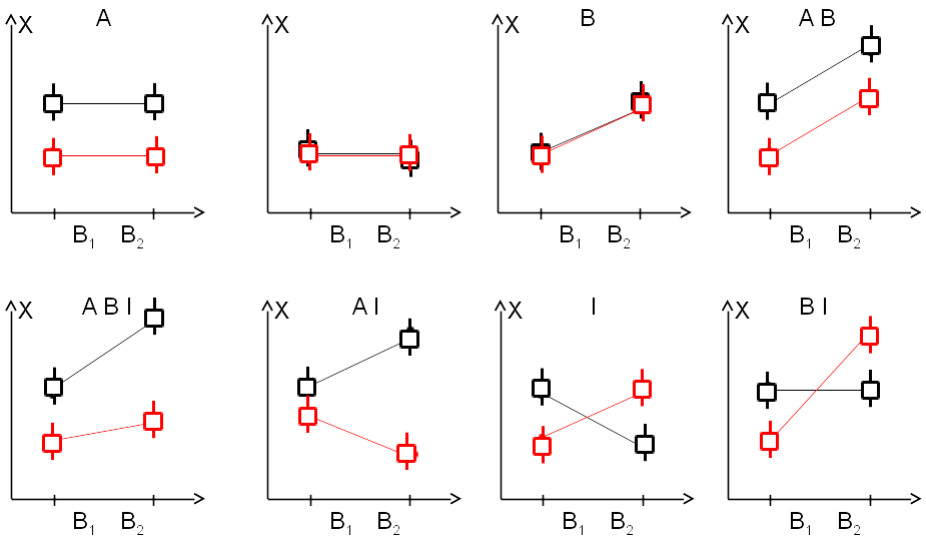


Figure 6.2: **Graph showing interaction in a case with two factors.** Each graph corresponds to a possible result of the ANOVA. Significant factors (A,B,I) are shown. An ANOVA result corresponds to a representation of the means of each combination of factors. The x axis is the level of factor B. The y axis is the calculated mean in the sample. For each situation, the means are shown by squares. The vertical line corresponds to the residual standard variation $\widehat{\sigma}_R$. The four graphs at the top correspond to a lack of interaction. The four graphs at the bottom correspond to situations with significant interaction.

We can notice several things:

- When the interaction is not significant, the differences between $A1$ and $A2$ are constant, whatever the level of factor B. Similarly, the differences between $B1$ and $B2$ are constant, whatever the level of factor A. The straight lines run in parallel.

- When no factor is significant, all means are equal.

- A significant interaction effect corresponds to deviations between the means of a factor that varies according to the level of the other factor. The straight lines are no longer parallel. In other words: the effect of factor A depends on the level of B and vice versa.

6.2.3 Two-way analysis of variance, unbalanced design

When the design is unbalanced, that is, when we do not have the same number of observations for each combination of factors, we cannot note the analysis of variance as before. In this case, we can test the effect of a factor and the interaction by comparing the interlocking models. Different approaches, which we will call Type I, Type II and Type III, are possible. Here we will show type I and type III.

a) Models
We will look at a series of models.

$$\begin{array}{lcl} \text{M0} & : & Y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ijk} \\ \text{M1} & : & Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \\ \text{M2} & : & Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk} \\ \text{M3} & : & Y_{ijk} = \mu + \epsilon_{ijk} \\ \text{M4} & : & Y_{ijk} = \mu + \alpha_i + \theta_{ij} + \epsilon_{ijk} \\ \text{M5} & : & Y_{ijk} = \mu + \beta_i + \theta_{ij} + \epsilon_{ijk} \end{array} \tag{6.6}$$

For each model, we assume (when the parameter is used in the model):

$$\epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2) \qquad \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \theta_{ij} = \sum_j \theta_{ij} = 0.$$

We say that two models interlock if, when we place constraints on the parameters of one of them, we obtain the other. For example, M1 is interlocked in M0, because when we make the hypothesis $\forall(i,j) \quad \theta_{ij} = 0$ in M0, we obtain M1.

b) Test of comparison of two interlocked models

We will take the example of the comparison between M0 and M1.

Model We will consider the model M0 : $Y_{ij} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ij}$

Hypotheses tested $H0 : \forall(i,j) \quad \theta_{ij} = 0$ versus $H1 : \exists(i,j) \quad \theta_{ij} \neq 0$

Test statistic and distribution under H0 If H0 is true, then the residual variance σ^2 is the same for both models. However, if H0 is false, we expect the residual variance of the model M0 to be bigger than that of the model M1, since it describes the data less well. So we use a test statistic to compare the residual variances in both models. In a linear model, the residual variance is estimated by the sum of squared residuals: For the model M0, $SSR_0 = \sum_i \sum_j \sum_k (Y_{ijk} - \widetilde{Y}_{ij_{M0}})^2$, where $\widetilde{Y}_{ij_{M0}} = Y_{ijk} = \widehat{\mu} + \widehat{\alpha_i} + \widehat{\beta_j} + \widehat{\theta_{ij}}$ is the value predicted by the model M0 with $\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}$ the estimators of α, β, γ. For the model M1, $SSR_1 = \sum_i \sum_j \sum_k (Y_{ijk} - \widetilde{Y}_{ij_{M1}})^2$, where $\widetilde{Y}_{ij_{M1}} = Y_{ijk} = \widehat{\mu} + \widehat{\alpha_i} + \widehat{\beta_j}$ is the value predicted by the model M1 with $\widehat{\alpha}, \widehat{\beta}$ the estimators of α, β. If H0 is true, then the difference $(SSR_0 - SSR_1)$ should be "close" to zero. We use the test statistic:

$$F = \frac{\frac{SSR_1 - SSR_0}{(I-1)(J-1)}}{\frac{SSR_0}{n-IJ}}$$

Under the hypothesis H0, F follows an F-distribution: $F \sim \mathcal{F}_{((I-1)(J-1),N-IJ)}$. In this statistic, we divide the difference $(SSR_0 - SSR_1)$ by the difference of degrees of freedom of each term and SSR_0 by its number of degrees of freedom.

Choice of risk and rejection region We will reject H0 for the large values of F. The limit of the rejection region is the quantile $1 - \alpha$ in the distribution $\mathcal{F}_{((I-1)(J-1),N-IJ)}$

c) Testing the effects
To test the effect of the factors A, B and their interactions, we can proceed in several different ways. By default, anova(lm()) in **R** uses Type I.

Type I We will add the effects as we go along. In this case, the order in which we add the factors is important. To test the factor A, we will compare the models M3 and M2. Then, to test the effect of factor B, we will compare the models M2 and M1. Lastly, to test the interaction, we will compare the models M1 and M0, as described above. We will use this method when there is a natural order of factors and there will be of no point in testing the effect of B without taking into account A.

Type III In this approach, to test the effect of interaction, we will compare M0 and M1, as in the analysis of Type I. However, to test the effect of A, we will compare M0 and M5. To test the effect of B, we will compare M0 and M4.

6.3 Linear regression

When studying a biological system (metabolic network, nervous system, forest ecosystem, etc.), we can characterise it using several variables with values collected through a specific study (experiment, field study, clinical observations, etc.). For example: we want to study life history strategies in the yeast *Saccharomyces cerevisiae*. We conduct experiments to measure growth rate, the size of cells, the biotic capacity, the rate of resource consumption and the mortality rate in different strains of yeast. Each strain will be characterised by these different totals. Often, the question asked is: "What are the relations between these variables?". One method used to answer this question is linear regression. With this method we can study the linear variations of a quantitative variable according to the variations of one or several other quantitative variables. We will begin by describing the relation between two variables, that is, the *simple linear regression*, then we will generalise by discussing *multiple linear regression*.

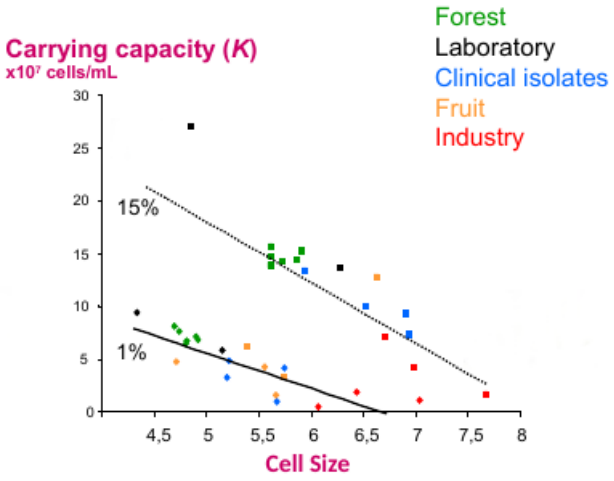


Figure 6.3: **Biotic capacity and average size of cells** of different strains of yeast *S. cerevisiae* measured using batch cultures initially containing 1% (lozenges) or 15% (squares) of glucose . The colour of the dots depends on the ecological niche where the strains have been sampled. (based on Spor *et al.*, 2009, BMC Evol. Biol. 9:296)

6.3.1 Simple linear regression

If we take the example presented in the introduction, we can measure the average size of cells in different strains of yeast and their biotic capacity, that is, the maximum size of a population at a given level of resources.
In the figure 6.3, if we study each environment (1% or 15% glucose) separately, we see that the biotic capacity decreases in a linear way with the size of cells. It seems "natural" to draw a straight line in the middle of the cluster of dots. To say there is a linear relation means that we know the average size of a cell for a given strain, and we will be able to predict its biotic capacity from a linear relation. Of course, this prediction will not be the same as the value observed, but the prediction should be better than if we used the mean biotic of all the strains as a prediction. So the linear regression can provide an equation for this straight line and quantify the margin for error on predictions.

