Intermediate Statistics

Christian M. Hafner
Université catholique de Louvain

LINGE1214 – *Statistique Approfondie*

2022/23

1 Probability

We start with recalling some notions of probability theory that will be essential and useful for statistical analysis. A more detailed treatment of the material is provided e.g. in Chapters 1 to 7 of Wackerly et al. (2008), and part I of Linton (2017). Also, most textbooks on econometrics have excellent summaries of the basics in probability and statistics, for example Baltagi (2011) or Stock and Watson (2012).

A random experiment is a real or artificial experiment whose result can not be predicted with certainty, and which can be repeated under the same conditions arbitrarily often. The possible results of this experiment are collected in the sample space Ω.

Example 1.1: Throwing a dice is a random experiment with sample space

$$\Omega = \{1, 2, \ldots, 6\}.$$

A random event A is a subset of the sample space, for example the event of an even outcome, $A = \{2, 4, 6\}$ when throwing the dice. We want to associate probabilities to any such random event. Probabilities can be defined as the limit of relative frequencies when the number of experiments grows to infinity. In the given example, we throw the dice many times and observe the proportion, or relative frequency, of observing an even number. If it is a correct dice, this proportion should converge to the probability 1/2 as the number of experiments increases. This is an implicit and intuitive definition of probability. More explicit, set-theoretic definitions are available in the literature.

A probability measure is a measure P that has the following properties:

1. $P(A) \geq 0$ for all $A \in \mathcal{A}$

2. $P(\Omega) = 1$

3. If $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

These are the famous Kolmogorow axioms of probability. One important property of probabilities that follows from these axioms is that $P(A^c) = 1 - P(A)$, where A^c is the complement of A. Another is the following addition property:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Let A and B be two random events of the sample space Ω. Then the conditional probability $P(A|B)$ is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

In the numerator is the joint probability that both events A and B occur, and in the numerator the probability of B. This conditional probability is the probability that A occurs, given that we know that B has occured. A simple example will illustrate this concept.

Example 1.2: What is the probability that the dice will show a 2, given that we know that the outcome is even? Here, the conditioning event is "number is even", which occurs with probability 1/2, hence $P(B) = 1/2$. The event A is "dice shows a 2". The probability that A and B occur simultaneously is the same as that of A, $P(A \cap B) = P(A) = 1/6$. The conditional probability is therefore

$$P(A|B) := \frac{1/6}{1/2} = \frac{1}{3}.$$

In other words, knowing that the outcome is even, the probability of obtaining a 2 is 1/3. In a sense, we discard the odd numbers of the potential outcomes from the sample space, and then look at the probability of obtaining A given this new space.

Independent events are defined by the property

$$P(A|B) = P(A).$$

In other words, for independent events, the probability of A conditional on B is equivalent to the unconditional (or marginal) probability of A. In the dice example, A and B are clearly not independent, since $P(A|B) = \frac{1}{3} \neq P(A) = 1/6$. Independent events would be, for example, $A = \{1,2\}$ and $B = \{even\}$, because $P(A \cap B) = P(2) = 1/6$, and hence $P(A|B) = P(A) = 1/3$.

A consequence of the definition of independent events is the following multiplication rule:

$$P(A \cap B) = P(A)P(B)$$

which holds only if A and B are independent, not in general.

A random variable (r.v.) is a mapping from the sample space to the real line, i.e.

$$X : \Omega \rightarrow \mathbb{R}.$$

It assigns to each member of the sample space, which is not necessarily numeric, a number on the real line. When tossing a coin, for example, heads and tails could be mapped to the values 0 and 1, respectively.

If Ω is at most countable, then we speak of a discrete random variable, otherwise of a continuous r.v. An example of a discrete r.v. in the dice example is the r.v. X that is defined to take the value 1 if the result is even, and 0 if it is odd. The mapping would therefore be:

$$\begin{aligned} A = \{2,4,6\} &\quad \rightarrow \quad X = 1 \\ A^c = \{1,3,5\} &\quad \rightarrow \quad X = 0 \end{aligned}$$

An example of a continuous r.v. is the height of a randomly sampled person of the population, if measured at arbitrary precision.

A probability function of a discrete r.v. X assigns to each possible outcome of X its probability. In the previous example, the probability function would be given by

$$P(X) = \begin{cases} 1/2, & \text{for } X = 1 \\ 1/2, & \text{for } X = 0 \end{cases}$$

This is a special case of a Bernoulli r.v. which is defined as a binary r.v. with probability function

$$P(X) = \begin{cases} p, & \text{for } X = 1 \\ 1 - p, & \text{for } X = 0 \end{cases}$$

where $0 \leq p \leq 1$.

The distribution function, or cumulative distribution function (cdf), is defined as

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

Note that X is the r.v., while x is a fixed non-random value. For the example of a Bernoulli r.v., the cdf is given by

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ p, & \text{for } 0 \leq x < 1 \\ 1, & \text{for } x \geq 1 \end{cases}$$

In general, the cdf has the following properties:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

2. F is non-decreasing in x.

3. F is everywhere right-continuous, i.e. for all $x_0 \in \mathbb{R}, \lim_{x \downarrow x_0} F(x) = F(x_0)$.

The quantile function $Q : [0,1] \to \mathbb{R}$ is in a sense the inverse of the cdf. It solves the equation

$$P(X \leq Q(\alpha)) = F(Q(\alpha)) = \alpha$$

where $\alpha \in [0,1]$. The median is just the quantile of X with $\alpha = 1/2$, i.e. half of the probability mass of X is smaller, the other half larger than the median. Lower and upper quartiles can be obtained with $\alpha = 1/4$ and $\alpha = 3/4$. The interquartile range, $Q(3/4) - Q(1/4)$, is sometimes taken as a measure for the dispersion of X.

The density $f(x)$ of a continuous r.v. X, if it exists, is defined as the derivative of the cdf, i.e.

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

The density function has the following properties:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

Probabilities for a continuous r.v. X can be calculated using the density function, for example

$$P(a < X < b) = \int_a^b f(x)dx,$$

for some constants a and b with $a < b$. This implies that the probability that X takes a specific value is equal to zero, e.g. $P(X = a) = 0$.

1.1 *Moments of a random variable*

The expectation of random variable Y is defined as

$$\mathbb{E}(Y) = \begin{cases} \int_{-\infty}^{\infty} yf(y)dy & Y \text{ is continuous} \\ \sum_k y_k P(y_k) & Y \text{ is discrete} \end{cases},$$

and gives information about the average value to be expected in a large number of samples. It may itself not be a member of the sample space. For example, for a Bernoulli r.v. Y with $\Omega = \{0,1\}$ and $P(Y = 1) = p$, the expectation is given by $\mathbb{E}(Y) = 1 * p + 0 * (1 - p) = p \notin \Omega$. The expectation is a linear operator in the sense that, for two r.v. X and Y and constants a, b,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Unlike the median, the expectation is not a robust measure for location, meaning that it is sensitive to outliers and the probability mass in the tails of the distribution.

The most common measure for the dispersion of a r.v. Y is the variance, defined as

$$\mathbb{V}\text{ar}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2].$$

It is the expectation of the squared deviance of Y from its mean. Obviously, it can not be negative, and zero only for degenerated distributions with all probability mass concentrated in a single point. The higher the variance, the more dispersed the r.v., and the lower the variance, the more concentrated it is about its mean. Being an expectation itself, the variance is not a robust measure for dispersion. If a robust measure is preferred, the above mentioned inter-quartile range, or the median absolute deviation, might be alternatives.

The skewness is defined as the third central and standardized moment:

$$Sk(Y) = \frac{\mathbb{E}[(Y - \mathbb{E}(Y))^3]}{\mathbb{V}\mathrm{ar}(Y)^{3/2}}$$

Note that, letting Z be the centralized and standardized r.v. Y, i.e.

$$Z := \frac{Y - \mathbb{E}(Y)}{\mathbb{V}\mathrm{ar}(Y)^{1/2}}$$

such that $\mathbb{E}(Z) = 0$ and $\mathbb{V}\mathrm{ar}(Z) = 1$ by construction, we can write the skewness of Y equivalently as $Sk(Y) = \mathbb{E}[Z^3]$, and in this sense it is the third central and standardized moment of Y. The skewness gives information about the symmetry of the distribution. If the distribution of Y is symmetric about its mean, $Sk(Y) = 0$. Distributions with a negative skewness are called left-skewed, and those with a positive skewness are called right-skewed.

The kurtosis is defined as the fourth central and standardized moment:

$$K(Y) = \frac{\mathbb{E}[(Y - \mathbb{E}(Y))^4]}{\mathbb{V}\mathrm{ar}(Y)^2}$$

Note that $K \geq 1$, which can be shown using Jensen's inequality and is left as an exercise.

Theorem 1.1 (Jensen's inequality): Let X be a random variable and g a convex function. Then,

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

In the kurtosis case, take Z^2 as X, and g as the square function, and simply apply Jensen's inequality to see that the kurtosis cannot be smaller than one.

Again, using the standardized r.v. Z, we can write the kurtosis equivalently as $K(Y) = \mathbb{E}[Z^4]$. The kurtosis gives information about the curvature of the distribution around its mean, and also about the thickness of the tails. Suppose that two distributions have the same mean, variance, and skewness, then the distribution with the higher kurtosis will typically have a higher probability mass in the tails, and also a higher curvature around the mean.

The defined moments do not necessarily exist. For the case of a Cauchy distribution, for example, even the mean does not exist. Assuming that the kurtosis exists implies that the lower moments (mean, variance, and skewness) exist as well. We will often make this assumption in later chapters.

1.2 Measures of association

For two r.v. X and Y, the covariance is defined as

$$\mathbb{C}\mathrm{ov}(X,Y) := \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]$$

Obviously, $\mathbb{C}\mathrm{ov}(X,X) = \mathbb{V}\mathrm{ar}(X)$, i.e. the covariance reduces to the variance in the case of $X \equiv Y$. It is also obvious that the covariance is symmetric, i.e. $\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{C}\mathrm{ov}(Y,X)$. The covariance is a measure of linear association between two random variables. If the covariance is positive (negative), then there tends to be a positive (negative) linear relationship between the two r.v. If it is zero, then there is no such relationship. The strength of the relationship is indicated by the size of the covariance. However, the covariance is not a standardized measure, in the sense that the covariance depends on the scale of the r.v. If either X or Y are multiplied by ten, for example, then the covariance is also multiplied by ten. If both X and Y are multiplied by ten, then the covariance is multiplied by 100.

To obtain a scale-free measure of linear association, we define the correlation between X and Y as

$$\mathrm{cor}(X,Y) := \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\mathbb{V}\mathrm{ar}(X)^{1/2}\mathbb{V}\mathrm{ar}(Y)^{1/2}}$$

By the Cauchy-Schwarz inequality, we can show that $-1 \leq \mathrm{cor}(X,Y) \leq 1$, which is left as an exercise.

Theorem 1.2 (Cauchy-Schwarz inequality): For random variables X, Y,

$$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

with equality if and only if $Y = aX$ for some $a \in \mathbb{R}$.

The closer the correlation is to plus or minus 1, the more the joint distribution will be concentrated on a straight line with positive or negative slope, depending on the sign of the correlation.

If the correlation is zero, the r.v. are usually called linearly unrelated, or linearly independent. However, this does not imply independence in general, which is a much stronger concept. Independence implies zero correlation, but not vice versa. Only in the case of a bivariate Gaussian distribution, the two concepts are equivalent. Because this is important to remember we will formulate it as a theorem.

Theorem 1.3: Independence between two r.v. X and Y implies that $\mathrm{cor}(X,Y) = 0$. However, $\mathrm{cor}(X,Y) = 0$ does not imply independence in general. Only in the case of a bivariate Gaussian distribution, we have: independence $\Leftrightarrow \mathrm{cor}(X,Y) = 0$.

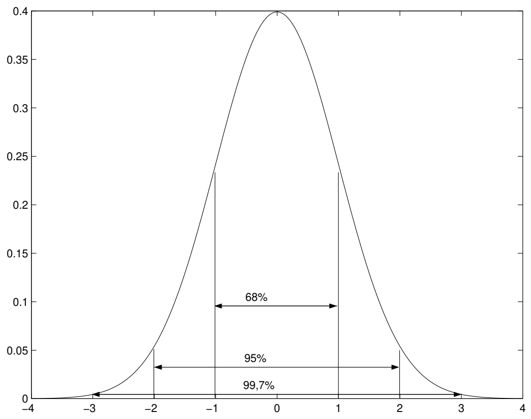


Figure 1.1: Standard Gaussian density.

1.3 *The Gaussian distribution*

The Gaussian distribution plays a central role in statistics, mainly thanks to the central limit theorem which we will recall shortly. It is also often called the normal distribution, but we prefer to be careful with this term, knowing that many real-world phenomena actually do not follow a Gaussian distribution.

The density of a Gaussian r.v., $Y \sim N(\mu, \sigma^2)$, is given by

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right), \text{ où } \sigma > 0, \quad \mu \in \mathbb{R}, \quad y \in \mathbb{R}.$$

It can be shown that $\mathbb{E}(Y) = \mu$ et $\mathbb{V}\text{ar}(Y) = \sigma^2$. Furthermore, the density is symmetric about its mean, and hence, $Sk(Y) = 0$, and the kurtosis is given by $K(Y) = 3$. Because of its importance in statistics, the kurtosis of the Gaussian distribution is often taken as a reference value, and the *excess kurtosis* of some r.v. X is defined as $K(X) - 3$. So a Gaussian distribution has an excess kurtosis of 0.

If $Y \sim N(\mu, \sigma^2)$, then the centered and standardized r.v. $Z = \frac{Y-\mu}{\sigma}$ follows the so-called standard normal, or standard Gaussian distribution, i.e., $Z \sim N(0,1)$. Figure 1.1 shows the standard Gaussian density, which has the well-known bell shape. As indicated in the graph, about 68% of the probability mass are in the interval $[\pm 1]$, about 95% are in $[\pm 2]$, and about 99.7% are in $[\pm 3]$. In other words, when sampling from a standard Gaussian distribution, it is quite rare to obtain values larger than 3 in absolute value.

1.4 *Calculation of probabilities: Example*

To understand the difference between probability theory and statistics, let us look at an example. Suppose we toss a coin ten times, and look at how often the coin shows "head". The random variable defined by the number of heads has a binomial distribution with parameters $p = 1/2$ for the success probability at each throw, and $n = 10$ for the number of independent experiments (tossing the coin). The probability of obtaining 7 times head can be then be calculated as

$$P(Y = 7) = \binom{10}{7} \left(\frac{1}{2}\right)^7 \left(1 - \frac{1}{2}\right)^3 = 0.1172.$$

This calculation is possible because two conditions are fulfilled: (i) the probability law is known (here the binomial distribution), and (ii) all parameters of the law are known numerically (here $p = 1/2$ and $n = 10$).

If we suppose, on the contrary, that some parameters of the distribution, such as p in our example, are unknown, then the probability cannot be calculated because it depends on this unknown parameter. The question to be asked is, then, how can we determine a plausible value for p, based on observed data, such that the calculation becomes possible again? This is what is called the problem of **estimation**.

Without information it is impossible to estimate p. Statistics proposes methods for data collection and treatment. In our example, this corresponds to tossing the coin several times, and noting the numbers of appearing heads.

Suppose that we have tossed the coin 20 times, and that we have obtained head 7 times. A natural estimator of p is the relative frequency, or proportion, of obtained heads, i.e.

$$\widehat{p} = \frac{7}{20}.$$

The symbol $\widehat{}$ usually indicates that this is an estimator. Note that the estimator $\hat{p}$ will quite likely be different from the true but unknown parameter p. Moreover, if we redo the experiment in exactly the same way, the resulting new estimator will quite likely be different from the first estimator. If in the first experiment we got 7 heads out of 20, in the second we might get 13 out 20, for example. In other words, the estimator $\hat{p}$ is not something that is fixed as the true parameter, but it is actually the result of a random experiment, and therefore itself a random variable. This is important to emphasize: Estimators are random variables, and as such they possess all characteristics of usual random variables, such as a distribution, moments, etc. It is however important to recognize that the properties, such as the distribution, depend on the characteristics of the random experiment. In particular, the number of coin tosses, or in general the number of observations, n, plays a crucial role. We will see that a good estimator has a distribution whose dispersion shrinks as the sample size increases, and that its mean approaches the true parameter. The precision of an estimator is supposed to become better as the sample size n grows.

In our example, suppose that we are allowed to toss the coin as many times as desired. Let us denote the number of coin tosses by n. Call the event

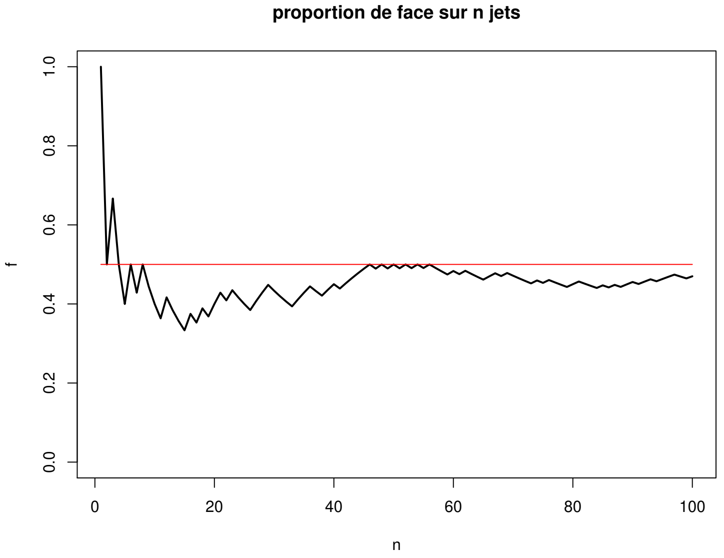


Figure 1.2: Possible trajectory of relative frequencies of heads for the coin tossing example.

of obtaining heads A. To estimate p we will divide n_A, the number of times heads appeared, by the total number n. Call this proportion $f_n(A)$:

$$f_n(A) = \frac{n_A}{n}$$

Now $f_n(A)$ can be regarded as a sequence indexed by n, where n increases from 1 to some large number. Consider the following possible sequence. If in the first coin toss we obtain heads, then $f_1(A) = \frac{1}{1} = 1$. If then in the second toss the result is tails, then $f_2(A) = \frac{1}{2}$. Suppose that the following results are sucessively heads, tails, heads, heads, tails, etc. We deduce that $f_1(A) = \frac{1}{1} = 1$, $f_2(A) = \frac{1}{2}$, $f_3(A) = \frac{2}{3}, f_4(A) = \frac{2}{4} = \frac{1}{2}$, $f_5(A) = \frac{2}{5}$; $f_6(A) = \frac{3}{6} = \frac{1}{2}$. It is interesting, then, to look at the evolution of the estimator of p as a function of n. This is shown in Figure 1.2.

If we now restarted the experiment from the beginning, quite certainly we would get another trajectory of relative frequencies as a function of n, see Figure 1.3 for an example of another possible trajectory.

In both cases, we observe essentially two phenomena: The first is the apparent approximation of $f_n(A)$ towards the horizontal line corresponding to $P[A] = \frac{1}{2}$ as n increases. The second is the higher variance of $f_n(A)$ for smaller n, and the variance decreases for larger values of n.

How can we formalize the concept of convergence of $f_n(A)$ towards $1/2$ in our case? We have to be careful here because $f_n(A)$ is not a deterministic sequence such as $1/n$, whose limit is simply defined by $\lim_{n\to\infty} 1/n = 0$.

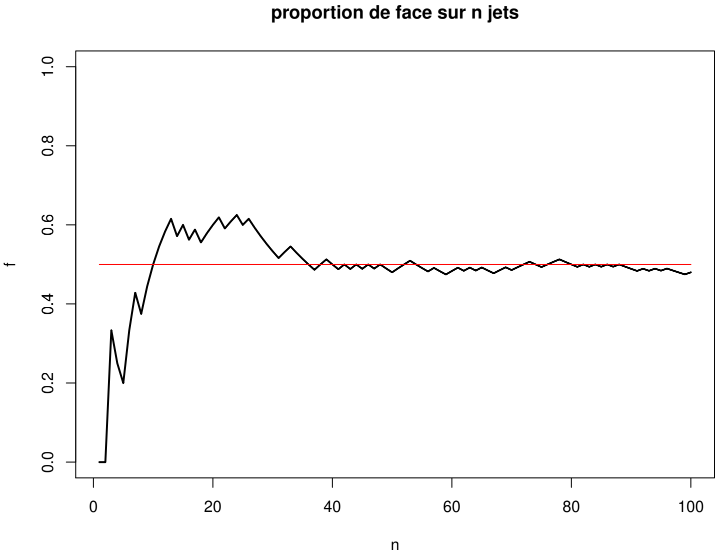


Figure 1.3: Another possible trajectory of relative frequencies of heads for the coin tossing example.

Since $f_n(A)$ is a random sequence, the lim operator is not defined in an unambiguous way and we need to specify what we mean by convergence to a limit. In fact several notions of convergence of random sequences exist in the literature, and we will discuss two of them in this lecture. The following section will introduce the notions of convergence in probability, and the convergence in distribution.

1.5 *Convergence of random sequences*

For a non-random sequence Z_n, $n = 1, 2, \ldots$, the limit as $n \to \infty$, if it exists, is well defined. For example, if $Z_n = 1/n$, then $\lim_{n\to\infty} Z_n = 0$. If Z_n is a sequence of random variables, however, this limit is not defined in an unambiguous way and we first need to specify what we mean by convergence. We distinguish two types of convergence: (i) Convergence in probability, denoted by the symbol $\to_p$, and (ii) convergence in distribution, denoted by $\to_d$.

1.5.1 *Convergence in probability*

The formal definition of convergence in probability is the following:

Definition 1.1: The sequence Z_n tends to a fixed value θ in probability, if

$$\lim_{n\to\infty} P(|Z_n - \theta| < \varepsilon) = 1, \quad \forall \varepsilon > 0$$

We will denote this as

$$Z_n \rightarrow_p \theta, \quad \text{or} \quad \text{plim}(Z_n) = \theta$$

Note first that, by the axioms of probability theory, $P(|Z_n - \theta| < \varepsilon) = 1 - P(|Z_n - \theta| \geq \varepsilon)$, so that an equivalent definition would be based on

$$\lim_{n\rightarrow\infty} P(|Z_n - \theta| \geq \varepsilon) = 0, \quad \forall\varepsilon > 0$$

The idea of this definition is the following. Convergence means that, intuitively, the sequence of "errors", $Z_n - \theta$, has to go to zero as n increases. For a given n and some given $\epsilon > 0$, there is some probability that the absolute value of the error, $|Z_n - \theta|$ is larger than ϵ. We want this probability to be small as n increases, and in fact to tend to zero asymptotically. If this holds for any ϵ, no matter how close to zero, then there is convergence in probability. Intuitively, the distribution of Z_n concentrates around θ more and more as n increases, so that the probability of any positive difference disappears asymptotically.

Now, the definition of convergence of probability as such may be difficult to check in practice. Fortunately, there is a simple sufficient condition that is much easier to check. It can be shown by Chebyshev's inequality below that convergence in probability is implied by the following sufficient condition:

$$\lim_{n\rightarrow\infty} \mathbb{E}[Z_n] = \theta \quad \text{and} \quad \lim_{n\rightarrow\infty} \mathbb{V}\text{ar}(Z_n) = 0.$$

So if the expectation of Z_n, which is non-random, converges in the usual sense to θ, and additionally, the variance of Z_n (also non-random) converges to zero, then this implies convergence in probability. In many of our examples, $\mathbb{E}[Z_n] = \theta$ for any finite n, so that of course this also holds asymptotically. It should be emphasised that this is only a sufficient condition. So there may be cases where a sequence is convergent in probability although this sufficient condition does not hold.

Theorem 1.4 (Chebychev's inequality): For $\eta > 0$,

$$P(|X - \mathbb{E}(X)| \geq \eta) \leq \frac{\mathbb{V}\text{ar}(X)}{\eta^2}.$$

1.5.2 *The law of large numbers*

In simple terms, the law of large numbers says that the average of n independent and identically distributed (i.i.d.) random variables converges in probability to the expectation. This is the simplest and weakest version, and many extensions have been made to allow for non-i.i.d. random variables, but which will not be relevant here.

Theorem 1.5: Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with $\mathbb{E}[Y_1] = \mu$ and $\mathbb{V}\text{ar}(Y_1) = \sigma^2 < \infty$. Then,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \rightarrow_p \mu$$

This is the law of large numbers. Averages converge to expectations. It is very easy to prove this theorem, because we can use the sufficient condition for the convergence in probability. Indeed, we can easily show (as an exercise) that $\mathbb{E}[\overline{Y}] = \mu$ and $\mathbb{V}\mathrm{ar}(\overline{Y}) = \frac{\sigma^2}{n}$, and hence, $\lim_{n\to\infty} \mathbb{V}\mathrm{ar}(\overline{Y}) = 0$, which is the sufficient condition. This implies convergence in probability of the average, $\overline{Y}$, to the expectation, μ.

Example 1.3: Consider the special case of a Bernoulli experiment with success event A. For example, tossing a coin is such an experiment, and A could be *heads*. Let p denote the probability of obtaining A at one such experiments, i.e. $p = P[A]$. In the coin example, $p = 1/2$.

We denote the relative frequency of observing A in n experiments as $f_n(A) := \frac{1}{n}\sum_{i=1}^n I_i$, where I_i are i.i.d. Bernoulli random variables defined as

$$I_i = \begin{cases} 1 & \text{if A is observed in the } i\text{th experiment} \\ \\ 0 & \text{otherwise} \end{cases}$$

By the definition of I_i, the sum $\sum_{i=1}^n I_i$ just gives the number of A events among the n experiments. Therefore, $f_n(A) = \frac{1}{n}\sum_{i=1}^n I_i$ is the proportion, or relative frequency, of A events. As we see, it can be written as the average of the underlying Bernoulli r.v. This is a special case of our general r.v. Y above, and therefore the law of large numbers applies:

$$f_n(A) = \frac{1}{n}\sum_{i=1}^n I_i \to_p \mathbb{E}[I_1] = p.$$

The proportion, which can be written as an average, converges to the expectation. The expectation is equal to p because, for a Bernoulli r.v., $\mathbb{E}[I_1] = 1 \cdot p + 0 \cdot (1 - p) = p$. To summarize this example, the law of large numbers can be used to show that relative frequencies converge to probabilities as the number of experiments increases.

1.5.3 *Convergence in distribution*

Our second type of convergence is in distribution. It is formally defined as the convergence of the cumulative distribution function (cdf) towards the cdf of the limiting r.v.:

Definition 1.2: Let V be a r.v. with cdf F. The random sequence Z_n converges in distribution to V if

$$\lim_{n\to\infty} P(Z_n < u) = F(u), \quad \forall u \in \mathbb{R}$$

We will write this as

$$Z_n \to_d V.$$

If, for example, V has a Gaussian distribution, we sometimes write more compactly,

$$Z_n \to_d N(\mu, \sigma^2).$$

1.5.4 The Central Limit Theorem (CLT)

The importance of the central limit theorem in statistics can hardly be overstated. It plays indeed a central role for statistical inference, therefore its name. Similar to the law of large numbers, it is an asymptotic result for the convergence of random sequences. Unlike the law of large numbers, however, it says something about the asymptotic distribution of this sequence, after having appropriately standardized the sequence. The term "asymptotic distribution" means the distribution of the random sequence in the limiting case as n goes to infinity.

Why is it necessary to standardize the random sequence? Well, if we take our simple example of sample average, $\overline{Y}$, this depends on the sample size n, and therefore is a sequence of r.v. We have calculated its mean and variance, and noted in particular that the variance goes to zero as $n \to \infty$. This means that the asymptotic distribution of $\overline{Y}$ degenerates into a single point. This is not useful. A better idea is to ask, what happens if we scale $\overline{Y}$ with something that depends on n such that the variance of the standardized average remains stable as n increases?

It turns out that the appropriate scaling factor, up to constants, is $\sqrt{n}$: The variance of $\overline{Y}$ is σ^2/n, as you were asked to show in the exercises. So if we multiply $\overline{Y}$ by $\sqrt{n}$, the variance will remain constant at σ^2. It is important to recognize that this scaling is important: If we scaled by n^α with $\alpha > 1/2$, then the variance will increase as n goes to infinity. If, on the other hand, $\alpha < 1/2$, then the variance goes to zero. In both cases, the asymptotic distribution is degenerated and not well defined. It is only with $\alpha = 1/2$ that a stable variance is ensured.

Apart from scaling, we also need a centering of $\overline{Y}$. This is because $\sqrt{n}\overline{Y}$ has a stable variance, but its expectation is $\sqrt{n}\mu$, which diverges to infinity. So in fact, to do both scaling and centering, the appropriate transformation is $\sqrt{n}(\overline{Y} - \mu)$ because then the expectation is zero and the variance remains constant, σ^2.

Theorem 1.6: Let $Y_1, Y_2, \ldots, Y_n$ be iid random variables with mean μ and variance σ^2. Then,

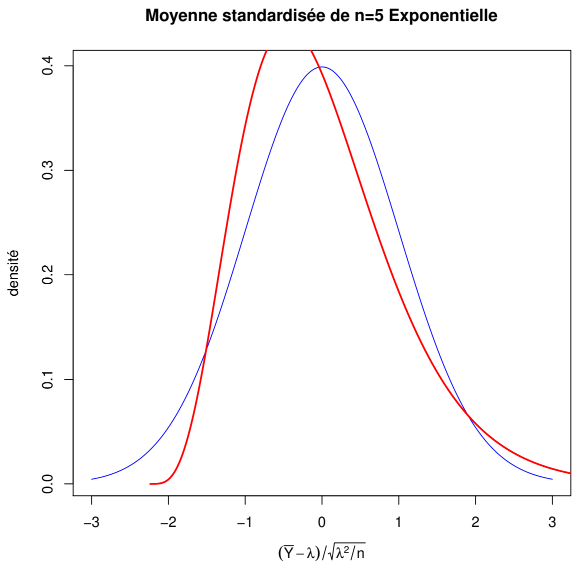
$$\sqrt{n}(\overline{Y} - \mu) \to_d N(0, \sigma^2)$$

as $n \to \infty$. Equivalently,

$$\sqrt{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right) \to_d N(0,1)$$

It is very important to recognize that this result does not depend on the distribution of the underlying Y_i r.v.!

Example 1.4: Let us consider an example of an exponential distribution, i.e. $Y_1, Y_2, \ldots, Y_n \sim$ i.i.d. exponential with density $f(y) = (1/\lambda)exp(-y/\lambda), y \geq 0$ and parameter $\lambda > 0$. As we know, $\mathbb{E}(Y) = \lambda$ and $\mathbb{V}\mathrm{ar}(Y) = \lambda^2$.



Moyenne standardisée de n=5 Exponentielle

Figure 1.4: Distribution of centered and standardized $\overline{Y}$ with $Y \sim Exp$, $\lambda = 1$, and $n = 5$ (in red), standard normal density (in blue).

The following graphs show the distribution of

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}-\lambda}{\lambda}\right)$$

for $Y_i \sim$ exponential with $\lambda = 1$ and $n = 5$ (Figure 1.4) and $n = 20$ (Figure 1.5). It is compared with that of a standard normal distribution. While the approximation for $n = 5$ is not yet very close, the distribution still being slightly skewed to the right, the approximation is already quite good for $n = 20$.

This example shows that in simple cases like this one, the CLT can be applied already in moderately large samples such as $n = 20$ or $n = 30$. If however the underlying r.v. have a more complicated structure, for example not identically distributed, then it might take a much higher sample size to justify an approximation by the CLT. So the applicability of this theorem depends very much on the given situation and no general recommendation can be made.