6.3.1.1 Model

To present the formulation, we will use data from a study designed to estimate the effect of water supply on the yield of parcels of wheat. In the figure 6.4 showing these data, we can see 30 dots corresponding to 30 parcels of land that have been watered to different degrees and where we can observe yields.

We will take a note of:

- n the number of observations ($n = 30$)
- Y_i the yield from the i^{th} parcel of land sampled ($i = 1$ to n)
- x_i amount of water supplied to the i^{th} parcel of land sampled ($i = 1$ to n)

We consider that we can predict the yield Y_i by the supply of water x_i with the possibility of error ϵ_i. So we can write:

$$Y_i = a + bx_i + \epsilon_i$$

where the residuals ϵ_i are independent Gaussian random variables and distributed identically according to $\mathcal{N}(0; \sigma^2)$ (σ^2 unknown). The part $a + bx_i$ is called the *prediction*. We say that the yield is the *explained variable* and the supply of water is the *explanatory variable*. Another way of writing down this model is to say that the variables Y_i are independent Gaussian random variables and distributed according to $\mathcal{N}(a + bx_i; \sigma^2)$.

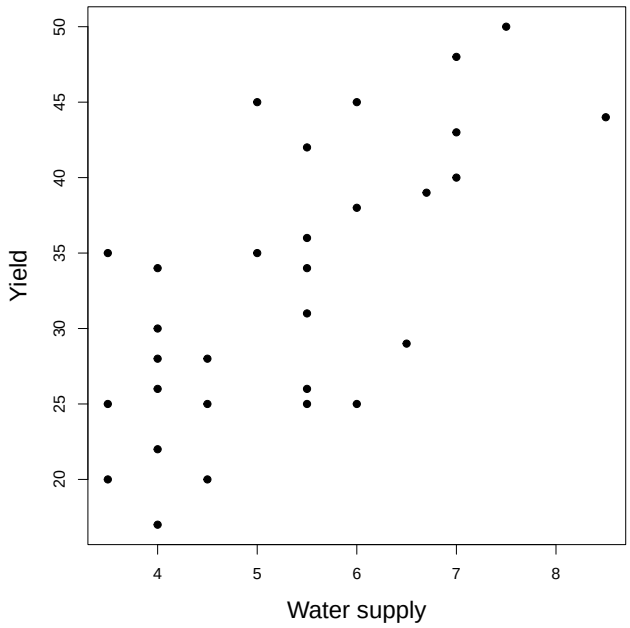


Figure 6.4: **Yield of wheat from parcels of land according to water supply.** Each dot links the supply of water in a piece of land to the yield measured on the same piece of land.

6.3.1.2 Parameter estimators

We want to estimate the unknown values a and b in data from the sample. We can look for the straight line that is "closest" to all the dots by defining a criterion with which to quantify the distance between the dots and the straight line (figure 6.5). We call this criterion the *squared error*, defined as the sum of the squared sum of distances between Y_i and its prediction $a + bx_i$.

$$\text{Squared error} = SSR = \sum_{i=1}^{n} (Y_i - (a + bx_i))^2$$

The aim is to find the values of a and b that will minimise the squared error. We can also define the mean of the sample for the variable to be explained $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, and the mean of the sample for the explanatory variable $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
The values of a and b that will minimise the squared error are values that cancel the derivatives $\frac{\partial SSR}{\partial a}$ and $\frac{\partial SSR}{\partial b}$ (providing that the second derivatives are positive). By resolving the system of two equations

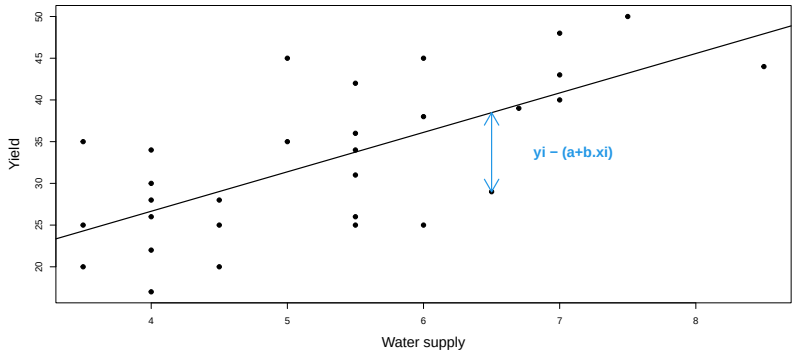


Figure 6.5: **Representation of the distance between a measurement and the linear regression.** The parameter estimators a and b of the linear regression are chosen so as to minimise the squared sum of these distances.

to two unknowns thus defined, we find that the estimators of a and b that minimise the squared error are:

$$\widehat{a} = \overline{Y} - \widehat{b}\overline{x} \tag{6.7}$$

$$\widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{6.8}$$

$\widehat{a}$ and $\widehat{b}$ are calculated from observations of x and y in the sample.

6.3.1.3 Hypothesis testing on the model

We want to test the hypothesis H0 : $b = 0$ compared with H1 : $b \neq 0$. We can note that under H0, the predicted value of Y is a constant a which is estimated as the empirical mean of Y, $\overline{Y}$.

Analysis of the variance. The linear regression model seeks to explain the variations of Y from the variation of x. To do so, we can analyse the variations of Y in two sources of variation, the variations explained by the variations of the explanatory variable and the residual variations. Below, we write $\widehat{Y_i} = \widehat{a} + \widehat{b}x_i$ the prediction of the model, and $\overline{Y}$ the prediction of the model under the hypothesis H0.

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 + \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2$$

If we have $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSres = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$ and $SSreg = \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2$, we can write:

$$SST = SSreg + SSres.$$

Test statistic and distribution under H0. Under the hypothesis H0, the test statistic

$$F = \frac{\frac{SSreg}{1}}{\frac{SSres}{n-2}}$$

follows an F-distribution $\mathcal{F}_{1,n-2}$.

6.3.1.4 Model's consistency with data

To assess the quality of the model, we need to: 1) check *a posteriori* if the hypotheses provided by the model are substantiated; 2) estimate the degree of variance of Y explained by the model.

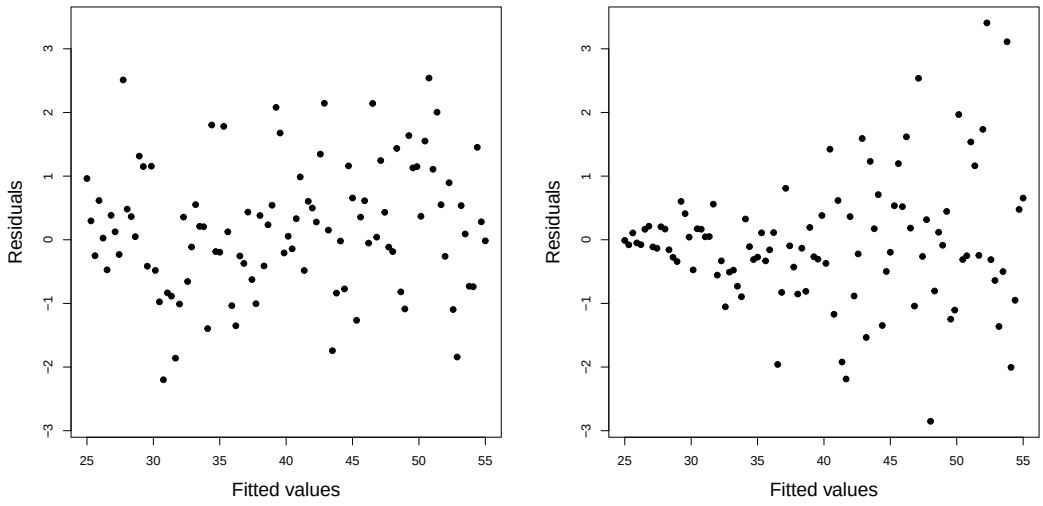


Figure 6.6: **Checking the residuals' homoscedasticity**. We need to check that the variance of the residuals does not depend on the predicted values. In these graphs, we cannot see the variance directly, but we can get an idea of the variance by observing the deviation of the residuals from the mean (0). On the left-hand graph, the residuals are homoscedastic. On the right-hand graph, the variance of the residuals increases with the predicted values.