As we have emphasized, the CLT is based on an appropriate centering and rescaling of the sample average $\overline{Y}$, such as $\sqrt{n}(\overline{Y}-\mu)$. Once we know the asymptotic distribution of this object, we can actually use it as an approximation in large but finite samples. Thus we can say that

$$\sqrt{n}(\overline{Y}-\mu) \sim N(0,\sigma^2)$$

holds not exactly, but approximately, for large but finite n. Consequently, we

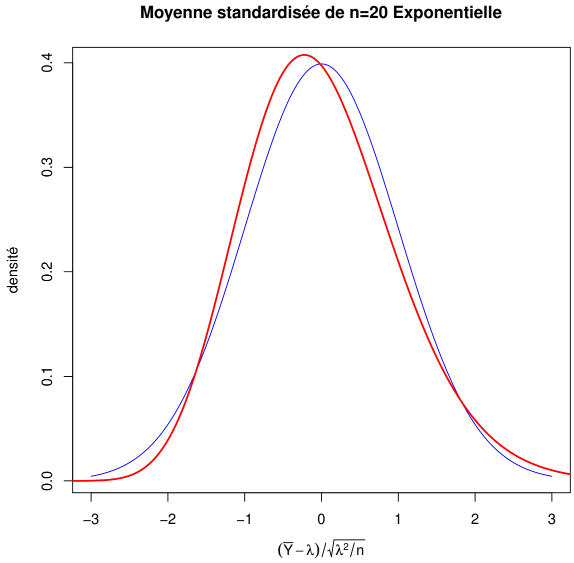


Figure 1.5: Distribution of centered and standardized $\overline{Y}$ with $Y \sim Exp$, $\lambda = 1$, and $n = 20$ (in red), standard normal density (in blue).

can also reformulate this result to obtain an approximate distribution for $\overline{Y}$ itself, not just its centered and standardized version:

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

where again, in this context, "$\sim$" means approximately distributed. We should stress that this is only valid for finite n. It does not make sense to consider the asymptotic case $n \to \infty$ for the uncentered and unstandardized $\overline{Y}$, because we would obtain $\overline{Y} \to_d N(\mu, 0)$, which is not well defined.

1.5.5 *Summary of the convergence concepts*

We have discussed mainly two concepts for the convergence of sequences of random variables, i.e. convergence in probability, and convergence in distribution. There are many other convergence concepts in the probability and statistics literature, but these two are perhaps the most often used and will suffice for this lecture.

Convergence in probability occurs when a random sequence converges towards a fixed value asymptotically. It can also be defined for a random limiting value, but in this lecture we only consider fixed limiting values. This also means that the distribution of the random sequences degenerates at a single point as n approaches infinity. The law of large numbers uses this notion of convergence to state that sample averages converge towards the expectation of the underlying random variables.

The second concept is that of convergence in distribution, which happens if the distribution of a random sequence does not degenerate and tends towards a given distribution, for example the Gaussian, as n increases. The CLT uses this concept and states that appropriately centered and scaled sample averages tend to a Gaussian distribution, irrespective of the underlying distribution of the random variables. Once we start taking averages or sums of random variables, there is a tendency of these objects to be closer to a Gaussian distribution than the original distribution. But keep in mind that this is only the case for the averages, not the individual r.v. that constitute the components of the average, which of course remain for example exponentially distributed even though the average is close to Gaussian.

In statistical inference, one is often interested in the distribution of a statistic for a given sample of n observations. If the sample size is large, the CLT can often be applied to approximate this distribution. If, however, n is small this is not possible. Many textbooks propose to add additional assumptions, such as normality of the underlying r.v., to obtain exact distributions in the small sample cases. We do not follow this strategy for various reasons that are summarized in the paper Hafner (2021). In this lecture, we will therefore not consider the case of small samples combined with (too) strong assumptions, but rather concentrate on the large sample cases.

1.6 Exercises

1.1: Show using Jensen's inequality that the kurtosis of a random variable is bounded below by 1.

1.2: Show using the Cauchy-Schwarz inequality that the correlation is bounded between -1 and 1.

1.3: Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $\mathbb{E}(Y_1) = \mu$ and $\mathbb{V}\mathrm{ar}(Y_1) = \sigma^2$. Show that $\mathbb{E}[\bar{Y}] = \mu$ and $\mathbb{V}\mathrm{ar}(\bar{Y}) = \frac{\sigma^2}{n}$, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

1.4: For a random sequence Z_n, suppose that

$$\mathbb{E}[Z_n] = \theta \quad \text{and} \quad \lim_{n\to\infty} \mathbb{V}\mathrm{ar}(Z_n) = 0.$$

Show using Chebychev's inequality that this implies convergence in probability, $Z_n \to_p \theta$.

1.5: (a) The lifetime of a machine component is a random variable having the following density

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{autrement} \end{cases}$$

As soon as the component breaks down, it is immediately replaced by another one. If X_i is the lifetime of the i^{th} component, S_n = $\sum_1^n X_i$ is the duration until the n^{th} break down and R_n is defined as

$$R_n = \frac{n}{S_n},$$

then, show that

$$R_n \rightarrow_p \frac{3}{2}$$

[Hint: Consider $G_n \rightarrow_p a$. If $g : \mathbb{R} \rightarrow \mathbb{R}$ is a continuous fonction in a, then $g(G_n) \rightarrow_p g(a)$.]

1.6: Suppose that X_1,...,X_n are $i.i.d$ continuous random variables with distribution $\mathcal{U}(0, \theta)$, for some $\theta > 0$. Show that $X_{(n)}$ = $\max_{1 \leq i \leq n}$ (X$_i$) converges in probability to θ.

1.7: Suppose that $X_1, ..., X_n$ and $Y_1, ..., Y_n$ are independent random samples from populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively. Show that the random variable

$$U_n = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}}$$

satisfies the conditions of the central limit theorem and thus that the distribution function of U_n converges to a standard normal distribution function as $n \rightarrow \infty$.

[Hint: Consider W_i = $X_i - Y_i$, for $i = 1, 2, ..., n$.]

2 Estimation

Consider a parameter of interest θ that characterizes a population $\mathcal{P}$. This could be, for example, the proportion of the Belgian population in favor of euthanisia. This parameter is generally unknown, unless one has a complete dataset about θ available for the entire population, which is unlikely. In order to estimate θ, i.e. obtain a plausible value $\hat{\theta}$, we draw a sample $\mathcal{E}$ from the population $\mathcal{P}$ and calculate an estimator based on the observed individuals in the sample. We call this a point estimator of the parameter θ.

The way we draw the sample is not innocuous, we have to ensure that it is a random sample, i.e. each member of the population has the sample probability of ending up in the sample. In that case, the sample is considered as representative for the population. Whether we can draw meaningful conclusions from the sample will also depend on other factors such as the sample size.

A second objective of estimation is to measure the uncertainty that is linked to our point estimator $\widehat{\theta}$. A way of doing this is to calculate an interval, typically centered at the point estimator, that contains the true parameter with a given high probability. If this interval is large, then the uncertainty is rather high, if the interval is narrow, the point estimator will be precise and our uncertainty low. This estimation problem is called interval estimation.

We therefore distinguish between point estimation and interval estimation. Point estimation is a rule to compute a plausible value for the parameter of interest only based on the observations in the sample. Interval estimation associates an interval to the point estimator such that it covers the true parameter with a given high probability. It is also only depending on the observations in the sample. This chapter deals with the problem of point estimators, what they are and what are their properties, while the problem of interval estimation is delegated to the next chapter.

The simplest example is the estimation of the population mean μ of some variable Y. This means that $\mu = \mathbb{E}(Y)$. We draw a sample of n observations and observe the values $Y_1, Y_2, ..., Y_n$. Because each member of the population has the same probability of ending up in the sample at Y_1, at Y_2, etc., these are independent and identically distributed (iid) random variables (r.v.). We can then construct an estimator of μ as the sample mean of the observations, i.e.

$$\overline{Y} := \frac{1}{n} \sum_{i=1}^{n} Y_i$$

This is an obvious and natural estimator of μ, however it is not the only one. An alternative could be the sample median. Because in general we have several potential estimators, we need to discuss the properties of the estimator in order to select the best one according to some criterion.

First some general statements about $\overline{Y}$. Unlike μ, which is a fixed parameter, $\overline{Y}$ is a function of the random variables $Y_1, Y_2, ..., Y_n$, and it is therefore random itself. This is because if we had taken another sample, we would have obtained also another estimator $\overline{Y}$. Because $\overline{Y}$ is random, it possesses a distribution which is called the sampling distribution of the estimator. This distribution will in general depend on the distribution of the underlying r.v. Y (which is not the same), but also on the sample properties and in particular the size of the sample (i.e. the number of observations, n). Using the sampling distribution, we can calculate moments such as the mean and the variance of $\overline{Y}$, which will turn out to be important properties to evaluate the quality of the estimator.

2.1 *Quality of point estimators*

We define a statistic to be a function of sample observations that does not depend on any unknown parameters. An estimator of an unknown parameter θ is defined as a statistic whose values are in the parameter space of θ, $\theta \in \Theta$. Because the sample is random, a statistic and an estimator will be random variables.

For example, to estimate the population mean $\mu = E[Y_1]$ based on a sample $Y_1, Y_2, \ldots, Y_n$, we could consider the estimators median(Y), $\overline{Y}$ or $Y_{(n)}$, where $Y_{(n)}$ is the observed maximum value of Y, also called the maximum order statistic. The latter is perhaps not the best estimator for the population mean, but it is at least a possible estimator. To give a counterexample, suppose that we have an estimator of the variance, S^2, which means that S^2 depends only on sample obervations, and it takes values in $\mathbb{R}_+$. Then, $-S^2$ is not an estimator of the variance. Although it is a function of sample observations, its values are not in $\mathbb{R}_+$, so it does not satisfy the second condition.

Our general notation for an estimator of an unknown parameter θ will be $\widehat{\theta}$. Because in general we have several possible estimators at hand, we need criteria to compare their advantages and disadvantages, and then decide for the optimal one. For us, the main statistical criteria are the following:

- bias

- mean squared error
- efficiency
- consistency

Additionally, we will often look at the sampling distribution of an estimator, because that is used for statistical inference. In the following we will define these concepts and give some examples.

Definition 2.1: The bias of an estimator $\hat{\theta}$ is defined as

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

An estimator is called unbiased if its bias is equal to zero.

The bias of an estimator gives information about whether the estimator correctly estimates the parameter on average, that is, in repeated samples. If the bias is negative, the estimator systematically (i.e., on average) underestimates the true parameter, if it is positive, then it systematically overestimates the true parameter. Of course, it is desirable to have an unbiased estimator. But it is not the only criterion, and we may often face situations where we prefer a slightly biased estimator because it has other advantages.

Definition 2.2: The mean squared error (MSE) of an estimator $\hat{\theta}$ is defined as

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

The estimation error is $\hat{\theta} - \theta$, which can be positive or negative. In order to measure the size or amplitude of this error, it is meaningful to consider its square, i.e. $(\hat{\theta} - \theta)^2$. The MSE now is defined as the average squared error in repeated samples. It is preferrable to have a small MSE because that implies small sizes of the estimation errors. Note that the MSE is not the same as the variance, which is defined as $\mathbb{V}\mathrm{ar}(\hat{\theta}) = E[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2]$. Only in the case of an unbiased estimator, i.e. $\mathbb{E}(\hat{\theta}) = \theta$, the MSE is equivalent to the variance.

We now have the following important result for the decomposition of the MSE in terms of bias and variance:

Theorem 2.1: Let $\hat{\theta}$ be an estimator of a parameter θ. Then,

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + B(\hat{\theta})^2 \tag{2.1}$$

So there is a very simple relationship between MSE, bias and variance. The MSE is equal to the variance plus the squared bias. This has several implications. First, as already mentioned, the MSE is equal to the variance if and only if the estimator is unbiased. Second, the MSE is never smaller than the variance, it usually will be larger in case of biased estimators. If we consider the MSE as a measure of our average size of error, then this theorem tells us something about the contributions to this measure: both bias and variance contribute. It can be that an estimator has a high MSE because it has a high variance, a high bias, or both. Essentially, the MSE gives a way to combine these two error sources, bias and variance, into a single criterion that allows us to compare estimators.

The proof of this theorem is very simple:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta}-\theta)^2] \\ &= \mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^2] \\ &= \underbrace{\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2]}_{\mathbb{V}\mathrm{ar}(\hat{\theta})} + \underbrace{(\mathbb{E}[\hat{\theta}]-\theta)^2}_{bias^2} + 2\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)] \end{aligned}$$

The third term is equal to zero because $\mathbb{E}[\hat{\theta}]-\theta)$ is a constant (the bias) that can be taken outside the expectation to obtain

$$\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)] = (\mathbb{E}[\hat{\theta}]-\theta)\underbrace{\mathbb{E}[(\hat{\theta}-\mathbb{E}[\hat{\theta}])]}_{0} = 0$$

This proves the theorem. In the following we consider two important examples, the estimation of the population mean and variance.

2.2 *Estimation of the population mean*

Consider the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as an estimator of the population mean, $\mu = \mathbb{E}(Y)$. First, we have

$$\mathbb{E}[\overline{Y}] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \mu,$$

and therefore the estimator is unbiased, i.e. its bias is zero. Second, we have already calculated its variance,

$$\mathrm{Var}[\overline{Y}] = \frac{\sigma^2}{n},$$

where $\sigma^2 = \mathbb{V}\mathrm{ar}(Y)$. Together, these results imply that

$$MSE(\overline{Y}) = \frac{\sigma^2}{n}.$$

Moreover, by the central limit theorem (CLT), we know that the sampling distribution can be approximated by

$$\overline{Y} \approx_d N(\mu, \frac{\sigma^2}{n})$$

in large samples.

The next example, the estimation of the population variance, is less trivial.

2.3 *Estimation of the population variance*

For a random sample $Y_1, Y_2, \dots, Y_n$, consider the following estimator of the variance σ^2:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

We begin by analyzing the bias properties of this estimator. Perhaps surprisingly, we have the following result.

Theorem 2.2:

$$\mathbb{E}[S^2] = \frac{n-1}{n}\sigma^2.$$

To prove this theorem, we first rewrite (left as an exercise)

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

Then,

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i^2\right] - n\mathbb{E}(\overline{Y}^2) = \sum_{i=1}^{n}\mathbb{E}Y_i^2 - n\mathbb{E}(\overline{Y}^2)$$

Note that Y_i is i.i.d. with mean μ and variance σ^2, and that by the definition of the variance, $\mathbb{E}(Y_i^2) = \sigma^2 + \mu^2$. Moreover, the mean and the variance of $\overline{Y}$ are given by, respectively, μ and σ^2/n, and therefore, $\mathbb{E}\overline{Y}^2 = \sigma^2/n + \mu^2$. Thus,

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\sigma^2$$

Dividing by n, we obtain the result. A consequence of this theorem is that S^2 is a biased estimator of σ^2, because

$$B(S^2) = \mathbb{E}[S^2] - \sigma^2 = -\frac{\sigma^2}{n}.$$

Two remarks are in order. First, the bias of S^2 is negative, i.e., S^2 on average underestimates the true variance. Second, the bias disappears as the sample size increases, and

$$B(S^2) = -\frac{\sigma^2}{n} \to 0 \text{ if } n \to +\infty.$$

The estimator S^2 is therefore asymptotically unbiased. In fact, in large samples, the bias of S^2 is negligible.

If we wanted to construct an unbiased estimator of σ^2, it is a simple exercise to show that the modified estimator

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

is indeed unbiased. In this class, however, we will mostly work with S^2 because it turns out to have other advantages, for example in terms of a smaller mean squared error.

We now turn to the variance of S^2 which is a little bit more tedious.

Theorem 2.3: Let $Y_1, \ldots, Y_n$ be an i.i.d. sample with $\mathbb{V}\mathrm{ar}(Y_1) = \sigma^2$ and kurtosis coefficient $K < \infty$. Then,

$$\mathbb{V}\mathrm{ar}(S^2) \approx \frac{\sigma^4(K-1)}{n}$$

where the approximation holds in large samples. The approximation error is of order $O(1/n^2)$, meaning that it decays rapidly in large sample.

The proof of this theorem requires several steps. It is beyond the scope of this class, but the interested reader is referred to Linton (2017, pp.165).

We can say something about the sampling distribution of S^2 in large samples, thanks to the central limit theorem:

Theorem 2.4: Under our conditions, as $n \to \infty$,

$$\sqrt{n}(S^2 - \sigma^2) \to_d N(0, \sigma^4(K-1))$$

This result enables us to do inference about the true population variance, i.e. constructing confidence intervals or do hypothesis tests.

2.4 Other estimators

Besides the population mean and variance, other parameters might be of interest such as:

- the "success" probability p of a Bernoulli experiment (such as tossing a coin where p would be $1/2$.

- The difference in means between two populations, μ_1 and μ_2. For example, the difference in average salaries of men and women. The point estimator would be the difference of sample averages for two independent samples of both groups, $\overline{Y}_1 - \overline{Y}_2$.

- The difference in variances between two populations, σ_1^2 and σ_2^2. The point estimator would be the difference of sample variances for two independent samples of both groups, $S_1^2 - S_2^2$.

- The difference between the probabilities p_1 and p_2 of a Bernoulli experiment for two populations, estimated by the difference in proportions, $\hat{p}_1 - \hat{p}_2$, for two independent samples.

The following table summarizes some of the considered parameters of interest, their point estimators, and first two moments.

parameter	sample	point	$E[\hat{\theta}]$	Variance
of interest θ	size	estimator $\hat{\theta}$		$\text{Var}[\hat{\theta}]$
μ	n	$\overline{Y}$	μ	$\frac{\sigma^2}{n}$
p	n	$\hat{p}$	p	$\frac{p(1-p)}{n}$
$\mu_1 - \mu_2$	n_1 et n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
$p_1 - p_2$	n_1 et n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
σ^2	n	S^2	$\frac{n-1}{n}\sigma^2$	$\approx \frac{\sigma^4(K-1)}{n}$

2.5 *Efficiency*

An estimator is called efficient in a certain class of estimators if it has the smallest MSE in that class. For unbiased estimators, this reduces to comparing their variances. Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of a parameter θ. Then $\widehat{\theta}_1$ is more efficient than $\widehat{\theta}_2$ if it has a smaller variance. We can express this in relative terms by computing the relative efficiency index,

$$\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{\text{Var}[\widehat{\theta}_2]}{\text{Var}[\widehat{\theta}_1]}.$$

If $\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) < 1$ (resp. > 1), we prefer $\widehat{\theta}_2$ to $\widehat{\theta}_1$ (resp. $\widehat{\theta}_1$ to $\widehat{\theta}_2$). If, for example, $\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = 1.2$, then we could say that $\widehat{\theta}_1$ is 20% more efficient than $\widehat{\theta}_2$.

Example 2.1: Estimation of a population mean

We estimate the population mean μ by the sample mean $\overline{Y}$. But we could also use the median $\widehat{q}_{1/2}$ defined as

$$\widehat{q}_{1/2} = \begin{cases} Y_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ is even.} \end{cases}$$

where $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ represent the order statistics of Y, i.e., the observations ranked in increasing order. The advantage of the median is that it is insensitive to outliers. However, it can be inefficient, depending on the distribution of Y. For example, if Y has a Gaussian distribution, then it can be shown that both estimators are unbiased, while

$$\text{Var}[\widehat{q}_{1/2}] = (1.2533)^2 \frac{\sigma^2}{n}.$$

and hence,

$$\text{eff}(\widehat{q}_{1/2}, \overline{Y}) = \frac{\sigma^2/n}{(1.2533)^2 \sigma^2/n} = 0.6366.$$

The variance of the sample mean is only about 63% of that of the sample median in the case of a Gaussian distribution. In this case, according to the efficiency criterion, the sample mean is the preferred estimator. Note that for other distributions the median might be preferable.

2.6 *Consistency*

Consider an estimator $\widehat{\theta}_n$ of an unknown parameter θ based on a sample of size n. To emphasize the dependence of $\widehat{\theta}_n$ on the sample size, we add the subscript n to it, and view it as a sequence of estimators, indexed by n, as n increases to infinity. One important question is whether the sequence of estimators converges, in a sense to be defined, to the true parameter. If that is the case, the estimator will be called consistent. Depending on the notion of convergence that is used, several definitions of convergence exist. We will use here the most common definition of convergence in probability, so that formally, an estimator is called consistent if, for any $\epsilon > 0$,

$$\lim_{n \to +\infty} P\left[|\widehat{\theta}_n - \theta| \leq \epsilon\right] = 1 \Leftrightarrow \lim_{n \to +\infty} P\left[|\widehat{\theta}_n - \theta| > \epsilon\right] = 0.$$

So if the sequence of estimators converges in probability to the true parameter as n increases to infinity, then the estimator is called consistent, and we usually denote this by $\widehat{\theta}_n \to_p \theta$. Intuitively, the probability that the estimation error $\widehat{\theta}_n - \theta$ is larger than an arbitrarily small ϵ, in absolute value, decreases in large samples, and in fact tends to zero asymptotically. The event $|\widehat{\theta}_n - \theta| > \epsilon$ becomes less and less likely in large samples, no matter how small ϵ. In other words, the distribution of the estimation error, $\widehat{\theta}_n - \theta$, concentrates more and more at zero.

As we already discussed in the first chapter, convergence in probability may be difficult to check in general. However, there is a very simple sufficient condition that is easy to verify. If both the bias and the variance of the estimator tend to zero as n increases, then the estimator is consistent. We emphasize that this is only a sufficient condition, so there may be estimators for which the sufficient condition does not hold but which nevertheless are consistent.

Example 2.2: Consider the estimation of the population mean μ by the sample mean $\overline{Y}$ of an i.i.d. sample of size n. We have shown that $B(\overline{Y}) = 0$ and

$$\text{Var}\left[\overline{Y}\right] = \frac{\sigma^2}{n} \to 0 \text{ as } n \to +\infty.$$

Thus, the sufficient condition holds and $\overline{Y}$ is a consistent estimator of the population mean μ. We can write shortly:

$$\overline{Y} \to_p \mu$$

The question arises if we can say something about the convergence of functions of consistent estimators, or combinations of several such estimators. For example, does the sum of two consistent estimators, the first converging to a and the second to b, converge to $a + b$? Fortunately, the answer is yes, and there are several rules for operations with consistent estimators which are summarized in the following. Let $\widehat{\theta}_n$ and $\widehat{\theta}'_n$ be two consistent estimators with $\widehat{\theta}_n \to_p \theta$ and $\widehat{\theta}'_n \to \theta'$. Then we have the following properties.

P1: $\widehat{\theta}_n + \widehat{\theta}'_n$ converges in probability to $\theta + \theta'$

P2: $\widehat{\theta}_n \times \widehat{\theta}'_n$ converges in probability to $\theta \times \theta'$

P3: $\widehat{\theta}_n / \widehat{\theta}'_n$ converges in probability to θ / θ' provided that $\theta' \neq 0$

P4: If $g : \mathbb{R} \to \mathbb{R}$ is a continuous function of θ, then $g(\widehat{\theta}_n)$ converges in probability to $g(\theta)$

Example 2.3: Let $Y_1, Y_2, \dots, Y_n$ be an i.i.d. sample with

$$E[Y_1] = \mu, \quad E[Y_1^2] = \mu'_2 \text{ and } E[Y_1^4] = \mu'_4 < \infty.$$

Then, S^2 is a consistent estimator of σ^2. To show this, first decompose S^2 as

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - \overline{Y}^2$$

By the law of large numbers, the first term on the right hand side converges to μ'_2 in probability. For the second term, $\overline{Y}^2$, recall first that $\overline{Y} \to_p \mu$ by the law of large numbers. Now what about $\overline{Y}^2$? Fortunately there is our property P4 which says that any continuous function g of $\overline{Y}$ will converge to $g(\mu)$ in probability. In our case, the function g is the square, so we have that $\overline{Y}^2 \to_p \mu^2$. Putting these things together,

$$S^2 \to_p \mu'_2 - \mu^2 = \mathbb{V}\text{ar}(Y) = \sigma^2.$$

This shows that S^2, although biased in finite samples, converges in probability to the true variance, σ^2.

We finish this chapter with a theorem that will turn out very useful for constructing confidence intervals, see the next chapter.

Theorem 2.5 (Slutsky): Let there be two random sequences U_n and W_n with the properties $U_n \to_d$ N(0,1) and $W_n \to_p 1$ as $n \to \infty$. Then,

$$\frac{U_n}{W_n} \to_d N(0,1) \text{ as } n \to \infty$$

Intuitively, the Slutsky theorem says that dividing an asymptotically Gaussian random sequence by another sequence that converges to one in probability does not change the distribution of the first sequence. This is good news because we will often be obliged to do this division, and can then use the same limiting distribution of the first sequence. We will give an example in the following.

Example 2.4: Let $Y_1, Y_2, \dots, Y_n$ be an iid sample of a population with mean μ and variance σ^2, and let

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Then we can use the CLT and Slutsky's theorem to show that

$$\sqrt{n}\frac{\overline{Y}-\mu}{S} \to_d N(0,1).$$

Indeed, thanks to the property P4 and to the fact that S^2 is a consistent estimator of σ^2, we know that

$$\frac{S}{\sigma} = \sqrt{\frac{S^2}{\sigma^2}} \to_p 1$$

This is the W_n sequence in the Slutsky theorem. Second, by the CLT we know that

$$U_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) \to_d N(0,1)$$

which is the sampling distribution of S^2. Then, the Slutsky theorem ensures that

$$U_n/W_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right)/\frac{S}{\sigma} = \sqrt{n}\frac{\overline{Y}-\mu}{S} \to_d N(0,1)$$

The usefulness of this result stems from the fact that we can replace the unknown standard deviation σ in the U_n statistic by the known S (which is an estimator of this standard deviation), without changing the asymptotic distribution. This can be used, e.g., for the construction of confidence intervals for μ, which will be discussed in the following chapter.

2.7 *Exercises*

1. Show that

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

2. The bias of the estimator S^2 for σ^2 is $-\sigma^2/n$. Show that the estimator

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$$

is unbiased.

3. Suppose that we want to estimate the difference between the proportions of men and women who are in favor of vaccination against some virus. To this end, we take two independent samples of men and women of size n_1 and n_2, respectively. The sample proportions of men and women in favor of vaccination are denoted by $\hat{p}_1$ and $\hat{p}_2$, respectively. Show that the estimator $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of the true difference, $p_1 - p_2$, and that

$$\mathbb{V}\mathrm{ar}(\hat{p}_1-\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Is the estimator consistent?

4. Suppose Y is a r.v. with $\mathbb{E}(Y) = \mu$ and $\mathbb{V}\text{ar}(Y) = \sigma^2$. We draw three values, Y_1, Y_2 and Y_3. To estimate μ, consider the sample mean,

$$\hat{\mu}_1 = \frac{1}{3}(Y_1 + Y_2 + Y_3).$$

Someone pretends that it may not be optimal to give equal weights (1/3) to each observation, and that it might be better to consider for example the estimator

$$\hat{\mu}_2 = \frac{1}{4}Y_1 + \frac{1}{2}Y_2 + \frac{1}{4}Y_3.$$

(a) Are both estimators unbiased?

(b) Calculate the relative efficiency $\text{eff}(\hat{\mu}_1, \hat{\mu}_2)$. In percentage terms, how much less efficient is the estimator $\hat{\mu}_2$ with respect to $\hat{\mu}_1$?

3 Confidence intervals

An interval estimator is a method to calculate the upper and lower limits of an interval as a function of observed data. As the data are obtained from a random sample, the interval will be random as well. Our objective in this chapter is to construct intervals that are as small as possible and cover the true parameter with a given high probability. This coverage probability is also called confidence level.

Formally, let us denote by $\hat{\theta}_L$ et $\hat{\theta}_U$ the lower and upper limits of the confidence interval for a parameter θ. The confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$ is then defined by the property

$$P\left[\theta \in [\hat{\theta}_L, \hat{\theta}_U]\right] = 1 - \alpha$$

where α is a small positive number, and $1 - \alpha$ is the confidence level. For example, if α is fixed at 5%, then the confidence level is 95%.

Confidence intervals are usually two-sided, i.e. bilateral, with an upper and lower limit. It is in principle possible also to construct unilateral confidence intervals such as $[\hat{\theta}_L, +\infty)$ or $(-\infty, \hat{\theta}_U]$, which would be defined by the properties

$$P[\hat{\theta}_L \leq \theta] = 1 - \alpha \quad \text{and} \quad P[\theta \leq \hat{\theta}_U] = 1 - \alpha.$$

Unilateral confidence intervals are, however, much less common and will not be considered further in this class.

In order to construct a confidence interval, we have to find probabilities related to the point estimator of θ. The problem is that the sampling distribution of $\widehat{\theta}$ itself depends typically on unknown parameters. For example, the sampling distribution of the sample mean depends on the unknown mean, or expectation, of the underlying variable. Hence, probabilities cannot be calculated using this partially unknown distribution. The trick is to center and standardize the point estimator in such a way that the resulting distribution is completely known, for example standard Gaussian. This new statistic is called a pivot. A pivot is a quantity that depends on the data and on the unknown parameter θ, but whose distribution does not depend on θ or any other unknown parameter. Thus, we can calculate probabilities related to the pivot and use them to construct confidence intervals. The following example illustrates the concept.

Example 3.1: Suppose $Y \sim N(\mu, 1)$ with unknown μ. Y is not a pivot because its distribution depends on an unknown parameter, in this case μ. After the transformation

$$Y - \mu \sim N(0, 1)$$

we obtain a distribution that is entirely known, i.e. it does not depend on any unknown parameters. Hence, $Y - \mu$ is a pivot and we can calculate probalities numerically. For example, by the properties of the standard Gaussian distribution, we know from the statistical tables, or by consulting a statistical software, that

$$P[-1.96 \leq Y - \mu \leq 1.96] = 0.95.$$

Can we use this information to say something about the likely range of values of the unknown μ, given an observation of Y? Indeed, consider the inequality

$$-1.96 \leq Y - \mu \leq 1.96$$

We would like to have just μ in the center of this inequality, and the bounds on the left and right hand sides only be dependent on Y. To this end, we first multiply the inequality by (-1):

$$1.96 \geq \mu - Y \geq -1.96$$

or equivalently

$$-1.96 \leq \mu - Y \leq 1.96$$

By adding Y on all sides,

$$Y - 1.96 \leq \mu \leq Y + 1.96$$

so that

$$P[Y - 1.96 \leq \mu \leq Y + 1.96] = 0.95$$

and we see that $[Y \pm 1.96]$ is the confidence interval for μ at confidence level 95%.

In general, as we have seen in Chapter 2, the point estimators $\widehat{\theta}$ of $\theta = \mu, p, \mu_1 - \mu_2$ and $p_1 - p_2$ are approximately Gaussian in moderate to large samples, thanks to the central limit theorem. We can write the approximate distribution as

$$\widehat{\theta} \approx_d N(\theta, \sigma^2_{\widehat{\theta}}) \tag{3.1}$$

where we have used the fact that the point estimators are unbiased, or at least approximately unbiased as in the case of S^2. The variance is denoted by $\sigma^2_{\widehat{\theta}} := \mathbb{V}\mathrm{ar}(\hat{\theta})$. Note that the form of $\sigma^2_{\widehat{\theta}}$ depends on the particular estimator, it will be different for the estimation of a mean, a proportion, a variance, a difference between means, etc.

Note that $\widehat{\theta}$ in (5.1) is not a pivot because its distribution depends on the unknown θ, and on $\sigma^2_{\hat{\theta}}$. But we can standardize it:

$$U = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \approx N(0, 1)$$

and because U has (approximately) a standard Gaussian distribution, we can calculate probabilities.

Now we could proceed as in the example above by setting up an interval for U that is realized with a certain probability, and then reformulate to obtain a corresponding interval for the unknown θ. The bounds of the interval would however depend on the variance of the estimator, $\sigma^2_{\hat{\theta}}$, which is unknown in general. So are we lost here?

Fortunately not, if we have a consistent estimator of $\sigma^2_{\hat{\theta}}$, which we call $\hat{\sigma}^2_{\hat{\theta}}$. Thanks to Slutsky's theorem, we have

$$Z = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{\hat{\theta}}} \approx N(0, 1)$$

which then can be used as a pivot, i.e. it depends only on data and the unknown θ, and its distribution is completely known (at least approximately). Because

$$P[-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}] \approx 1 - \alpha$$

we can obtain a confidence interval for θ by substituting $\frac{\hat{\theta}-\theta}{\hat{\sigma}_{\hat{\theta}}}$ for Z, and then isolating θ at the center of the inequality, similar to the example above, which will then give us

$$[\hat{\theta} - z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}}, \hat{\theta} + z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}}]$$

which is sometimes written shortly as $[\hat{\theta} \pm z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}}]$. So this is the confidence interval for the estimation of θ at a given confidence level of $1 - \alpha$. We see that its construction is quite simple: Take the point estimator, and then add and subtract the quantile of the Gaussian distribution multiplied by the estimated standard devitation of the estimator. This concept applies to all special cases that we consider, and so the structure of all confidence intervals that we consider remains the same.

Note that the half-length of a confidence interval, i.e. $z_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}}$, is sometimes called the error margin of the estimator $\hat{\theta}$ at a given confidence level.

Example 3.2: This example is about confidence intervals for a proportion p. Suppose that an institute surveys $n = 1000$ individuals about their political preferences. As a response to the question "Would you vote for Mr Dupont?", 560 persons say yes. The point estimate of the proportion p of the population who vote in favor of Mr Dupont is

$$\hat{p} = \frac{560}{1000} = 0.56.$$

How do we obtain a confidence interval for the unknown proportion p? First, we use the fact that, by the central limit theorem,

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

This is a special case of our general result in(5.1), where θ is now the proportion p, and the variance of the estimator, $\sigma^2_{\hat{\theta}}$ is given by $\frac{p(1-p)}{n}$. The unknown variance can be estimated consistently by $\frac{\hat{p}(1-\hat{p})}{n}$, because $\hat{p}(1-\hat{p}) \rightarrow_p p(1-p)$ (by property P2).