Analysis of residuals. One of the hypotheses underlying the model is that the residuals are independent, Gaussian and identically distributed. In particular, we make the hypothesis that the mean and the variance of the residuals do not depend on the predicted values. When the variance does not depend on the predicted values, we use the term *homoscedasticity*. In the opposite case, we use the term *heteroscedasticity*. To find out which is the case, we can examine a Q-Q plot of the standardised residuals, as well as the graph showing the observed residuals $e_i = y_i - (\widehat{a} + \widehat{b}x_i)$ according to the predicted values $\widehat{y}_i = \widehat{a} + \widehat{b}x_i$ where $\widehat{a}$ and $\widehat{b}$ are the estimations calculated in line with the equations (6.7) and (6.8).

Coefficient of determination. The degree of variance explained by the model is:

$$R^2 = \frac{SSreg}{SST}.$$

We call it the *coefficient of determination*. It varies between 0 and 1.
The figure 6.7 shows two models of linear regression, estimated from two different sets of data.

Estimator of the residual variance. An estimator for the residual variance of the model is:

$$\widehat{\sigma^2} = \frac{SSres}{n-2}$$

6.3.1.5 Tests and confidence interval for the parameters

Estimator distribution $\widehat{a}$ and $\widehat{b}$, confidence intervals and tests on parameters. The estimators of the a and b parameters are random variables. The estimator distribution is given respectively by:

$$\frac{\widehat{a} - a}{\widehat{Var}(\widehat{a})} \sim \mathcal{T}_{n-2} \quad \text{and} \quad \frac{\widehat{b} - b}{\widehat{Var}(\widehat{b})} \sim \mathcal{T}_{n-2},$$

where $\widehat{Var}(\widehat{a})$ and $\widehat{Var}(\widehat{b})$ are the variances of $\widehat{a}$ and $\widehat{b}$.
We can deduce an interval of confidence of $1 - \alpha$ for each of the parameters:

$$IC(a) = [\widehat{a} - t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{a})}; \widehat{a} + t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{a})}]$$

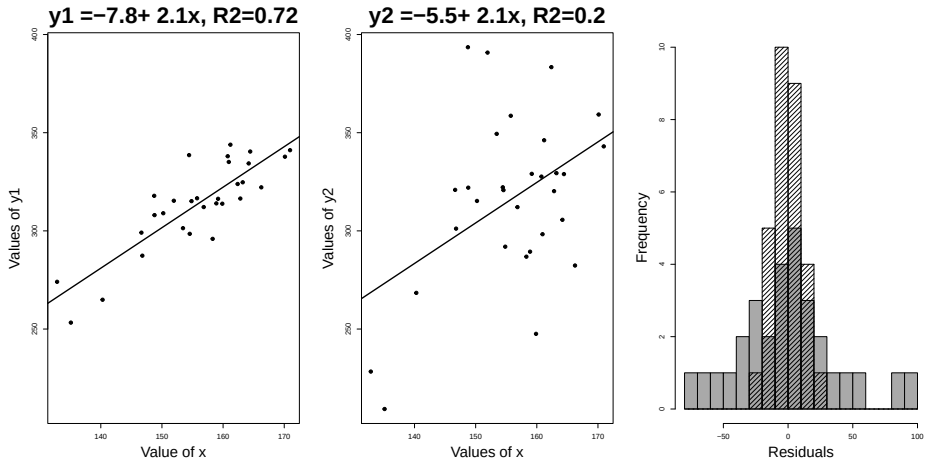


Figure 6.7: **Models of regression for two sets of simulated data.** The equations of the two straight lines are very similar, but the dots are less scattered for $y1$ than for $y2$. The degree of explained variance is higher in the first case. The right-hand graph represents the empirical distribution of residuals for the variable $y1$ (hatched) and for $y2$ (grey).

$$IC(b) = [\widehat{b} - t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{b})}; \widehat{b} + t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{b})}].$$

We can also test for the nullity of each parameter. For example, for the nullity of parameter a:

$$\text{H0} : a = 0 \text{ as against H1} : a \neq 0$$

We can take as a test statistic: $T = \dfrac{\widehat{a}}{\sqrt{\widehat{Var}(\widehat{a})}}$. Under the hypothesis H0, the test statistic follows a Student's t-distribution $\mathcal{T}_{n-2}$.

R Output

The figure 6.8 represents the output obtained with the R software using the function `lm(yield~water,data=tab)`.

In the table `tab`, `yield` is a column showing the observations of the yield and `water` a column with the water supplied. The line `intercept` concerns the parameter a. The line `water` concerns the parameter b (the parameter multiplying the supply of water in the model). The column `Estimate` gives estimations of the two parameters a and b, the column `Std.Error` gives an estimation of the standard errors in the estimators $\sqrt{\widehat{Var}(\widehat{a})}$ and $\sqrt{\widehat{Var}(\widehat{b})}$. The columns `t-value` and `p-value` give, respectively, the observed value of the test statistic and of its p-value for the tests for nullity of a and b described above. In the second boxed text, `Residual standard error` is an estimation of σ and `Multiple R-squared` provides the determination coefficient R^2.

The estimation of a (7.75) corresponds to the original y-axis. The estimation of b (4.73) corresponds to the gradient of the straight line. When we examine the p-value, we can see that the parameter a is not significantly different from 0, whereas the parameter b is significantly different from 0. So, we can interpret its value: if we increase the supply of water by one unit, the yield will increase by 4.73. Independently of the fact that a is not significantly different from 0, we can notice that the original y-axis has no biological meaning: the model has been validated for values of water between 3 and 9, and the linear relation is probably not true if the supply of water is too low. According to this model, the variations in the supply of water explain $R^2 = 48\%$ for the variance in the yield. We can thus ask whether other variables might not provide extra information that would help improve the quality of our prediction.

```
> reg <- lm(rendement~eau,data=tab)
> summary(reg)

Call:
lm(formula = rendement ~ eau, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max
-11.1109  -4.2246  -0.0836   3.5482  13.6164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7473     5.1099   1.516    0.141
eau           4.7273     0.9352   5.055 2.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.667 on 28 degrees of freedom
Multiple R-squared:  0.4771, Adjusted R-squared:  0.4585
F-statistic: 25.55 on 1 and 28 DF,  p-value: 2.388e-05
```

Figure 6.8: **R output for linear regression**

6.3.2 Multiple linear regression

Often the variations in the explained variable depend linearly on the variations of several variables and not just on one, as in the previous section. For example, the yield may depend not only on the water supply, but also on the supply of nitrogen, on the temperature or soil acidity. We will use this example to show the procedure for the multiple regression model.

6.3.2.1 The model

As in the case for a single regression, we write Y_i for the yield of the i^{th} piece of land sampled ($i = 1$ at n). This time we consider that the yield depends in a linear way on the supply of water to the piece of land x_{1i}, the supply of nitrogen x_{2i}, soil acidity x_{3i} and temperature x_{4i}. The previous model can thus be modified as follows:

$$Y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i} + \epsilon_i$$

where, as in the case of the model for simple regression, the residuals ϵ_i are independent Gaussian random variables and identically distributed according to $\mathcal{N}(0; \sigma^2)$.

6.3.2.2 F-test

The hypotheses tested are:

$$\text{H0} : b_1 = b_2 = ... = b_p = 0 \text{ as against H1} : \exists j, \quad b_j \neq 0$$

The test is based on the analysis of the variance shown in the previous section, which is still valid in the case of the multiple regression model. The element $SSreg$ has p degrees of freedom, the element $SSres$ has $n - p - 1$ degrees of freedom. The test statistic is:

$$F = \frac{SSreg/p}{SSres/(n-p-1)}.$$

Under the hypothesis H0, this statistic follows an F-distribution F^p_{n-p-1}. To choose the rejection region and to perform the test, the procedure is the same as for the ANOVA tests above.

Validation of conditions of application. As before, we need to check that the residuals really are independent, Gaussian and identically distributed by examining the observed residuals. We can also run a Fisher's exact test to see whether the predictions from the model are better than if we predicted yield only by average yield.