We obtain the pivot

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1).$$

which uses the CLT and Slutsky's theorem. We proceed exactly as in the general case and obtain the confidence interval

$$P\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \approx 0.95$$

That is, the confidence interval at level 95% is given by:

$$[0.53; 0.59],$$

and the corresponding error margin is 3%.

Another example of a parameter of interest θ may be the unknown variance σ^2, for which we proposed the point estimator S^2. Can we also construct a confidence interval for σ^2? Indeed, this is just another special case of our general approach. In the case of the variance, we have discussed the following approximation in large samples, based on the central limit theorem:

$$S^2 \approx_d N(\sigma^2, \frac{\sigma^4(K-1)}{n})$$

So now the variance of the estimator, $\sigma^2_{\hat{\theta}}$ in our general notation, is equal to $\frac{\sigma^4(K-1)}{n}$. We do not know this variance as it depends on the unknown σ^2 and K, the population kurtosis. However, we can replace both by their sample equivalents, and the distribution of S^2 in large samples will be unaffected, thanks to Slutsky's theorem. Thus, we estimate the variance of S^2 by $\frac{S^4(\hat{K}-1)}{n}$.

Using again the same approach as in the general case, we then construct the following confidence interval at level $1-\alpha$ for σ^2:

$$\left[S^2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{S^4(\hat{K}-1)}{n}}\right]$$

Note that this is only an approximation based on the asymptotic distribution. It may be that the interval covers negative values, which is not useful as we know that the true variance cannot be negative. However, this issue disappears in large samples, because the error margin of the confidence interval decreases, and so the interval concentrates more and more around the (positive) point estimator S^2.

After having discussed some special cases, we want to make some general observations about the determinants of the length of confidence intervals, or equivalently, of the error margin, which is half the length of the confidence interval.

The error margin of a bilateral interval at level $1 - \alpha$ is equal to

$$z_{\alpha/2}\ \hat{\sigma}_{\hat{\theta}}$$

It depends on two things: First, the confidence level $1 - \alpha$: the higher it is, the wider is the confidence interval. This makes sense as you become more confident that the true parameter is in the interval when it is wider. Second, the precision of the estimator. The more the estimator is precise (i.e. the smaller its standard deviation), the narrower will be the confidence interval. Again, this makes sense, because if an estimator is precise, it provides more information about the true parameter and the corresponding interval will be narrow.

Finally, a word about the interpretation of confidence intervals. We have to be a bit careful. Having observed a sample and calculated a confidence interval, such as the [0.53;0.59] in the preceding example, it is not true to say that "the probability that this interval contains the true parameter is 95%", because this probability is either 1 (when it is inside), or 0 (when it is not). The correct interpretation is in terms of repeated samples: If one draws many independent samples of the population, and each time recalculates the confidence interval, then in about 95% of the cases these intervals will cover the true parameter.

To summarize this chapter, the point estimators of chapter 2 are often associated with an interval, called the confidence interval, which is a function of the observations, and for which one is "confident" that it contains the true parameter. In general, a confidence intervals is obtained by taking the point estimator and adding a subtracting an error margin which depends on two factors: the confidence level and the estimator precision. The confidence interval widens with increasing confidence level, or lower estimation precision. Confidence intervals can be used in various circumstances. In particular, we will see that there is a close link with hypothesis tests, which we will discuss in Chapter 5.

3.1 Exercises

1. For a confidence interval at level 98% for the population mean, with an estimated standard deviation of 15, what is the minimum number of observations such that the error margin is smaller than 5?

2. Consider a confidence interval for the variance σ^2 based on the asymptotic approximation. Suppose that with a sample of $n = 20$ observations, the sample variance is $S^2 = 2$ and sample kurtosis $\hat{K} = 3$. What is the maximum confidence level $1 - \alpha$ such that the interval does not cover negative values?

3. Someone claims that the reaction time of men in video games is faster than that of women. To investigate this question, two independent groups of 50 men and 50 women are given a particular task in a video game, and the reaction times of men (X) and women (Y) is measured (in milliseconds). The sample means are $\bar{X} = 3.6$ and $\bar{Y} = 3.8$, and the corresponding sample variances $S_X^2 = 0.18$ and $S_Y^2 = 0.14$.

   (a) Calculate a 95% confidence interval for the difference in reaction times between men and women. Based on this interval, would you be confident (at level 95%) that the difference is positive? Explain.

   (b) Now recalculate the same confidence interval but at a level of 99%. Would you now still be confident that there is a difference? Explain.

4
Estimation methods

In the previous chapters we have presumed that point estimators exist, in a sense they fall from the sky and we have looked at their statistical properties. This is plausible in simple situations, for example the estimation of the population mean, where the sample mean is a plausible and natural estimator with, as we have seen in Chapter 2, good statistical properties. In general, however, it may not always be so easy to come up with a suitable suggestion for an estimator. In complex nonlinear models, for example, no natural estimator is available a priori, and we have to search for methods to obtain an estimator. This chapter deals with such estimation methods.

We will mainly discuss three estimation methods: (1) the method of moments, (2) the method of maximum likelihood, and (3) the method of least squares. Since the latter method, least squares, is linked to linear regression models, we delay its discussion to Chapter 7. The first two methods are discussed in the following.

4.1 *Method of moments*

The idea of the method of moments is to match sample moments (which can be calculated using the data) with population moments (which are unobserved but depend on the unknown parameters), obtaining a system of equations that can be solved for the parameter.

Suppose we have a random sample $Y_1, \ldots, Y_n$ of a population variable Y whose distribution depends on an unkown p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. The first p moments of Y, $\mu'_k = E[Y_1^k]$, depend on $\boldsymbol{\theta}$. To express this dependence, we write $\mu'_k \equiv \mu'_k(\boldsymbol{\theta})$. The idea is now to match $\mu'_k(\boldsymbol{\theta})$ with the corresponding sample moments, for $k = 1, \ldots, p$. We obtain a system of p equations with p unknowns, that can be solved with respect to the parameter. This parameter can then be calculated numerically as a function of the sample information, and so it is an estimator that is called $\widehat{\boldsymbol{\theta}}$.

Formally, the method of moments consists of choosing as an estimator $\widehat{\boldsymbol{\theta}}$

of $\boldsymbol{\theta}$ the solution of the system of equations

$$\mu_k'(\boldsymbol{\theta}) = m_k', \ \ k = 1, \ldots, p.$$

where $m_k' = \frac{1}{n}\sum_{i=1}^n Y_i^k$ is the k-th sample moment, which is an estimator of $\mu_k'(\boldsymbol{\theta})$. Thus, the first equation matches population and sample mean, the second equation the second moments, etc. until one has as many equations as the dimension of the parameter vector. The method is justified by the fact that sample moments will converge to population moments in large samples, thanks to the law of large numbers. Therefore, the methods of moments estimator will also, under quite weak conditions, be a consistent estimator of the true parameter.

Example 4.1: Suppose that the population variable is Gaussian, i.e. $Y \sim N(\mu, \sigma^2)$, with two unknown parameters, μ and σ^2. So here the parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$, and its dimension is $p = 2$. We will therefore need to consider two equations to obtain the method of moments estimator:

$$\begin{aligned} \mu_1' = E[Y_1] = \mu \quad &= \quad m_1' = \frac{1}{n}\sum_{i=1}^n Y_i \\ \mu_2' = E[Y_1^2] = \sigma^2 + \mu^2 \quad &= \quad m_2' = \frac{1}{n}\sum_{i=1}^n Y_i^2 \end{aligned}$$

In this system of two equations, we have the population moments μ_1' and μ_2' on the left hand side. These depend on the unknown parameters: $\mu_1' = \mu$ and $\mu_2' = \sigma^2 + \mu^2$. The latter equation follows because by the definition of the variance, $\sigma^2 = \mu_2' - \mu^2$. On the right hand side of the equations, we have the corresponding sample moments m_1' and m_2'. These can be calculated from the data in the sample, so these are known quantities. We now have to solve this system for μ and σ^2, which will give us the estimators $\widehat{\mu}$ and $\widehat{\sigma^2}$. For $\widehat{\mu}$ this is trivial as we obtain from the first equation directly:

$$\widehat{\mu} = m_1'.$$

Note that m_1' is nothing but our usual sample mean of Y, which we have also called $\bar{Y}$ previously. Plugging this estimator into the second equation, and solving for σ^2, we then obtain the estimator for σ^2 :

$$\widehat{\sigma^2} = m_2' - (m_1')^2.$$

Note that this estimator is the same as our variance estimator S^2. In other words, the method of moments confirms our choice of using the sample mean and variance as estimators of the population mean and variance, respectively.

In the previous example, there were two unknown parameters, and both the first and second moments depended on at least one of those parameters. In that case, one obtains two equations with two unknowns that can be solved. It may be the case, however, that the first moment does not depend on the unknown parameter, in which case it is simply skipped and one proceeds with the second moment. This is the case in the next example.

Example 4.2: A famous distribution is the so-called student-t distribution whose density $f(x)$ is proportional to $(1 + x^2/\nu)^{-(\nu+1)/2}$, where ν is a parameter. If $\nu > 2$, then one can show that $\mathbb{E}[X] = 0$ and $\mathbb{V}\mathrm{ar}(X) = \frac{\nu}{\nu-2}$, and therefore

$$\nu = \frac{2\mathbb{V}\mathrm{ar}(X)}{\mathbb{V}\mathrm{ar}(X) - 1}$$

Thus, we have a single equation, because there is only one parameter, and this expression does not depend on the mean but on the variance of X. To obtain the method of moments estimator, we can now simply replace the unknown variance by the sample variance of the data,

$$\widehat{\nu} = \frac{2\widehat{\sigma}^2}{\widehat{\sigma}^2 - 1},$$

which is very simple to compute.

In general, the method of moments has some good properties: First, it is often quite easy to obtain. Second, it provides consistent parameter estimators, because thanks to the law of large numbers, $m_k' \rightarrow_p \mu_k'(\theta)$, which typically implies that $\hat{\theta} \rightarrow_p \theta$, so that the estimator is consistent. However, a drawback of the method of moments is that it is often not efficient, and it may be possible to find other, more efficient, estimators. This means that it would be possible to find other estimators with a smaller variance than the method of moments. It is also not necessarily an unbiased estimator of the true parameter, because sample moments are not necessarily unbiased estimators of the corresponding population moments. Indeed, we have seen that $\hat{\mu}$, the methods of moments estimator of the population mean μ, is an unbiased estimator, but $\widehat{\sigma}^2$, the method of moments estimator of σ^2, is biased, although the bias decreases towards zero as the sample size increases, i.e., it is asymptotically unbiased.

4.2 *The method of maximum likelihood*

The method of maximum likelihood is a very general estimation method that can be applied in many different contexts in statistics. Its main idea is to define as estimator that quantity that is most likely to having produced the data that are available in the sample. This is a probabilistic statement, and so the essential feature of maximum likelihood estimation (MLE) is the specification of the joint density of the observations, which will depend on the parameter. The question is then, which value of the parameter maximizes this joint density, given the observations of the sample.

To illustrate the maximum likelihood principle, let us first look at an example.

Example 4.3: Consider the problem of estimating the unknown probability p of Bernoulli experiment, for example obtaining heads when tossing a coin. Suppose that we toss the coin ten times, and obtain 7 times heads out of these 10 tosses. Without any prior information, what would be a plausible estimator of p? The idea of MLE is to look for that parameter value p that makes it most likely of having observed 7 heads out of 10 tosses. In this case, the MLE will be the proportion of heads in the sample, so

$$\widehat{p} = \frac{7}{10}.$$

We will see later that indeed this is the MLE in this example.

In general, suppose that we want to estimate a parameter θ that characterizes the distribution f of a random variable Y. For an i.i.d. sample $Y_1, \ldots, Y_n$, the joint probability (or density) function is given by

$$\prod_{i=1}^{n} f(Y_i, \theta).$$

Because of the independence of Y_i and $Y_j, i \neq j$, the joint density is just the product of the marginal density functions. For a given sample $Y_1, \ldots, Y_n$, this function is now interpreted as a function of θ, and called the likelihood function:

$$L(\theta) := \prod_{i=1}^{n} f(Y_i, \theta)$$

In practice, replacing the r.v. Y_i by their realizations in the sample, this function will indeed only depend on θ. The idea of MLE is then to find the parameter value θ that maximizes the plausibility, or likelihood, of having observed the data in the sample. Technically, this means that MLE maximizes the likelihood function with respect to θ. Formally, the MLE is defined by

$$\widehat{\theta} := \arg\max_{\theta} L(\theta)$$

Note that often it is more convenient to maximise the logarithm of the likelihood function, because the logarithm of a product is the sum of the logarithms. Maximising the log likelihood function gives the same result as maximising the likelihood function directly, because the logarithm is a monotone increasing function. Hence, $\arg\max_\theta L(\theta) = \arg\max_\theta \log L(\theta)$.

Let us now come back to our example.

Example 4.4: Let $Y_1, Y_2, \ldots, Y_n$ be an i.i.d. sample of a Bernoulli population with unknown parameter p (as above in the coin example). That is,

$$Y_1 = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p. \end{cases}$$

The MLE of p maximizes the likelihood function $L(p)$ given by

$$L(p) = \prod_{i=1}^{n} f(Y_i, p) = p^{\sum_{i=1}^{n} Y_i} (1-p)^{n - \sum_{i=1}^{n} Y_i}$$

or, equivalently, the log likelihood

$$\ln L(p) = \ln p \sum_{i=1}^{n} Y_i + \ln(1-p)\left(n - \sum_{i=1}^{n} Y_i\right).$$

This function can easily be maximized by setting the derivative of $\ln L(p)$ with respect to p to zero, i.e.

$$\frac{1}{p}\sum_{i=1}^{n} Y_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} Y_i\right) = 0 \Leftrightarrow \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

We see that $\widehat{p}$ is just the sample proportion of obtained successes. So in our case above, $\sum_{i=1}^{n} Y_i$ is the number of times heads was obtained among 10 tosses, which was 7, and dividing this quantity by 10 gives the proportion 0.7 as the MLE estimator.

In general, the MLE has very good statistical properties. Under quite weak conditions, it is asymptotically unbiased, consistent, efficient, and asymptotically Gaussian. In particular, in many cases it will be more efficient than the method of moments estimator. However, it may require some more efforts to obtain the MLE because it involves a maximization problem. In simple cases such as those we discuss in this class, this is easy, but in more complex situations, no simple analytical solution of the maximization problem might be available. In that case, one might employ numerical algorithms such as Newton-Raphson to obtain the MLE. This is actually the normal procedure in most practically relevant models and applications.

MLE have an important invariance property: If $\widehat{\theta}$ is the MLE of θ, then $g(\widehat{\theta})$ is the MLE of $g(\theta)$, for some function g. For example, if Y has a Poisson distribution with parameter λ, then $P(Y=0) = e^{-\lambda}$. We may wish to estimate this probability. Having an MLE for λ, say $\widehat{\lambda}$, we then know by the invariance property that the MLE of $P(Y=0)$ is given by $e^{-\widehat{\lambda}}$.

4.3 Exercises

4.1: If $Y_1, Y_2, ..., Y_n$ denote a random sample from the normal distribution with known mean μ = 0 and unknown variance σ^2, find the method-of-moments estimator of σ^2.

4.2: If $Y_1, Y_2, ..., Y_n$ denote a random sample from the normal distribution with unknown mean μ and variance σ^2, find the method-of-moments estimator of μ and σ^2.

4.3: Let $Y_1, \ldots, Y_n$ denote a random sample from a population Y with density function

$$f(y|\theta) = (\theta+1)y^{\theta}, \quad 0 < y < 1, \quad \theta > -1$$

Find an estimator for θ by the method of moments, and show that the estimator is consistent.

4.4: Let $Y_1, \ldots, Y_n$ denote a random sample from a population Y with a Poisson distribution and parameter λ. Find an estimator of λ by the method of moments.

4.5: Let $Y_1, \ldots, Y_n$ denote a random sample from a population Y with a Poisson distribution and parameter λ. Find an estimator of λ by the method of maximum likelihood, and show that the estimator is consistent for λ.

5
Hypothesis tests

Hypothesis tests are an important tool for statistical inference that allow to gain information about the underlying population, and to take decisions concerning hypotheses that may be of interest. At the core is the formulation of hypotheses concerning parameters of the population. These hypotheses may reflect theories of the domain of application, such as biology, psychology, or economics. These theories are then confronted with the data, and a decision will be made in favor, or against, the hypotheses and related theories.

The hypothesis under study will be rejected if the data clearly contradict the hypothesis. If it is not rejected it may be that either the hypothesis is correct, or that there is simply not sufficient information in the data to be able to reject it. Intuitively, with only very few observations, one will be prudent to reject a hypothesis even if it does not seem to compatible with the data. On the other hand, with a lot of observations, and hence information, in the sample, one will more likely reject it in that case.

We already see that there is some kind of asymmetry in the decision process: The hypothesis under study should only be rejected if there is clear evidence against it. If it is not rejected, however, it does not mean that we take it for granted, we may actually still be convinced that it is wrong. It is just that the statistical information of the sample may not be sufficient to be able to claim that the hypothesis is wrong.

To better grasp this asymmetry, it is helpful to look at an analogy of hypothesis tests with a criminal trial with an accused and a jury. In this case, the population corresponds to the universe of facts, which are not all known, and the truth, which means the culpability or not of the accused (also unknown). The sample corresponds to all available information and witnesses during the trial. The hypothesis to test is the hypothesis of innocence. We will later call this the "null hypothesis".

The jury will find the accused guilty if the available information implies a very high probability of his or her culpability. In this case, it will reject the hypothesis of innocence. It is important to recognize that the hypothesis of innocence has to be protected, that is, without sufficient evidence the jury cannot find the accused guilty. In this case, it cannot reject the hypothesis of innocence, although it may be convinced of the culpability.

In our terminology: The null hypothesis should only be rejected if there is strong evidence against it, because in that case we can be confident that it is indeed incorrect. On the other hand, if we do not reject the null hypothesis, it does not imply that we consider it as true. It may just mean that we do not have enough evidence against it, perhaps because the information in the data is not sufficient.

Consider an introductory example of a politician, Mr. Jones, who claims to have the majority of electoral votes in the population. So in our terminology, the null hypothesis (denoted H_0) is that $p \geq 50\%$, where p is the proportion of electoral votes in favor of Mr. Jones. If we do not believe his claim, we can formulate an alternative hypothesis (denoted H_1), $p < 50\%$, i.e. that Mr. Jones does not have the electoral majority, and try to reject the null hypothesis given statistical evidence.

Formally, we can write the system of hypotheses as

$$\begin{cases} H_0 : p \geq 0.5 \\ H_a : p < 0.5 \end{cases}$$

where p is the proportion of electoral votes in favor of Mr. Jones.

Now, H_0 is an inequality and it is not clear a priori which value of p should be considered under H_0. It turns out that most "protects" H_0 is just at the border, i.e. $p = 0.5$, because that value is closest to the alternative hypothesis. If H_0 is rejected for this value of p, then it will be rejected for any other value of p under H_0, i.e. $p \geq 0.5$. However, this is not true the other way around: It might that H_0 is rejected for $p = 0.7$, but not for $p = 0.5$. So the idea is to replace the inequality of H_0 by an equality to that value that minimizes the probability of rejecting H_0 when it is true. Thus, we reformulate the system of hypotheses as:

$$\begin{cases} H_0 : p = 0.5 \\ H_a : p < 0.5 \end{cases}$$

In our example of Mr. Jones, we are looking for statistical evidence in order to take a decision concerning his claim of having the majority. To this end, we may ask $n = 15$ voters, in a representative survey, and denote by Y the number of voters in favor of Mr. Jones. Intuitively, we should reject H_0 if Y is "small", because then it would not be likely that Mr Jones has the majority in the population. We will need to determine what is "small" using statistical criteria. Whatever decision we take, it is possible that we are wrong: For example, it is not impossible to obtain $Y = 0$ even if H_0 is correct, but it is quite unlikely. It would be much more likely to observe $Y = 0$ when H_a is correct. Therefore, in this case we would reject H_0 in favor of H_a. On the other hand, it is not impossible to obtain $Y = 15$ even if H_0 is incorrect, but it is again quite unlikely. It would be much more likely to observe $Y = 15$ when H_0 is correct. Therefore, in this case we would not reject the null hypothesis.

5.1 Elements of a hypothesis test

All hypothesis tests share the same procedure and are composed of the following four elements:

- The null hypothesis, H_0
- The alternative hypothesis, H_a
- The test statistics, T
- The rejection region, RR

As statisticians, we are in the role of the jury at a criminal trial: We want to find evidence against the null hypothesis (innocence of the accused), in which case we would reject it. So in fact, we typically would like to emphasize the alternative hypothesis (culpability). However, we can only do so if the evidence against the null hypothesis is sufficiently strong, because of the necessary presumption of innocence.

The decision will be based on statistical evidence, which is obtained by the numerical result of a test statistic. This statistic can be calculated using the data at hand, just like parameter estimators. The rejection region RR is the set of values of the test statistic that lead to a rejection of H_0 in favor of H_a. This region needs to be defined prior to observing the test results, and it depends primarily on our tolerance of making erroneous decisions.

The decision rule is therefore the following: If the test statistic falls into the rejection region RR, we reject H_0, which will be denoted by RH_0. If, on the other hand, the test statistic does not fall into RR, then we do not reject H_0, which will be denoted by $\overline{R}H_0$.

After having specified the two hypotheses, we have to define a test statistic and a rejection region in order to take a decision. The question is, how? In the example of Mr Jones, the test statistic will be the number of people in the sample in favor of Mr Jones. it is quite clear that the rejection region will be of the form

$$RR = \{0, 1, \ldots, k\}$$

for some small integer k, because if the number of people in the sample that are in favor of Mr Jones is small, we have evidence against the null hypothesis that he has the majority in the population. The question is, how to fix k?

However we choose k, two types of error can occur:

- Type I error: We reject H_0 although it is true: $\alpha = P[RH_0|H_0]$ is called the level of the test

- Type II error: We do not reject H_0 although it is false: $\beta = P[\overline{R}H_0|H_a]$ is called the type II error probability

Altogether, there are four scenarios, two with the right decisions, and two with wrong decisions. These cases are summarized in the following table:

	Decision	
Reality	RH_0	$\overline{R}H_0$
H_0	Type I error (α)	OK! $(1-\alpha)$
H_a	OK! $(1-\beta)$	Type II error (β)

The table gives for each case, in parentheses, the corresponding probability conditional on the reality. So, for example, α is the probability of rejecting H_0 conditional on H_0 being true, and β is the probability of not rejecting H_0 conditional on H_a being true. These are the two error probabilities. For a given reality, i.e., a given row of the table, the probabilities have to sum to 1. That means that the probability of not rejecting H_0 when it is true, which is the right decision, must be equal to $1-\alpha$ (such that $\alpha + 1 - \alpha = 1$), and similarly, the probability of rejecting H_0 when it is false, again a correct decision, must be equal to $1-\beta$ (such that $1-\beta+\beta = 1$).

In the example of Mr Jones, the type I error consists in concluding that he will lose the election (RH_0) although he will win, and the type II error occurs if one concludes that he will not lose the election although he does.

Because of the asymmetry of the hypotheses, illustrated with the criminal trial analogy, it is much more important to control for the type I error probability, so to keep α at a small level. The standard procedure of the statistical methodology is to fix α at some small value close to zero, for example 0.01 or 0.05. This is because one is forced to protect the null hypothesis, and only reject if one is almost certain that it is not true (with a very small error probability α). In a second step, provided that several tests are available, one searches for the one that minimizes the type II probability, β. But very often only one test is available. If, then, β turns out to be large, then one has to careful when interpreting the results in the case of $\overline{R}H_0$: In the case of the accused, it could simply mean that, although he is guilty in reality, we cannot find him guilty due to lacking evidence. We will later call this a lack of power of the statistical test.

In the case of Mr Jones, consider the choice of $k = 2$, i.e. the rejection region becomes: $RR = \{0, 1, 2\}$. What are the corresponding values of α and β? We can calculate this using our knowledge of the distribution of Y, the number of people in favor of Mr Jones, under the null hypothesis that $p = 0.5$. The random variable Y follows a binomial distribution with parameters $n = 15$

and $p = 0.5$. We can therefore calculate:

$$\begin{aligned} \alpha &= P[\text{Type I error}] \\ &= P[RH_0 | H_0 \text{ is true}] \\ &= P[Y \leq 2 | Y \sim Bin(15, 0.5)] \\ &= \binom{15}{0}(0.5)^{15} + \binom{15}{1}(0.5)^{15} + \binom{15}{2}(0.5)^{15} = 0.004. \end{aligned}$$

Thus, for our decision rule "RH_0 if $Y \leq 2$", the probability to erroneously reject H_0 is 0.004. This is quite negligible.

But what about β? To calculate β, we have to assume that the alternative hypothesis is correct. The alternative is, however, composed of an infinite number of possible values for p, because $p < 0.5$ under H_a. To calculate β we have to specify for which value of p, $p < 0.5$, we want to do it, because as we will see, β will depend on p. For example, we could fix p at 0.3. For this case, we can calculate β for our decision rule "RH_0 if $Y \leq 2$" as follows:

$$\begin{aligned} \beta &= P[\text{Type II error}] \\ &= P[\overline{R}H_0 | H_a \text{ is true}] \\ &= P[Y > 2 | p = 0.3] \\ &= \sum_{y=3}^{15} \binom{15}{y} (0.3)^y (0.7)^{15-y} = 0.873 \end{aligned}$$

We see that the decision rule "RH_0 si $Y \leq 2$" leads to a negligible type I error probability, but a very high type II error probability of 87.3% if the true proportion p is 0.3. In other words, if the true proportion in favor of Mr Jones is 30%, then it is quite unlikely to discover that the null hypothesis is incorrect using our decision rule.

Two remarks are important: First, β is obviously a decreasing function of the difference between p and 0.5 (i.e. H_0). The reason is that, if the distance is small, for example $p = 0.49$ for H_a, it is very difficult to distinguish between H_0 and H_a, and we will often not reject $H_0 : p = 0.5$ even though $H_a : p = 0.49$ is true. Here, β will be high. But once p decreases, the difference between the two hypotheses increases and it becomes easier to distinguish them. Then, we will more often reject H_0 when it is false, and hence β decreases.

The second remark is about the relation between α and β: for a given sample size, these error probabilities are conflicting: One cannot decrease both simultaneously. If one wishes to decrease α by choosing another rejection region, then β increases, and vice versa. Only by increasing the sample size it is possible to decrease both α and β simultaneously, as we will see.

In our example, given that α is very small, it might seem acceptable to increase it a little bit such that β is reduced. For example, if we replaced the rejection region $RR = \{0,1,2\}$ by $RR^* = \{0,1,\ldots,k\}$ with $k \geq 3$, then

$$\begin{aligned} \alpha^* &= P[T \in RR^*|H_0] \geq P[T \in RR|H_0] = \alpha \\ \beta^* &= P[T \notin RR^*|H_a] \leq P[T \notin RR|H_a] = \beta. \end{aligned}$$

For example, if we choose $k = 5$, we can calculate:

$$\alpha = P[Y \leq 5|p = 0.5] = \sum_{y=0}^{5} \binom{15}{y} (0.5)^{15} = 0.151$$

and, if the true value of p under H_a is 0.3,

$$\beta = P[Y > 5|p = 0.3] = \sum_{y=6}^{15} \binom{15}{y} (0.3)^y (0.7)^{15-y} = 0.278.$$

If a type I error probability of slightly more than 15% is still acceptable, then this might be a suitable choice because the type II error probability has been substantially reduced. In general, α is typically fixed at values of 0.01 or 0.05, because it is reasonably close to zero, but not too close such that there is not an excessive type II probability. In our example, one might choose $k = 4$, and obtain an α that is close to these typical values (left as an exercise).

5.2 *Tests of a parameter θ*

We now consider the problem of testing a parameter θ in general, where θ is our generic symbol for the parameter of interest that might represent a mean, a proportion, a difference in means, a variance, etc. As we have seen with many examples, the point estimator $\widehat{\theta}$ can be assumed to be normally distributed in large samples, thanks to the central limit theorem, and is at least approximately unbiased. We denote its variance by $\sigma^2_{\widehat{\theta}}$, but note that the particular form of this variance depends on the type of estimator at hand. So we have

$$\widehat{\theta} \approx_d N(\theta, \sigma^2_{\widehat{\theta}}). \tag{5.1}$$

Of course, the true value of θ is unknown, but it might be of interest to test a particular value of θ, because for example that value corresponds to a certain economic theory. Let us call this value of interest θ_0, and $\theta_0 \in \Theta$, where Θ is the parameter space of θ. We want to test the following hypotheses:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_a : \theta > \theta_0. \end{cases}$$

The alternative hypothesis is here formulated as ">", but this only one possibility and it depends on the particular situation of the study. Another possibility is, of course, to specify the alternative hypothesis as $H_a : \theta < \theta_0$. Both of these possibilities are so-called unilateral tests, because a rejection of H_0 will be possible only on one side of θ_0, either to the left or to the right. It would also be possible, or course, to use a so-called bilateral test as $H_a : \theta \neq \theta_0$, because here a rejection would be possible to both sides of θ_0. Note that for all three possibilities, the null hypothesis $H_0 : \theta = \theta_0$ remains unchanged.

To test the null hypothesis, we have to use a test statistic whose distribution is known under H_0. The starting point is (5.1), but the distribution is not pivotal because it depends on unknown parameters. But we can standardize $\widehat{\theta}$ to obtain

$$\frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \approx_d N(0,1)$$

Here we can calculate the probabilities, but the statistic still depends on other unknown parameters than μ, namely $\sigma_{\widehat{\theta}}$. But it can be replaced by an estimator, say $\hat{\sigma}_{\widehat{\theta}}$. If this estimator is consistent, then Slutsky's theorem tells us that the asymptotic distribution does not change, so that we also have

$$\frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}} \approx_d N(0,1)$$

This is a pivotal statistic for μ, because the statistic only depends on the unknown parameter μ, everything else is known or can be calculated, and the distribution of the statistic is entirely known (standard normal), so that corresponding probabilities can be obtained from the tables or a statistical software.

So far, this is exactly the same starting point as for the construction of confidence intervals. For testing a null hypothesis, we now make a crucial step: We assume that H_0 is correct, which fixes the value of θ at θ_0, such that we can calculate the test statistic and know its distribution if H_0 is indeed correct. That is, our test statistic is

$$Z = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} \approx_d N(0,1) \quad \text{if } H_0 \text{ is correct.} \tag{5.2}$$

In practice, we obtain a certain value, or realization, of our test statistic, i.e. a number. The question to ask, then, is the following: is the obtained value compatible with the distribution of Z under H_0, i.e. a standard normal distribution, or would it rather be compatible with a distribution under the alternative distribution? To answer this question, we have to consider what happens with Z if H_a is true: then, Z is not centered correctly because θ_0 does not correspond to the true value of θ, which under H_a is larger than θ_0. Thus, the numerator of Z does not have an expectation of zero:

$$\mathbb{E}[\widehat{\theta} - \theta_0] = \mathbb{E}[\widehat{\theta}] - \theta_0 = \theta - \theta_0 > 0$$

so that Z will take positive values on average, if H_a is true. Our decision rule will be the following: If Z is large and positive, then it is more likely that is has been produced by a population whose parameter θ is indeed larger than θ_0, i.e. H_a. We will therefore reject H_0 using a rejection region whose form is

$$RR = \{z \in \mathbb{R} \mid z > k\},$$

that is, all values of Z that are larger than some threshold k that still has to be determined as a function of our tolerance with respect to a type I error.

In particular, to determine the threshold k, suppose that we want to fix the type I error probability at some small α, close to zero (e.g. 0.01 or 0.05). Then,

$$P[Z > k|H_0] = \alpha$$

But this equation just means that k must be the quantile of the standard normal distribution corresponding to a probability level of α, so in our notation,

$$k = z_\alpha.$$

So the procedure is very simple: Prior to the test, fix a type I error probability such as 1% or 5%, and obtain the corresponding critical value, i.e. z_α. Then, calculate the test statistic Z using the data in the sample, and compare it with z_α. If $Z > z_\alpha$, then reject H_0, otherwise do not reject.