6.3.2.3 Estimators of parameters and tests for nullity on each parameter

We will not give details here about the estimators for each parameter. If we give $b_0 = a$ and p the number of explanatory variables (here $p = 4$), the distribution of these estimators is given by:

$$T = \frac{\widehat{b_j} - b_j}{\widehat{Var}(\widehat{b_j})} \sim \mathcal{T}_{n-p} \text{ with } j = 1, ..., p$$

We can thus build the p tests:

$$\text{H0} : b_j = 0 \text{ as against H1} : b_j \neq 0 \text{ with } j = 1, ..., p$$

6.3.2.4 Selection of a model

Issue. We can try to find the combination of explanatory variables which best predict the observations. This is a complicated problem since the number of possible combinations increases very quickly along with the number of descriptors.
For example, we may want to compare a model predicting the yield from the water and nitrogen content in the soil (Ma) with a model predicting the yield from the temperature alone (Mb):

$$\text{Ma} \quad : \quad Y_i = a + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$$

$$\text{Mb} \quad : \quad Y_i = a + b_4 x_{4i} + \epsilon'_i$$

Since the explanatory variables are not the same, the residuals will differ in the two models. A criterion for choosing could be the model with the lowest residual variance or the best determination coefficient. However, when we add descriptors with a significant effect on the characteristic of interest to us, the residual variance necessarily decreases, since we will describe the observations better and better. Thus, if we knew absolutely all the factors determining the yield of wheat, we would be able to predict Y with no error. *A good compromise consists in choosing the model that best describes the data with the fewest possible parameters.*

Likelihood. The likelihood of a model is the probability of observing the values of a sample with the given values of the parameters. We can calculate it by using an estimation of the parameters of the model and the probability distribution of the random variables. Within the framework of a linear model, if the residuals are Gaussian, there is a relation between the likelihood of the model and the sum of squared residuals.

Akaike information criterion. The AIC is calculated as the difference between the model's likelihood function and a penalty proportionate to the number of parameters to be estimated:

$$AIC = 2p - 2\ln(L)$$

where p is the number of parameters of the model and L the likelihood. The best model is the one with the lowest AIC.
A strategy for the multiple linear regression consists in testing all the models and choosing the best one on the basis of the AIC.
Other strategies exist, which involve selecting models repeatedly. We can start with the most simple model and try at each step to remove or add a descriptor, basing ourselves on the AIC for choosing the descriptor to remove or add.
We should end up with a subset of descriptors giving the best observations about the variables we are interested in.

Example. In the example about wheat, here is the result of the final stage of selection. The procedure involves successively adding all the explanatory variables. The last step consists in trying to remove one of them:

```
Step:  AIC=109.11
yield = water + nitro + temp + acid
```

Model	Df	Sum of Sq	RSS	Cp
<none>		816.31	388.15	109.11
- acid	1	39.59	855.90	108.53
- temp	1	80.59	896.90	109.93
- water	1	232.35	1048.66	114.62
- nitro	1	384.21	1200.52	118.68

In the column `Model`, the line `<none>` corresponds to the model chosen at the previous step. In this case, it is the complete model. The other lines correspond to the complete model minus one of the descriptors. For each sub-model, `Df` is the number of degrees of freedom gained or lost, `Sum of Sq` is the regression sum of squares ($SSreg$), `RSS` is the residual sum of squares ($SSres$) and `Cp` gives the Akaike Information Criterion for each sub-model.
We can see that the best model is the one including all the descriptors except soil acidity ($AIC = 108.53$).
The following step consists in examining the estimated coefficients.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.98903	9.58214	-2.086	0.04693 *
water	3.54548	1.05064	3.375	0.00233 **
nitro	0.07036	0.02078	3.386	0.00226 **
temp	1.26031	0.85770	1.469	0.15372

We see that all the coefficients are positive. The descriptors are thus positively correlated to the yield. On can note that the coefficient of the temperature variable does not significantly differ from zero. It is likely that the fact of adding this variable helps to increase the likelihood of the model (and thus decreasing the AIC), even if we cannot show that the coefficient is significantly different from 0 with a test.

6.4 Analysis of covariance

6.4.1 Example: allometric relationships in fish

The article by Yu et al. (BMC Evolutionary Biology, 2014, 14:178) presents a study about the allometric relationships between body size and brain size in vertebrates. In a single species, we can typically see an allometric relationship in the form:

$$W = CP^b$$

where W is the brain mass, P the body mass and C a constant. b is called the *allometric coefficient*. We can carry out a change in the variables to arrive at a logarithmic scale:

$$\ln(W) = \ln(C) + b\ln(P)$$

On the log scale, we can see that brain mass increases proportionally to body mass. The authors looked at the factors that might explain the changes in the allometric coefficient between species, helping us to understand the big difference in brain size between, for example, mammals and birds. Their hypothesis is that there is an energy cost in increasing brain size. They show that we can differentiate endothermic species (mammals, birds, insects) from ectothermic species (fish, reptiles, amphibians), with higher coefficients in endothermic species ($C = 0.078$, $b = 0.689$) than in ectothermic species ($C = 0.014$, $b = 0.578$). To check on the role of temperature in the body-brain allometric relationships, they sampled databases (`http://fishbase.org`) to represent the different fish species and wide-ranging environments in terms of water temperature (Table 7.2).

Environment	Average temperature	Nb species
Polar	1°C	34
Temperate	15°C	70
Tropical	25°C	88
Sub-tropical*	20–30°C	17

* The species taken into account in this environment are 17 species of sharks maintaining a body temperature between 20 and 30 °C, using a system of muscular contractions.

Table 6.1: **Sample of 209 species living at different temperatures.**

The figure 6.9 is taken from the article and shows the allometric relationships for each group of species on a logarithmic scale.
We can see a linear relationship for each environment between body mass and brain mass. However, species living in different environment seem to behave differently, with variations in the constant C and in the allometric coefficient b.

6.4.2 Nested models and hypotheses tests

To simplify, we call Y the random variable corresponding to the logarithm of the brain mass, X the logarithm of the body mass, and $a = \ln(C)$. We have a random sample of species in each environment. We note the environments by i ($i = 1, ..., p$, $p = 4$), and the species in an environment by j. The pair (Y_{ij}, X_{ij}) constitutes two continuous random variables. We can put forward the following model, which takes into account a difference depending on the habitats for the constant a and the allometric coefficient b:

$$Y_{ij} = a_i + b_i X_{ij} + \epsilon ij \tag{6.9}$$

with ϵ_{ij} the random variable describing the residuals of the model which we presume are independent and have the same distribution $\mathcal{N}(0; \sigma^2)$.

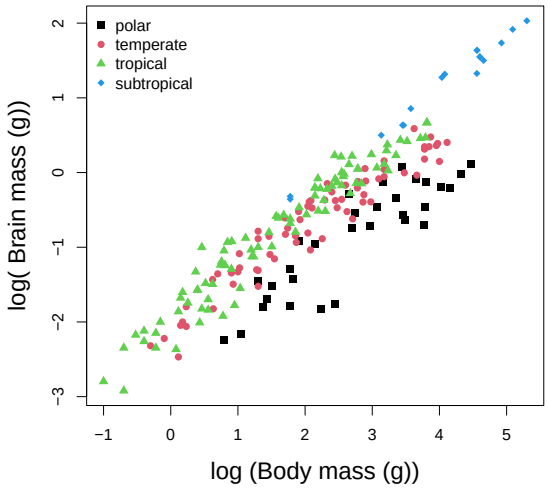


Figure 6.9: **Body-brain allometric relations in fish.** Each dot corresponds to a species, and each colour to a type of habitat, as described in the Table 7.2. We recall that the "sub-tropical" environment corresponds to a sample of 17 species of sharks.

6.4.2.1 Hypothesis test

The questions raised are as follows:

- Are there any differences between the environments for the y scale at the origin a ? The two hypotheses are

$$\begin{aligned} \text{H0} \quad &: \quad a_1 = a_2 = ... = a_p = a \\ \text{H1} \quad &: \quad \exists (i, i'), \quad a_i \neq a_{i'} \end{aligned}$$

- Are there any differences between the environments for the allometric coefficient b? The two hypotheses are

$$\begin{aligned} \text{H0} \quad &: \quad b_1 = b_2 = ... = b_p = b \\ \text{H1} \quad &: \quad \exists (i, i'), \quad b_i \neq b_{i'} \end{aligned}$$

Let us look at the second question. We can use a model to describe each hypothesis:

$$\text{M0} \quad : \quad Y_{ij} = a_i + bX_{ij} + \epsilon_{ij}$$

and

$$\text{M1} \quad : \quad Y_{ij} = a_i + b_i X_{ij} + \epsilon_{ij}^{'}$$

If H0 is true, then the residual variance σ^2 is the same for both models. However, if H0 is false, we expect that the residual variance of the model M0 should be bigger than the variance of model M1, since it describes the data less well.
M0 and M1 are two nested models. The model M0 is a sub-model of M1, with $b_i = \text{Cte}$. We can thus develop a test statistic to compare the residual variances in the two models.
In a linear model, the residual variance is estimated by the sum of squared residuals:
For the model M0, $SSR_0 = \sum_i \sum_j (Y_{ij} - \widehat{Y}_{ij_{M0}})^2$, where $\widehat{Y}_{ij_{M0}} = \widehat{a}_i + \widehat{b} X_{ij}$ is the value predicted by the model M0.
For the model M1, $SSR_1 = \sum_i \sum_j (Y_{ij} - \widehat{Y}_{ij_{M1}})^2$, where $\widehat{Y}_{ij_{M1}} = \widehat{a}_i + \widehat{b}_i X_{ij}$ is the value predicted by the model M1.
If H0 is true, then the difference $(SSR_0 - SSR_1)$ should be close to zero. We can note that $SSR_0 - SSR_1$ corresponds to the sum of squared residuals corresponding to the effects b_i.

In this case, the tests statistic is

$$F = \frac{\frac{SSR_0 - SSR_1}{p-1}}{\frac{SSR_1}{n-p-1}} \sim_{H0} \mathcal{F}_{p-1,n-p-1}.$$

6.4.2.2 Comparison of the two models: results table.

The table of results is presented as follows. The figures given here correspond to the allometric data described in the Table 6.1 for $n = 209$ species.

Model	res.df	SSR	model.df	SSM	F	pvalue
M0	204	11.30				
M1	201	10.78	3	0.52	0.23	0.023

For each model, we give the sum of squared residuals (`SSR`) and the corresponding degrees of freedom (`res.df`). For the most complete model, we show the number of additional parameters (`model.df`) and the deviation $SSM = SSR_{\text{M0}} - SSR_{\text{M1}}$, then the value of the test statistic `F` and the associated `pvalue`. We find here that the model M1 has three more parameters than the model M0 (`model.df`). When we study different gradients we can slightly reduce the sum of squared residuals (compare the column `SSR` for M0 and M1). Here, under H0, the statistic F follows a distribution $\mathcal{F}_{3.201}$. The p-value is 0.023. So we can reject H0 at the threshold of 5%.

6.4.2.3 Comparison of several nested models and AIC

In the present example, we can write several nested sub-models to test out different hypotheses on the allometric coefficients or on the constants. As before, we can then compare the models two by two using a Fisher's exact test. We can also choose the model that best describes the data using a criterion such as the Akaike Information Criterion (see the section on multiple regression).
The table below shows different models derived from the (6.9) model, the estimated number of parameters (including σ^2) and the value obtained for the AIC.

Model	Predicted value	Nb parameters	AIC
M111	$\widehat{Y}_{ij} = \widehat{a}_i + \widehat{b}_i X_{ij}$	9	–8,50
M110	$\widehat{Y}_{ij} = \widehat{a}_i + \widehat{b} X_{ij}$	6	–4,66
M011	$\widehat{Y}_{ij} = \widehat{a} + \widehat{b}_i X_{ij}$	6	8,32
M100	$\widehat{Y}_{ij} = \widehat{a}_i$	5	508,3
M001	$\widehat{Y}_{ij} = \widehat{a} + \widehat{b} X_{ij}$	3	202,6

Table 6.2: **Analysis of covariance: comparison of nested models.** For each model, the residual variance may be different. The complete model (M111) here is the one with the lowest AIC.

As an example, the figure 6.10 gives the predicted values for the three models tested: the complete model (M111), the model M110 and the model M100. We can see that the best model is the one that is most complete, M111, which predicts both the differences between the environments for the allometric coefficient and for the constant. The graph of the model M110 shows the predicted values by assuming a single allometric coefficient for all the species. Visually, we have the impression of a correct adjustment, which explains the relatively high p-value of 0.02 when we compare the models M111 and M110. Lastly, the graph M100 shows the differences on the y-axis, above all with a big difference between the subtropical sharks and the other species, irrespective of their environment.

6.4.2.4 Validation of the conditions of application

Once we have chosen the best model, we just need to confirm the conditions of application. Here, as is always the case for the linear model, we need to check the independence and normality of the residuals.

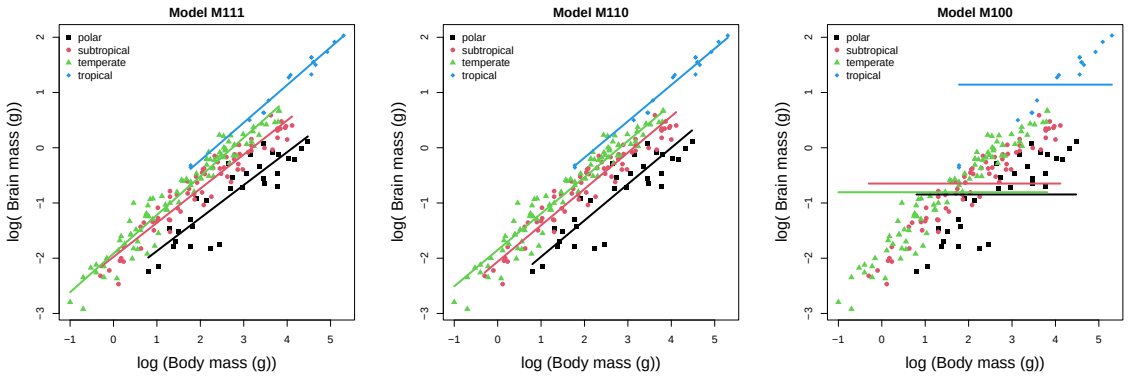


Figure 6.10: **Analysis of covariance: comparison of models.** Each graph represents data (relation between body weight and brain mass) and the values predicted for a model. The dots are the observed values, the straight-line segments are the values predicted in each environment. The colour code corresponds to the different environments.

To do so, we create a graph showing the relation between predicted values and residuals, and carry out a Q-Q plot of the residuals.

6.4.3 Estimation of the model's parameters

In our example, the best model is the complete model. We can then ask, for each type of parameter (constant and allometric coefficients), which coefficients are significantly different from the others. We can use the estimations of coefficients and their interval of confidence to carry out compliance tests (for example, $a_1 = 0$ as against $a_1 \neq 0$) or the comparison tests two by two (for example $a_1 - a_2 = 0$ as against $a_1 - a_2 \neq 0$).
The R software has a number of statistical tests about estimated coefficients. It is important to be able to decipher the function's output `summary`. By default, R lists the coefficients in alphabetical order of the level of factors. Here, there are four environments called "polar", "subtropical", "tropical" and "temperate" in the data file. The polar environment is the first in alphabetical order and will be considered as a reference. The parameters of the model M111 are:

Environment	Coefficients
Polar	a_1, b_1
Subtropical	a_2, b_2
Temperate	a_3, b_3
Tropical	a_4, b_4

R gives the estimated value of coefficients $\widehat{a}_1$ and $\widehat{b}_1$, then the differences between the other coefficients and the reference. A difference between two coefficients is called a contrast. For each contrast, R gives the estimated value, the standard error, and carries out a compliance test in relation to a value of zero. The software displays the value of the test statistic (Student's t-test) and the p-value.

Biological conclusion of the study. When we interpret the results in the table 6.3, we can see that all the constants are lower in the polar environment. For the same body mass, the average brain size will be larger in species living in a tropical or temperate habitat that among those living in a polar habitat. For sharks living in a subtropical environment, brain size will be much bigger.
We can see that we have not highlighted the difference between the allometric coefficients of the polar species and the subtropical or temperate species (the differences $b2 - b1$ and $b3 - b1$ are not significant). However, the species living in a tropical environment have a higher allometric coefficient than the species living in polar habitats. For these species, we can indeed see in the figure 6.10 that the gradient of the regressive straight line is steeper.

Estimation	Coefficients	Estimate	Std.	Error	t value	Pr(>|t|)
a_1	(Intercept)	-2.46539	0.11		-21.53	$< 2e-16$***
b_1	X	0.59653	0.04		15.07	$< 2e-16$***
$a_2 - a_1$	envsubtropical	0.87230	0.25		3.39	0.00085**
$a_3 - a_1$	envtemperate	0.48939	0.13		3.81	0.00018***
$a_4 - a_1$	envtropical	0.54941	0.12		4.52	0.00001***
$b_2 - b_1$	envsubtropical:X	0.08587	0.07		1.26	0.21085
$b_3 - b_1$	envtemperate:X	0.02137	0.05		0.463	0.64363
$b_4 - b_1$	envtropical:X	0.10262	0.04		2.31	0.02167*

Table 6.3: **model 111: tests on the estimated coefficients obtained using the function *lm* in the R software.**

Chapter 7

Principal Components Analysis

7.1 Introduction

In biology, we often have quantitative data associated with individuals. To study this type of data, we generally represent individuals with dots placed on a frame according to the values of each variable (figure 7.1).