Now consider the other possible unilateral test, where the alternative hypothesis is "<" rather than ">", so the hypotheses are now:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_a : \theta < \theta_0. \end{cases}$$

The procedure is similar to the previous case, and the test statistic is the same, given by (5.2). But now, if the alternative hypothesis is true, Z will take too small values on average to be compatible with a standard normal distribution, because its expectation is negative. The centering using θ_0 is not correct, because under H_a, $\theta < \theta_0$. Thus, we have subtracted a value that is too large, and Z will tend to be negative on average. We will therefore, now, reject the null hypothesis if Z is large and negative, so the rejection region becomes

$$\Rightarrow RR = \{z \in \mathbb{R}|z < k\}.$$

To determine the threshold k, we again fix the type I error probability at some small α, so that

$$P[Z < k \mid H_0] = \alpha$$

But this equation means that

$$k = -z_\alpha,$$

due to the symmetry of the standard normal distribution. Again, the decision rule is very simple: Having fixed α and obtained a critical value $-z_\alpha$, one compares the observed test statistic Z with the critical value: If $Z < -z_\alpha$, then reject H_0, otherwise do not reject.

Finally, we have the third possibility of a bilateral test, i.e., now the hypotheses are given by

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_a : \theta \neq \theta_0. \end{cases}$$

Again, the same test statistic (5.2) can be used, because the null hypothesis has not changed. The test is called bilateral because the alternative hypothesis is composed of two parts: one where $\theta > \theta_0$, and one where $\theta < \theta_0$.

Now, if the alternative hypothesis is true, Z will be either "too large" or "too small" to be compatible with the presumed standard normal distribution under H_0. So the rejection region is now composed of two sub-regions:

$$RR = \{z \in \mathbb{R} | z \notin [k_1, k_2]\}.$$

where k_1 and k_2 are two thresholds. Again, by fixing the type I error probability at some small value α, we have the equation

$$P[Z \notin [k_1, k_2] | H_0] = \alpha$$

which leads for the thresholds to $k_1 = -z_{\alpha/2}$ et $k_2 = z_{\alpha/2}$. So in other words, under H_0, there is an $\alpha/2$ probability that $Z < -z_{\alpha/2}$ and another $\alpha/2$ probability that $Z > z_{\alpha/2}$. The overall rejection probability, assuming that H_0 is true, is therefore $\alpha/2 + \alpha/2 = \alpha$. We have simply distributed equally the desired α probability to the left hand and right hand side rejection sub-regions.

The decision rule for the bilateral test is equally simple: Calculate the Z statistic for the observed sample data, then compare it with the rejection region. If Z is inside one of the two sub-regions, then we reject H_0, if it is not, then we do not reject.

5.3 *The power of a test*

An important criterion for the quality of a test is its so-called power. The power of a test about a parameter θ using a test statistic Z and a rejection region RR is defined as

$$\text{power}(\theta) = P[Z \in RR | \theta] = P[RH_0 | \theta].$$

So the power is the probability to reject the null hypothesis. It is the probability of taking the right decision in the case the null hypothesis is incorrect, i.e. $\theta \neq \theta_0$. If, however, the null hypothesis is correct, i.e. $\theta = \theta_0$, then we find

$$\text{power}(\theta_0) = \alpha = \text{level of the test}.$$

If, on the other hand, the null hypothesis is incorrect, i.e. $\theta = \theta_a \in H_a$,

$$\text{power}(\theta_a) = P[RH_0 | \theta_a] = 1 - P[\overline{R}H_0 | \theta_a] = 1 - \beta(\theta_a).$$

In other words, the power of a test is equal to 1 minus the probability of committing a type II error, or $1-\beta$. Note that the power is a function of θ and therefore not constant, unlike the type I error probability that is fixed at α.

To see how the power depends on the distance between $\theta_a \in H_a$ and θ_0, we represent it explicitly as a function of this distance, as shown in the following. In the case of a unilateral test with a ">" alternative, the rejection region is given by $RR = \{z \in \mathbb{R} | z > k\}$, which gives

$$\begin{aligned}
\text{power}(\theta_a) &= P[Z > k | \theta = \theta_a] = P\left[\frac{\widehat{\theta}-\theta_0}{\widehat{\sigma}_{\widehat{\theta}}} > k \middle| \theta = \theta_a\right] \\
&= P[\widehat{\theta} > \theta_0 + k\widehat{\sigma}_{\widehat{\theta}} | \theta = \theta_a] \\
&= P\left[\frac{\widehat{\theta}-\theta_a}{\widehat{\sigma}_{\widehat{\theta}}} > \frac{\theta_0 + k\widehat{\sigma}_{\widehat{\theta}} - \theta_a}{\widehat{\sigma}_{\widehat{\theta}}} \middle| \theta = \theta_a\right] \\
&\approx P\left[N(0,1) > k + \frac{\theta_0-\theta_a}{\widehat{\sigma}_{\widehat{\theta}}}\right]
\end{aligned}$$

Now, the second term on the right hand side of the inequality, $\frac{\theta_0-\theta_a}{\widehat{\sigma}_{\widehat{\theta}}}$, is negative because, under the ">" alternative, $\theta_a > \theta_0$. We now want to investigate how the power depends on the distance between θ_0 and θ_a. To see that, consider what happens if we let θ_a decrease, i.e. θ_a moves away from θ_0. In that case, $k + \frac{\theta_0-\theta_a}{\widehat{\sigma}_{\widehat{\theta}}}$ decreases and the probability that a standard normal r.v. is larger than this value increases. This means that the power of a test increases as the distance between the null and alternative hypotheses increases. This makes intuitive sense: If the two hypotheses are very different, then it should be easy, with high probability, to reject a wrong null hypothesis. If, however, the distance is small, i.e. θ_0 is close to θ_a, then the hypotheses are difficult to distinguish and the power of the test is low.

We can depict the power of a test as a function of θ, which is often simply called the power function. An ideal power function takes the value α at the null hypothesis, $\theta = \theta_0$, and then increases steeply towards 1. If several potential tests are available, one might compare the different power functions and choose the test whose power function is superior uniformly to the power functions of other tests.

5.4 *Link between hypothesis tests and confidence intervals*

There is an intricate link between hypothesis tests and confidence intervals. In fact, confidence intervals can be used for testing in a very simple way. Suppose that the set of hypotheses is given by

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_a : \theta \neq \theta_0. \end{cases}$$

Our decision rule is to reject H_0 if

$$\frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\widehat{\theta}}} \notin [-z_{\alpha/2}, z_{\alpha/2}]$$

But obviously this is equivalent to the following condition:

$$\theta_0 \notin [\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\ ;\ \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}]$$

But $[\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\ ;\ \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}]$ is just our confidence interval for θ at confidence level $1 - \alpha$. So this condition means that θ_0, the value of θ presumed under the null hypothesis, is not contained in the confidence interval for θ at confidence level $1 - \alpha$. Hence, rejecting H_0 is equivalent to the fact that the value under H_0 is not contained in the confidence interval, formally:

$$RH_0 \Leftrightarrow \theta_0 \notin \text{ Confidence interval for } \theta$$

We also see that the level of the hypothesis test, α, is directly related to the level of the confidence interval, $1 - \alpha$. It is the same α. So, for example, if we set the type I error probability of our test to $\alpha = 0.05$, then we reject H_0 if the parameter under H_0 is not contained in a 95% confidence interval for θ.

5.5 The p-value

An often used tool in hypothesis tests is the so-called p-value. It is defined as the smallest test level α for which the data lead to a rejection of H_a. So, for example, if one rejects at a level $\alpha = 0.05$, one may ask, what if I had chosen a smaller α, e.g. $\alpha = 0.01$, would I still reject H_0? Knowing that in that case, the rejection region shrinks (because the critical value, $z_{\alpha/2}$ in the bilateral case, increases), it may actually be that we do not reject for a level of $\alpha = 0.01$. So the p-value is that value of α for which we are just at the boundary of our decision to reject. For example, if the p-value is 0.03, it means that we reject for all levels of α that are larger or equal to 0.03, but we do not reject for all levels α that are smaller than 0.03.

The decision rule of our test can therefore also formulated in terms of the p-value:

$$RH_0 \text{ if } p\text{-value is } < \alpha.$$

In a sense, the p-value is a degree of credibility of the null hypothesis. the smaller it is, the less credible is H_0, and the stronger we are led to reject it. If the p-value is tiny, for example 10^{-8}, then we will reject H_0 without hesitation because this is below every usual value of α. If however the p-value is moderately close to zero, such as 0.03, the decision crucially depends on our choice of α.

The p-value gives us more information than just the fact that a test statistic is beyond a critical value or not. It delivers not a zero-one information, but a value between zero and one, and hence a continuum of information. For example, if our α is fixed at 0.05, it does not matter for the decision whether the p-value is 0.049 or 10^{-8}, in both cases we reject H_0. However, we would be much more confident in our decision in the latter case than in the former.

In the example of Mr Jones, what would be the p-value of the test? To recall, our system of hypotheses was

$$\begin{cases} H_0 : p = 0.5 \\ H_a : p < 0.5 \end{cases}$$

To calculate the p-value, suppose that 3 people among 15 had expressed to vote for Mr Jones. Because the p-value is the smallest value of α for which we still reject H_0, the rejection region corresponding to that α at the margin must be bounded by the observed test statistic, i.e. $Y = 3$ in our case. That means the p-value is given by

$$P[Y \leq 3|H_0] = \sum_{y=0}^{3} \binom{15}{y} (0.5)^{15} = 0.018.$$

If one accepts a type I error probability of $\alpha \geq 0.018$, then RH_0, but if one requires $\alpha < 0.018$, then $\overline{R}H_0$.

5.6 Accept H_0 versus $\overline{R}H_0$

The risk we take by accepting H_0 when the test suggests $\overline{R}H_0$ is measured by $\beta = P[\overline{R}H_0|H_a]$. However, β is not controlled, unlike α which is typically fixed at a small value. As we have seen, β depends on the particular value of the parameter under the alternative hypothesis. Moreover, β can be quite high even in straightforward classical situations.

This means that if the data suggest that $\overline{R}H_0$, then the data do not provide sufficient evidence against H_0, but it does not imply that we take H_0 for granted. It could simply be a consequence of a high type II error probability (β), or equivalently, a low power of the test (i.e., $1 - \beta$).

5.7 Specification of H_0

Recall that in the example of Mr Jones, we have replaced the initial system of hypotheses

$$\begin{cases} H_0 : p \geq 0.5 \\ H_a : p < 0.5 \end{cases} \text{by} \begin{cases} H_0 : p = 0.5 \\ H_a : p < 0.5 \end{cases}$$

Why have we done that? Because in this way, we can control for the type I error probability, α, that we want to keep small. The reason is that, for our rejection region $RR = \{0, 1, \ldots, k\}$, and the original null hypothesis $H_0 : p \geq$

0.5, α would be bounded by the probability of rejecting H_0, given p at the boundary value $p = 0.5$. That is,

$$\max_{p \geq 0.5} P[Y \leq k|p] = P[Y \leq k|p = 0.5]$$

In other words, by replacing the "$\geq$" hypothesis by "=", we ensure that $\alpha = P[RH_0|H_0] \leq 0.05$ for every $p \in H_0$. This is what is meant by "controlling for α". We replace the possible values of p under H_0 by that value that is closest to the alternative hypothesis, because at that point, the probability of erroneously rejecting H_0 is maximal.

In general, testing some parameter θ, we replace the hypotheses

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_a : \theta < \theta_0 \end{cases} \text{by} \begin{cases} H_0 : \theta = \theta_0 \\ H_a : \theta < \theta_0 \end{cases}$$

to control for the type I error probability, α, and a similar adjustment will be made for $H_0 : \theta \leq \theta_0$. Weak inequalities are replaced by equalities at the point of H_0 which has the highest type I error probability.

5.8 Hypothesis tests for the variance

We have seen examples of parameters of interest θ, such as the mean, a proportion, a difference between two means, etc. Another example may be the variance σ^2 of a random variable Y. Because it represents risk in particular if Y is related to some financial variable, one may want to do inference about the variance directly. We will see that this is just a special case of the general procedure above.

Let $Y_1, Y_2, \ldots, Y_n$ be an *i.i.d.* sample of Y with mean μ and variance σ^2, both unknown. We would like to test

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_a : \sigma^2 > \sigma_0^2. \end{cases}$$

Just as for confidence intervals for σ^2, the starting point is the asymptotic distribution of the estimator S^2 (by the CLT):

$$\sqrt{n}\frac{S^2 - \sigma^2}{\sqrt{(K-1)\sigma^4}} \rightarrow_d N(0,1)$$

where K is the kurtosis of Y. By Slutsky's theorem, this result still holds when replacing the unknown kurtosis by a consistent estimator, i.e.

$$\sqrt{n}\frac{S^2 - \sigma^2}{\sqrt{(\widehat{K}-1)\sigma^4}} \rightarrow_d N(0,1)$$

This is a pivotal statistic, because it only depends on the unknown parameter of interest, σ^2, and its asymptotic distribution is fully known.

The test statistic now replaces the unknown σ^2 by the value under H_0, i.e.

$$Z = \sqrt{n}\frac{S^2 - \sigma_0^2}{\sqrt{(\hat{K}-1)\sigma_0^4}} \approx_d N(0,1) \text{ under } H_0$$

where the approximate distribution holds in large samples. For a given sample, everything in Z is known or can be calculated. If H_a is true, then Z will tend to be "too large" to be compatible with H_0, which leads to a rejection of H_0. Hence, the rejection region is given by $RR = \{z \mid z > z_\alpha\}$, where z_α is the complementary α-quantile of Z, i.e. $P(Z > z_\alpha) = \alpha$.

5.9 Exercises

5.1: Suppose that $X \sim \mathcal{N}(\mu, 1)$ and $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$. The test statistic is $T = \sqrt{n}(\bar{X} - \mu_0)$, and we reject H_0 at level α if $T > z_\alpha$. Show that the type II error probability is given by $\beta(\mu) = \Phi(z_\alpha - \sqrt{n}(\mu - \mu_0))$, where $\Phi(\cdot)$ is the cdf of a standard normal distribution. What is the corresponding power function? Make a graph of these functions, and explain the effects of changing the sample size n on these functions.

5.2: In the previous exercise, what is the value taken by $\beta(\mu)$ for the case $\mu = \mu_0$? Explain.

5.3: In the first exercise, suppose that for the case $\alpha = 0.05$ and $\mu - \mu_0 = 1$ it is required to have a power of 0.9. Calculate the necessary sample size.

6

Analysis of categorical data

Many empirical studies involve categorical data, i.e. an observation consists of an association with a distinct category among a finite set of categories. For each category one obtains the number of observations that fall into this particular category. Our task will be to understand the probabilistic mechanism that drives the number of observations per category. In the case of only two possible categories, one speaks of a binomial scheme, in the case of more than two categories, of a multinomial scheme.

Let us consider an example for a categorical variable.

Example 6.1: A marketing campaign gives as a result the number of persons

- having followed the campaign and bought the product
- having followed the campaign but not bought the product
- not having followed the campaign and bought the product
- not having followed the campaign and not bought the product

If there are observations for each one of the four categories, the company might investigate the question whether the campaign was a success. Note that, in this particular example, one might also view the categories as a crossing between two binomial variables: follow the campaign (yes/no), and buy the product (yes/no). So instead of having a multinomial scheme with four categories, one would have a crossing of two binomial variables.

A multinomial scheme is a random experiment consisting of n repetitions of an elementary experiment, such as throwing a dice. The result of the elementary experiment is the observation of exactly one category among a set of k categories (or cells), where k is a finite integer. The probability that an observation falls into the i-th cell is denoted by p_i, $i = 1, 2, ..., k$, and this probability is assumed to be identical for all n repetitions of the experiment. We also assume mutual independence of the individual experiments. We are interested in the number of observations n_i that fall into the i-th cell, $i = 1, 2, ..., k$. Of course, we must have that $n_1 + n_2 + ... + n_k = n$. Note that if $k = 2$, then we obtain a special case the classical binomial scheme.

This chapter investigates two types of questions:

1. How can we test a probabilistic model for the distribution of individual observations on the different categories? This question will lead to the chi-square goodness-of-fit test for a multinomial scheme.

2. How can we test for the independence of two categorical variables? This question will lead to the chi-square test of independence for a contingency table.

We discuss these two concepts in the following. Before that we introduce a new distribution that will turn out to be necessary.

Definition 6.1: Suppose $X_1, X_2, \ldots, X_\nu, \nu \geq 1$, are independent standard normal random variables. Then Y defined as $Y = \sum_{i=1}^{\nu} X_i^2$ follows a chi-square, or χ^2-distribution with parameter ν.