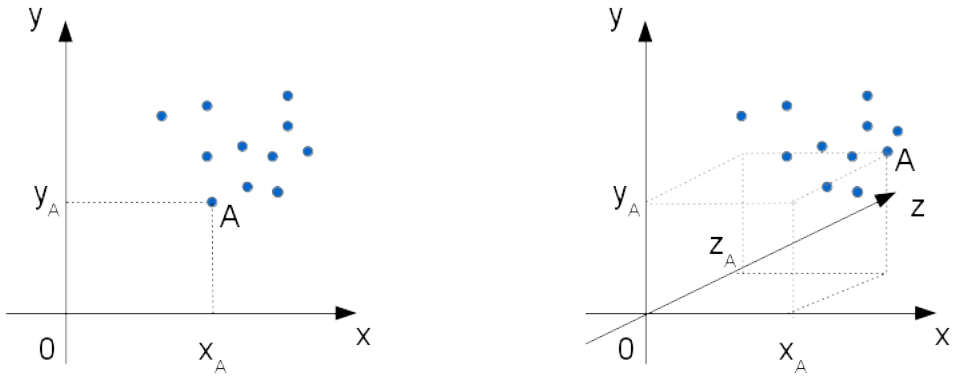


Figure 7.1: **Frames in two and three dimensions**

Although this representation is perfectly adapted to a set of data made up of two or three variables, it can prove more complex when we are looking at a larger number of variables. In this case, we will make use of varied methods of analysis that will take into account the distribution of several different variables. With these methods we can simplify the set of data by identifying the most informative combinations of variables: the principal components.

By carrying out an analysis on the principal components, we will be able to

- Visualise correlations between the variables
- Visualise relations between the individuals
- Identify the variables that can differentiate between individuals.

7.2 Principle

Principal component analysis (PCA) is a method of factor analysis in multivariate statistics.
Based on data made up of individuals (in a line) described by quantitative variables (in columns), the method is used to calculate latent variables, that is, linear combinations of observed variables and thus to identify the ones that best sum up the variance contained in the initial set of data. PCA can also reduce the number of descriptive variables while limiting information loss.

We can begin to grasp PCA by using a geometric approach. Based on data placed in a frame, we will show them in a new system of coordinates, maximising the distances between individuals.
The example in the figure 7.2 illustrates the changed frame allowing us to maximise the dispersion of the object according to the first axis Y_1, then Y_2, etc., of the new frame Y. PCA will not deform the initial object, since the distances between the points are preserved, but our viewpoint of the object is modified.

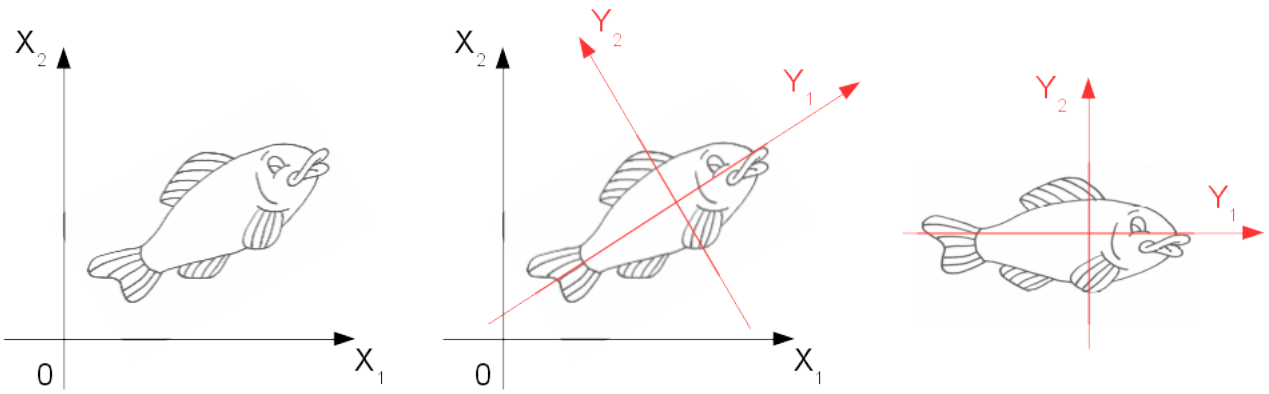


Figure 7.2: **Principle of changing the frame during PCA.**

7.3 Method

7.3.1 Reminder about the distances between two points

PCA is a method based on distances using the Euclidean distance between individuals in a set of data. The Euclidean distance $d(A, B)$ between two points A and B located in a two-dimensional space (figure 7.3) is given by the relation:

$$d(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

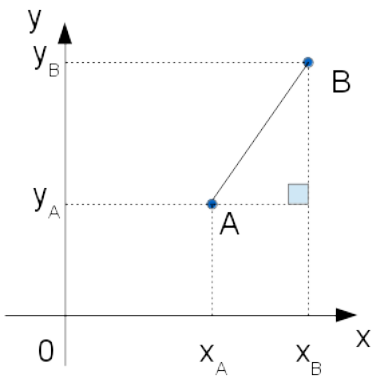


Figure 7.3: **Distance between two points in a two-dimensional frame**

In a space with J dimensions, we can generalise the expression:

$$d(A_i, A_{i'}) = \sqrt{\sum_{j=1}^{J} (x_{i'j} - x_{ij})^2}$$

- A_i and $A_{i'}$ are two points in a frame X
- x_{ij}, the coordinate of point A_i on the axis X_j
- $x_{i'j}$, the coordinate of point $A_{i'}$ on the axis X_j

7.3.2 The initial frame

We will consider a set of data made up of n individuals and J variables. We can represent these data using a scatter plot in J dimensions with each dot representing an individual.

The individuals

We can describe the scatter plot of individuals by its centre of gravity G (*i.e.* the point of which the coordinates are the mean values of the variables) and its inertia I_G (*i.e.* its dispersion).

The coordinates of the **centre of gravity** are:

$$G = [x_{.1}, x_{.2}, x_{.3}, ...x_{.J}]$$

where $x_{.j}$ is the mean of the coordinates of n dots on the axis X_j.

The **inertia** of a scatter plot is defined by the sum of squared distances of each dot in the centre of gravity, weighted by the m_i weight of each dot:

$$I_G = \sum_{i=1}^{n} m_i d^2(G, x_i).$$

Here, we consider that each individual has the same weight and we write $\sum_{i=1}^{n} m_i = 1$, so $m_i = \frac{1}{n}$. We then have:

$$I_G = \sum_{i=1}^{n} \frac{1}{n} d^2(G, x_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} (x_{ij} - x_{.j})^2 = \sum_{j=1}^{J} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_{.j})^2 = \sum_{j=1}^{J} s_j^2$$

We find the variance of individuals according to the axis X_j, written as s_j^2. In the case of standardised variables, the variance of each variable equals 1, so $I_G = J$ (number of variables).

The variables

We can describe the variables and their relations by calculating the variances of each of them and the covariances of each pair of variables.
These data are stored in a variance-covariance matrix (squared matrix with dimensions $J \times J$). The diagonal of this matrix contains the variances (always positive), and the other cells contain the covariances between pairs of variables. The matrix is thus symmetric. To go beyond units of measurement, we often use standardised variables, and in this case the variance-covariance matrix becomes the correlation matrix. This means we can give the same weight to all the variables, and the weight does not depend on the size of the variable.

7.3.3 Calculation of new axes

Based on the representation of data according to the J dimensions, we will carry out changes on the frame. We define G as the origin of the new frame. The axes $Y_1, Y_2, ..., Y_k, ..., Y_K$ will be calculated sequentially. They will all go through G, and since they are independent, they are orthogonal. We should note that $K = J$, since changing the frame does not change the number of dimensions of the space. Constructing new axes means constructing new variables, known as latent or virtual variables, which are linear combinations of the variables observed in the initial set of data, such that:

$$Y_{ik} = a_{k1}x_{i1} + a_{k2}x_{i2} + ... + a_{kJ}x_{iJ}$$

with:

- Y_{ik}, the coordinate of the i^{th} dot on the k^{th} axis Y
- a_{kj}, the coefficient associating the initial variable j with the latest variable k in the new frame.

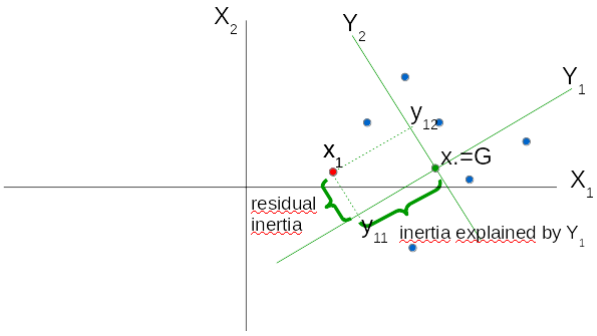


Figure 7.4: X **and** Y **frames and projection of the dot** x_1 **on the new frame** Y**.**

7.3.3.1 The main axis

The main axis is written as $Y1$. When we change frames, the total inertia I_G of the scatter plot stays the same, but we will seek to maximise the inertia of the scatter plot along the new axis Y_1, while minimising the residual inertia around the latter (figure 7.4).

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, x_i) = \frac{1}{n}\sum_{i=1}^{n} d^2(G, y_{i1}) + \frac{1}{n}\sum_{i=1}^{n} d^2(y_{i1}, x_i)$$

$$I_G = \text{total inertia} = \text{inertia explained by } Y_1 + \text{residual inertia}$$

The aim is to maximise the distance $d^2(G, y_{ik})$ bearing in mind that

$$\sum_{k=1}^{K} I_{Y_k} = I_G = \text{cte.}$$

and that Y_k are independent. The solution involves calculating vector space and optimisation under constraint (*cf.* Appendix). The resulting solution is:

- the eigenvalues $\lambda_k = I_{Y_k}$
- the eigenvectors $\overrightarrow{a_1} = (a_{11}, a_{12}, ...a_{1J}), ..., \overrightarrow{a_k} = (a_{k1}, a_{k2}, ...a_{kJ}), ..., \overrightarrow{a_K} = (a_{K1}, a_{K2}, ...a_{KJ})$.

Each virtual variable (or PCA axis) has a **characteristic value** λ_k called eigenvalue and representing its empirical variance ($\lambda_k = I_{Y_k} = \sigma^2_{Y_k}$). The $\overrightarrow{a_k}$ are **characteristic vectors**, called eigenvectors. For the first axis, we thus have the eigenvalue λ_1 and the eigenvector $\overrightarrow{a_1}$. The eigenvalues and eigenvectors are the ones from the variance-covariance matrix of the initial variables.

7.3.3.2 The following axes

The maximum inertia not supported by the first axis Y_1 will be supported by the following axis Y_2, etc. The new variables will pass through G and, since they are independent of each other, Y_2 will be perpendicular to Y_1. The axes will thus be listed by decreasing level of inertia, and the sum total of inertia is equal to the total inertia of the scatter plot.

$$I_G = I_{Y_1} + ... + I_{Y_k} + ... + I_{Y_K}$$

$I_G = \sum_{k=1}^{K} \sigma_k^2$ in the case of a PCA on non-standardised variables
$I_G = K$ in the case of a PCA on standardised variables, since each standardised variable has a variance equal to 1
I_{Y_k} : inertia explained by the axis Y_k
So, the share of inertia explained by the first p components can be written as $\frac{\sum_{k=1}^{p} I_{Y_k}}{I_G}$.

7.3.4 Interpreting the results of a PCA

7.3.4.1 Selection of main components

The previously calculated latent variables have the number K and are listed according to their decreasing eigenvalues. To interpret the results of a PCA, we will look for the latent variables that best explain the dispersion of the scatter plot.
To do so, we make a bar chart of the eigenvalues, also known as a scree plot.

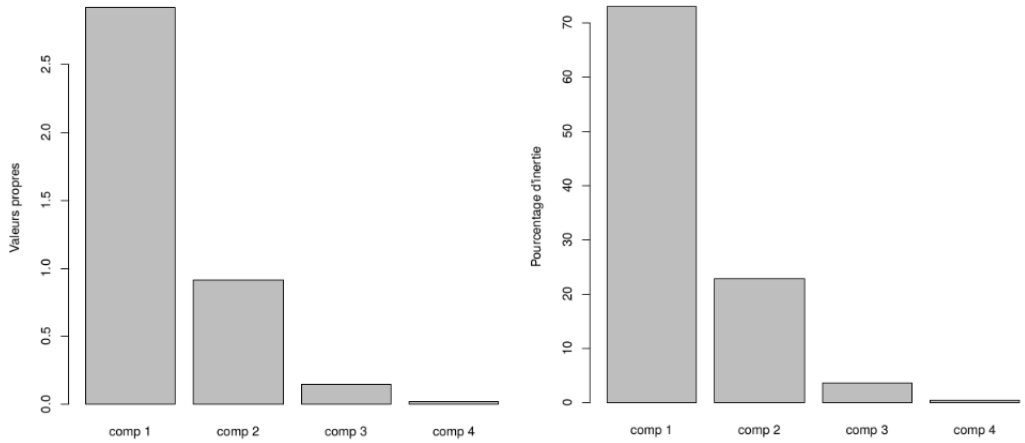


Figure 7.5: **Scree plot of characteristic values (left), shown in terms of percentage of inertia (right)**

The latent variables are on the horizontal axis and the eigenvalues are on the vertical axis (or the percentage of inertia of each component). The bar chart always decreases and the shape of the graph will help us determine which v variables should be taken into account. We can, for example, choose v such that $> 50\%$ of the inertia is explained. Sometimes, we see "an elbow" in the bar chart, that is, the eigenvalues fall sharply. This helps us locate the component coinciding with the beginning of the fall. On the figure 7.5, this criterion would lead to selecting the first component of the PCA. In practice, we will consider at least the first two components so that we can show the variables and individuals on a 2D graph.

7.3.4.2 Representation of variables

Above we saw that the main components were linear combinations of initial variables. We show the relations between the different initial variables and the main components in a correlation circle (figure 7.6).

Interpreting the relations between the initial variables and principal components

The two perpendicular axes in this graph show two principal components of the PCA: Y_1 and Y_2. Each initial variable is represented by a vector with its origin in the frame's origin. The outer limit of these vectors is the point whose coordinates on the main axes are associated with the measurement of the correlation with each of the principal components.
With this 2D representation we can interpret the relations between the initial variables and the two components. In fact, in the area of dimension K, the vectors corresponding to the initial variables are all of length 1 and describe a sphere of a radius equal to 1 (in the case of standardised variables).
The more a variable is represented in the 2D graph, the more the outer limit of the vector will be close to the correlation circle (figure 7.7a)

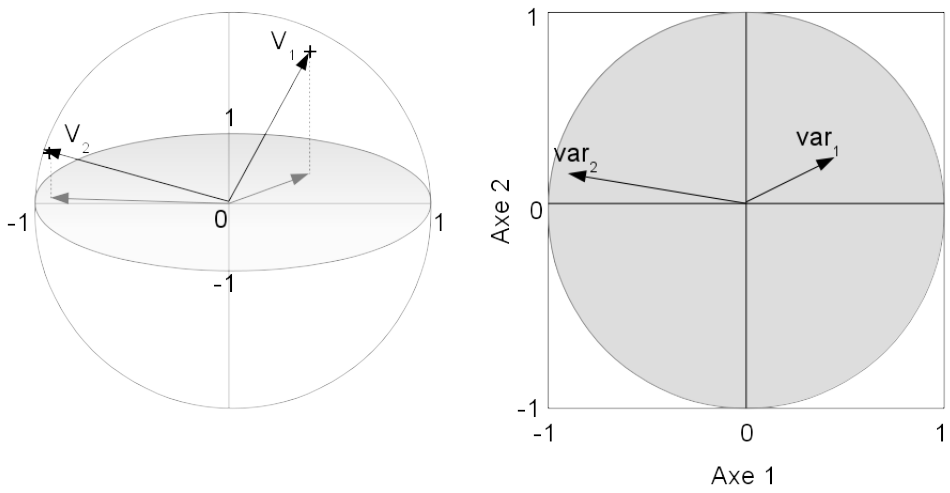


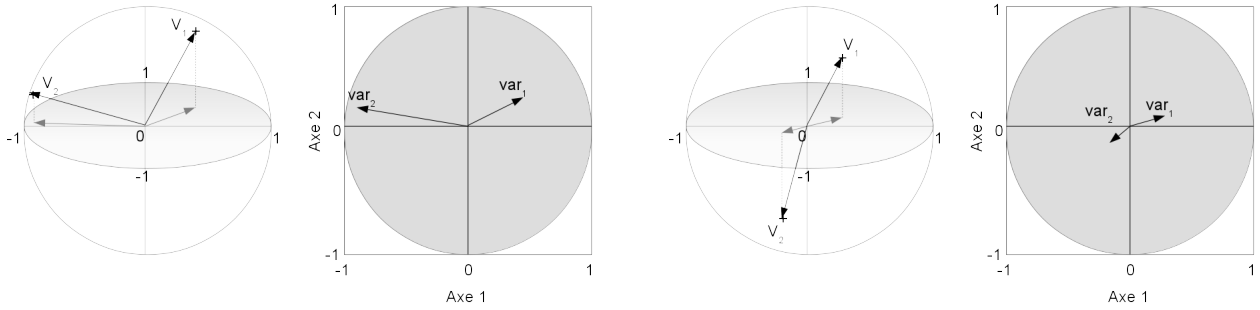
Figure 7.6: **Correlation circle**

Interpreting the relations between the initial variables

In the correlation circle, we can also interpret the relation between two variables. The relations between projected variables can only be interpreted when the two variables are close to the correlation circle (figure 7.7). This is the case for the example in the figure 7.7a but not for the one in the figure 7.7b. The angle formed by two vectors shows a correlation between the variables that they represent according to the relation:

$$\cos(\overrightarrow{X1}, \overrightarrow{X2}) = r(X1, X2)$$

If the two vectors are superimposed the correlation is maximal (= 1). When they form an angle at 180° the variables they represent are perfectly anti-correlated, and when they are orthogonal the variables are not correlated.



(a) (b)

Figure 7.7: **Two examples of correlation circles**

7.3.4.3 Representation of individuals

The principal components correspond to new quantitative variables that will be used to place the individuals from the set of data in a frame formed by the two principal components of the PCA.
The origin of the frame is the point $G(0.0)$, and each individual i is shown by a dot on the graph with its coordinates being the values y_{ik} on the axis Y_k. By making a projection on each of the axes of the PCA, we will be able to identify the modes of individuals that have been isolated according to each of the axes.

Please note

- *Using a representation in the form of a correlation circle, we can only interpret the behaviour of variables with a vector close to the edge of the circle, since these dots are projected best in the factorial area.*
- *The 2D projection can bring out proximity between two dots that are in fact far apart.*
- *In the case where we identify more than three main components, we will need to interpret the graphic representations for each pair of components.*

7.4 Example of interpretation of a PCA

We have data on the morphology of flowers in three species of iris (*Iris setosa*, *I. versicolor* and *I. virginica)* (en cm).

Species	Sepal length	Sepal width	Petal length	Petal width
I. versicolor	5	3.1	1.7	0.3
I. virginica	4	3.7	1.5	0.2
I. versicolor	4.2	2.8	1.8	0.2
I. setosa	4.7	3.1	1.6	0.2
...	...	...	...	...

Table 7.1: **Set of data on the morphology of irises**. Out of 150 individuals belonging to three species of iris, we have a qualitative variable (the species) and measurements for four quantitative variable.

We want to know if the variables can be used to differentiate the three species.
After calculating the principal components, we look at the graph of characteristic values in order to identify the principal components to be taken into account (figure 7.8).

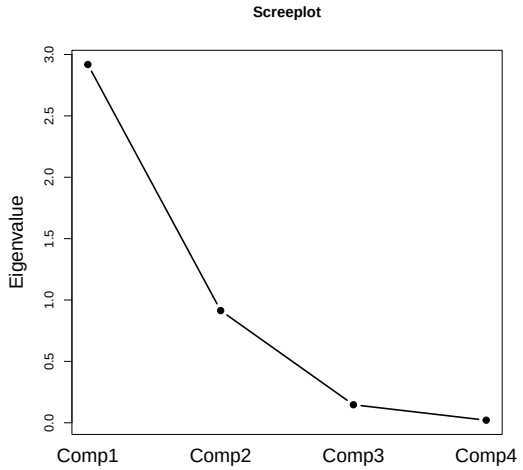


Figure 7.8: **Graph of characteristic values on the Iris set of data.**

The first principal component seems to explain most of the variance in the initial set of data. But we will still project the individuals on the map described by the first two components.
If we colour the dots according to the plant species and project them along the axis PC1 (Y1) and PC2 (Y2), we can see that PC1 is almost perfectly distinguished from the three species, whereas this is not the case for the PC2 axis. The reference axes also show the percentage of variance in the set of data for each component. We note that axis 1 has 73% and axis 23% of the variance in the set of data, or a total of 96% for map 1-2.

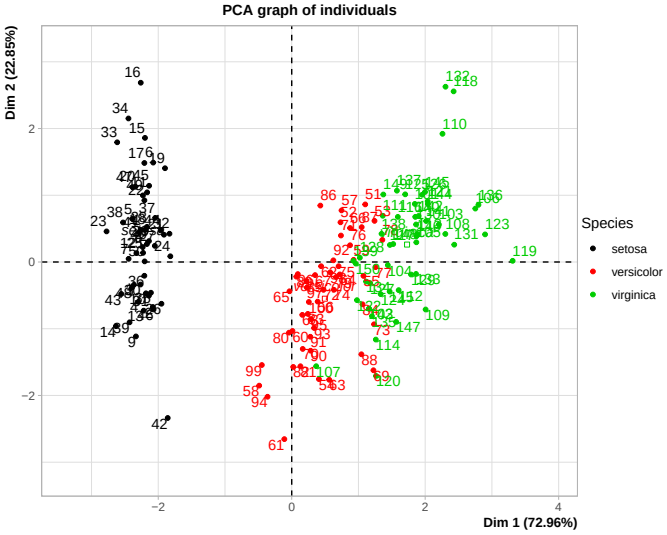


Figure 7.9: **Individuals factor map** ($PC1, PC2$)

We want to find out which initial variables are represented on each of the two components. The correlation circle shows us that the variables are projected near the correlation circle, so we can be confident about the angles between the variables to deduce the correlations from them. Axis 1 represents 3 of the 4 initial variables, while axis 2 mainly represents sepal width.

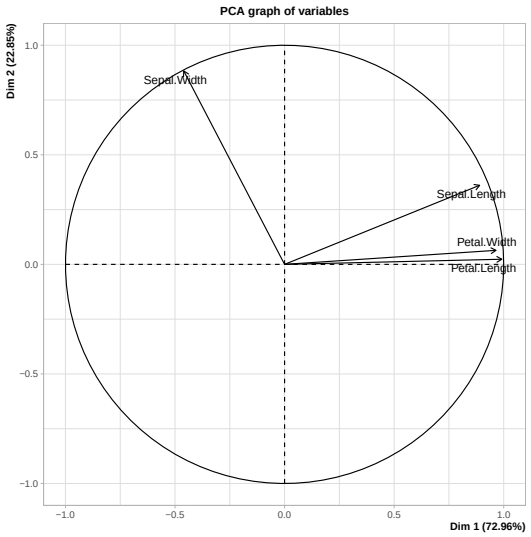


Figure 7.10: **Correlation circle for the Iris data**

Conclusion In this example, the three species mainly differ by the lenth of petals and sepals and the width of petals, with *Virginica* having the highest lengths, and *setosa* the lowest ones. Within each species, individuals vary for sepal width, independently of the three other traits.

7.5 Appendix: calculation of the inertia on the first axis

To calculate the inertia I_{Y_1} shown on the first axis, we will first calculate the distance G-y_{i1}

$$\begin{aligned} d(G, y_{i1}) &= \sqrt{(\overrightarrow{Ga_1}.\overrightarrow{Gx_i})^2} \\ &= \sqrt{(\overrightarrow{Ga_1}.\overrightarrow{Gx_i})(\overrightarrow{Gx_i}.\overrightarrow{Ga_1})} \text{ (since the dot product is symmetrical)} \\ &= \sqrt{a_1^T X_i X_i^T a_1} \text{ (the dot product written in matrix form)} \end{aligned}$$

So that:

- y_{i1}, the orthographic projection of the i^{th} point on the axis Y_1
  We thus deduce I_{Y_1},

$$\begin{aligned} I_{Y_1} &= \frac{1}{n}\sum_{i=1}^{n} d^2(G, y_{i1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} a_1^T X_i X_i^T a_1 \\ &= a_1^T \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right] a_1 \\ &= a_1^T \Sigma a_1 \end{aligned}$$

- Σ, the variance / covariance matrix

We will thus try to maximise $a_1^T \Sigma a_1$. We also know that $\|\overrightarrow{Ga_1}\|^2 = a_1^T a_1 = 1$ since $\overrightarrow{Ga_1}$ is the unitary vector of the axis Y_1.
Here we have a problem of optimisation under constraint, which could be processed using the Lagrange multiplier method by means of which we obtain the equation:

$$a_1^T \Sigma a_1 - \lambda_1 a_1^T a_1 = 0$$

By using $a_1^T a_1 = 1$, we get

$$a_1^T \Sigma a_1 = \lambda_{Y_1}$$

Each latent variable (or axis of the PCA) has an **eigenvalue** λ_k representing its empirical variance ($\lambda_k = I_{Y_k} = \sigma^2_{Y_k}$). The $\overrightarrow{a_k}$ are the **eigenvectors**.