The simplest case is of course $\nu = 1$, in which case Y is just a square of one standard normal random variable. Because Y is the sum of non-negative random variables, its support is only on the positive real line. The first moments of Y are $\mathbb{E}(Y) = \nu$, and $\mathbb{V}\mathrm{ar}(Y) = 2\nu$. Moreover, the χ^2-distribution is asymmetric, i.e. its skewness is different from zero. We can imagine, however, that the distribution becomes more symmetric as ν increases, because the CLT becomes effective, as Y is the sum of independent and identically distributed random variables. This is indeed what happens.

6.1 Chi-square goodness-of-fit test

The expected number of observations in the i-th category is

$$\mathbb{E}[N_i] = np_i \quad , \quad i = 1, 2, ..., k.$$

Suppose that we have a "model" for the k probabilities $p_1, p_2, ..., p_k$ and want to know if this model is reasonable, i.e. confirmed by the data, or whether the data contradicts our model. For the decision, it is natural to consider the differences between the observed numbers n_i and the expected numbers of observations, np_i, i.e.

$$n_i - np_i \quad , \quad i = 1, 2, ..., k.$$

If the model probabilities are reasonable, then these differences should be small in large samples. Of course, the sign of the difference $n_i - np_i$ is not important, what matters is the size, so we should take the absolute value, or better, the squared differences, $(n_i - np_i)^2$, as a criterion, and aggregate these over the different categories. As usual in hypothesis tests, we want to construct a test statistic, based on these aggregated squared differences, that follows some known distribution under the null hypothesis of the model probabilities being true. It turns out that this can be achieved by the following test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{\{n_i - E[N_i]\}^2}{E[N_i]} = \sum_{i=1}^{k} \frac{\{n_i - np_i\}^2}{np_i}$$

So to obtain a pivotal statistic it is important to standardize the squared differences in the numerator by the expected number of observations in the denominator. This will then indeed give a pivotal test statistic whose distribution under the null hypothesis is fully known. It is however not a standard normal distribution, as was the case prior to this chapter, which obviously would not be possible here because the statistic X^2 can only take positive values as it is essentially the sum of squared differences. So it must be another distribution, and it is in fact the chi-square or χ^2-distribution.

If the probabilities p_i are indeed correct, and if n is sufficiently large, then the test statistic X^2 follows approximately a χ^2-distribution with parameter $k-1$. In order to have a good approximation, the cell-wise observations n_i must not be too small. One generally considers that the expected number of observations for each category should at least be 5, i.e. $np_i \geq 5$ for all i. If the statistic X^2 takes large values, then one rejects the hypothesis that the probability model is correct, and one has to look for some other model.

Formally , we want to test the hypotheses

$$\begin{cases} H_0 : p_i = p_i^{(0)}, \ \ i = 1, 2, ..., k, \\ H_a : \exists i \text{ such that } p_i \neq p_i^{(0)} \end{cases}$$

where the $p_i^{(0)}$ are known probabilities, which is the probability model to be tested. We use the test statistic

$$X^2 = \sum_{i=1}^{k} \frac{\{n_i - np_i^{(0)}\}^2}{np_i^{(0)}}$$

As usual for one-sided tests, we reject H_0 if the observed value of the statistic X^2 exceeds the $1-\alpha$ quantile (or equivalently the α complementary quantile) of the χ^2 distribution with parameter $k-1$. Why is the parameter $k-1$ and not k? Because, intuitively, of the k probabilities, only $k-1$ can be chosen freely, the last one is determined by the restriction that the probabilities must sum to one. Therefore, one often speaks of "degrees of freedom", and the parameter of the χ^2 distribution is also often called the degrees of freedom parameter.

Example 6.2: Suppose that the number of weekly accidents at a cross-roads has been observed during $n = 50$ weeks. The following table shows the obtained data:

number of accidents	number of weeks
0	32
1	12
2	6
≥ 3	0

Suppose that the observations are independent. One wishes to know if the Poisson distribution describes well the data. The random variable Y is here the number of weekly accidents at this cross-roads. If $H_0 : Y \sim$ Poisson is true, then we know that

$$\begin{aligned} p_1 &= P[Y = 0] = \exp(-\lambda) \\ p_2 &= P[Y = 1] = \lambda \exp(-\lambda) \\ p_3 &= P[Y \geq 2] = 1 - \exp(-\lambda) - \lambda \exp(-\lambda) \end{aligned}$$

Now the parameter λ is unknown, but we know from the last chapter that an estimator by the method of moments, and the same by maximum likelihood, is given by the sample mean of the observations of Y. We therefore estimate λ by

$$\hat{\lambda} = \frac{12 \times 1 + 6 \times 2}{50} = \frac{24}{50} = 0,48$$

and this gives us for the probabilities:

$$\hat{p}_1 = 0.619, \quad \hat{p}_2 = 0.297 \quad \text{and} \quad \hat{p}_3 = 0.084.$$

With these probabilities we can calculate the expected number of accidents:

$$n\hat{p}_1 = 30.95, \quad n\hat{p}_2 = 14.85 \quad \text{and} \quad n\hat{p}_3 = 4.20.$$

The value of the test statistic is then obtained as

$$X^2_{obs} = \frac{(32 - 30.95)^2}{30.95} + \frac{(12 - 14.85)^2}{14.85} + \frac{(6 - 4.20)^2}{4.20} = 1.354$$

Under H_0, this is the realization of a χ^2 random variable with 1 degree of freedom. In this special case we lose another degree of freedom for the estimation of the parameter λ, so that the degrees of freedom is not $k - 1 = 2$, but $k - 2 = 1$. The 95% quantile of the χ^2 distribution with 1 degree of freedom is 3.841, which is the critical value of the test. Since the observed test statistic is not beyond the critical value, we do not reject the null hyothesis. The data confirm our hypothesis of a Poisson distribution for the number of weekly accidents.

In this example, the probability model was the Poisson distribution, but of course any other model could be tested, for example the simple model of equal probabilities for all categories, a binomial distribution, or any other discrete probability law.

6.2 *Contingency tables*

We now suppose to have two categorical variables that are crossed, so that for each individual we observe the association to a combination of categories of both variables. For example, for a sample of individuals one observes the gender (first categorical variable) and the opinion concerning the vaccination policy of the government (second categorical variable). A natural question, then, would be the independence of the two categorical variables. In our example, if the independence hypothesis is rejected by the data, then the opinion concerning the vaccination policy would be different between men and women. If it is not rejected, then one could conclude that the opinion is independent of gender (but be aware of the type II error).

We will introduce the independence test with an example. Suppose we want classify the defaults of produced machines according to the type of default (with four categories: A,B,C and D), and according to the production line (with three categories: 1, 2 and 3). In total, $n = 309$ defaults have been registered, and the data are summarized in the following table:

Ligne	Type de défaut				
de production	A	B	C	D	Total
1	15 (22,51)	21 (20,99)	45 (38,94)	13 (11,56)	94
2	26 (22,99)	31 (21,44)	34 (39,77)	5 (11,81)	96
3	33 (28,50)	17 (26,57)	49 (49,29)	20 (14,63)	119
Total	74	69	128	38	309

We want to test the hypothesis H_0 that the default type is independent of the production line. Let p_A (resp. p_B, p_C, p_D) be the probability that the default is of type A (resp. B, C, D). Of course, we have the restriction $p_A + p_B + p_C + p_D = 1$. Similarly, let p_1 (resp. p_2, p_3) be the probability that a default occurs in production line 1 (resp. 2, 3), and $p_1 + p_2 + p_3 = 1$.

If the two categorical variables are independent, then the probability that a default is of type A and in production line 1 is

$$p_A \times p_1$$

.

Note that the probabilities $p_A, p_B, p_C, p_D, p_1, p_2$ and p_3 are not specified. The null hypothesis requires that the cell probabilities are obtained as a product of the two corresponding marginal factors, i.e. the row and the column factor.

Let n_{ij} be the number of observations in cell (i, j), situated at the intersection of line i and row $j, i = 1, 2, ..., r, j = 1, 2, ..., c$. Similarly, let p_{ij} be the probability that an observation falls in this cell, which is estimated by

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

.

Let r_i (resp. c_j) be the number of observations in line i (resp. column j). The probability that an observation falls in the i-th line (resp. j-th column) is r_i/n (resp. c_j/n). Under H_0, the expected number of observations in the cell (i, j) is

$$n \times \frac{r_i}{n} \times \frac{c_j}{n} = \frac{r_i c_j}{n};$$

These numbers are noted in parentheses underneath the observed numbers for each cell of the table. For example, for the cell $(1, 1)$, the expected number of observations under H_0 is

$$\frac{r_1 c_1}{n} = \frac{94 \times 74}{309} = 22{,}51.$$

The test statistic is therefore

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\{n_{ij} - \frac{r_i c_j}{n}\}^2}{\frac{r_i c_j}{n}}.$$

Under the null hypothesis of independence, X^2 follows a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom. We will discuss below why this is the degrees of freedom here. The test is a one-sided test, i.e. we reject for large values of X^2. What is "large" is determined by a critical value corresponding to the $1-\alpha$ quantile of the χ^2 distribution. In our example, we obtain

$$\frac{(15-22{,}51)^2}{22{,}51} + \frac{(26-22{,}99)^2}{22{,}99} + ... + \frac{(20-14{,}63)^2}{14{,}63} = 19{,}17$$

and the critical value at level $\alpha = 0.05$ is $12{,}592$. The test statistic is larger than the critical value, therefore we reject the null hypothesis of independence. The data give us evidence that the type of default depends on the production line.

In general, a contingency table has the following form:

Row	Column factor					
factor	1	...	j	...	c	Total
1	n_{11} $(\frac{r_1 c_1}{n})$	...	n_{1j} $(\frac{r_1 c_j}{n})$	...	n_{1c} $(\frac{r_1 c_c}{n})$	r_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1} $(\frac{r_i c_1}{n})$	...	n_{ij} $(\frac{r_i c_j}{n})$	...	n_{ic} $(\frac{r_i c_c}{n})$	r_i
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1} $(\frac{r_r c_1}{n})$	...	n_{rj} $(\frac{r_r c_j}{n})$	...	n_{rc} $(\frac{r_r c_c}{n})$	r_r
Total	c_1	...	c_j	...	c_c	n

Now we come back to the parameter of the χ^2 distribution, also called degrees of freedom. Why is it equal to $(r-1)(c-1)$? This is equal to the total number of cells, i.e. $r \times c$, minus the number of linear constraints. There is one constraint because

$$n_{11} + n_{12} + \ldots + n_{rc} = n.$$

But for each parameter estimator we lose another degree of freedom, in our case for each estimated marginal probability. For the row probabilities we have $r - 1$ parameters to estimate (there are r probabilities, but the last one is determined by the restriction that the probabilities sum to one, so there are only $r - 1$ free parameters). Similarly, there are $c - 1$ column probabilities to estimate. This makes a loss of $(r - 1) + (c - 1)$ degrees of freedom. In total, we therefore have

$$rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1)$$

as the degrees of freedom of the χ^2 independence test.

Example 6.3: In order to study the effectiveness of a new vaccine against the influenza virus, the vaccination is offered for free to a community. Certain members of the community do not wish to be vaccinated, others only receive one injection, and the remaining members receive two injections. The following spring, the pharmaceutical company asks 1,000 members of the community and ontains the following results:

	No vaccine	1 injection	2 injections	Total
Infection	24	9	13	46
	(14.4)	(5.0)	(26.6)	
No infection	289	100	565	954
	(298.6)	(104.0)	(551.4)	
Total	313	109	578	1,000

One wishes to detect whether there is an impact of the vaccine on the development of the virus, i.e. to test the hypothesis H_0 that the infections are independent of the vaccine. In other words, whether the row factor, infection, is independent of the column factor, vaccine. To this end, one calculates the observed test statistic X^2:

$$\begin{aligned} X^2_{obs} &= \frac{(24 - 14.4)^2}{14.4} + \frac{(289 - 298.6)^2}{298.6} + \ldots + \frac{(565 - 551.4)^2}{551.4} \\ &= 17.35 \end{aligned}$$

which is beyond the critical value $\chi^2_{0.05} = 5.991$ of the χ^2 distribution with $(r - 1)(c - 1) = 2$ degrees of freedom. Hence, we reject H_0. The vaccine has an impact on the infections, or both categorical variables can be rejected to be independent, at a level of $\alpha = 5\%$.

To summarize this chapter, we have seen two types of questions concerning categorical variables: First, goodness-of-fit tests for a probability model that explains the frequency of observations per category. And second, the question of independence of two categorical variables. This is of course far from exhaustive, and categorical variables can play important roles in other contexts as well. They can also serve as explanatory variables in a regression model, see later in Chapter 7.

6.3 Exercises

6.1: For a multinomial experiment with $k = 2$ (i.e. a binomial scheme), show that the statistic

$$X^2 = \sum_{i=1}^{2} \frac{(n_i - np_i)^2}{np_i}$$

converges to a χ^2 distribution with 1 degree of freedom as $n \to \infty$, where n is the number of repetitions of the experiment, n_i the number of observations in the i-th category, and p_i is the probability of an observation falling into category i.

6.2: Suppose that the entries in a contingency table that appear in row i and column j are denoted n_{ij},for $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$; that the row and column totals are denoted r_i,for $i = 1, 2, \cdots, r$, and c_j, for $j = 1, 2, \cdots, c$; and that the total sample size is n.

(a) Show that

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left[n_{ij} - \widehat{E(n_{ij})}\right]^2}{\widehat{E(n_{ij})}} = n \left(\sum_{j=1}^{c} \sum_{i=1}^{r} \frac{n_{ij}^2}{r_i c_j} - 1 \right).$$

Notice that this formula provides a computationally more efficient way to compute the value of X^2.

(b) Using the preceding formula, what happens tot the value of X^2 if every entry in the contigency table is multiplied by the same interger constant $k > 0$?

7
Linear regression

Linear regression models attempt to explain the expectation of a random variable Y by another variable X, which typically is random as well. Take, for example, the household expenditures explained by the household revenues. For various economic reasons, one might be interested in how much households increase their expenditures when their revenues increase by a given amount. In general, the variable Y is the variable of interest, also called the response variable, and X is called the explanatory variable. Although this imposes some direction going from X to Y, this direction does by no means imply a causality, it merely represents an association, or dependence, of Y on X. The starting point is the conditional distribution of Y given X, $f(Y|X)$. If Y depends on X, then this conditional distribution will be different from the marginal distribution, i.e. $f(Y|X) \neq f(Y)$. Therefore, Y is no longer i.i.d., because it is not identically distributed.

Rather than looking at the entire conditional distribution $f(Y|X)$, we instead concentrate on the conditional expectation of Y given X, because that will be the important quantity to be used for prediction, for example. So the question is, given a particular value of X, what would be our best predictor of Y? In a mean square error sense, the optimal predictor is the conditional expectation, so this is our main object of interest: $\mathbb{E}[Y|X]$.

If Y and X are not independent, then the conditional expectation $\mathbb{E}[Y|X]$ is a function of X. It is the purpose of regression models to specify this functional relationship. In general, this function can take various forms. Linear regression models specify a linear function for the conditional expectation:

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X,$$

where β_0, β_1 are unknown coefficients. The coefficient β_0 is called the intercept term, because it measures the conditional expectation at $X = 0$. The coefficient β_1 is called the slope parameter, because it represents the slope of the regression line $\beta_0 + \beta_1 x$. It is often of particular interest, because it is the marginal effect of a small change in X on Y. So for example, if $\beta_1 = 0.8$, and one increases household revenues X by 1 euro, then one would expect households to spend 0.8 euro of that additional euro.

Of course the conditional distribution will not be concentrated at the conditional expectation, but there will be some dispersion. Let us call the difference between the actual Y and the conditional expectation ε:

$$\varepsilon := Y - \mathbb{E}[Y|X],$$

This error term is random and has certain properties that we still need to assume. Having defined the error term, we can write the linear regression model in the equivalent form

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

The systematic part of the model that is explained by X is $\beta_0 + \beta_1 X$, while the unsystematic, or noise part of the model is given by the error term ε. Note that the error term, as a consequence of its definition, is centered at zero, because $\mathbb{E}[\varepsilon|X] = 0$. It can be i.i.d., but this is not necessary. We will however assume in the following that is has a constant variance: $\mathbb{V}\text{ar}(\varepsilon|X) = \sigma^2$, where σ^2 is a positive constant. Note in particular that the conditional variance is not a function of X, unlike the conditional expectation. This can be relaxed, but for simplicity we work with this assumption in the following.

In order to work with the model, use it for predictions, do inference about the unknown parameters, etc., we first need to obtain estimators of these parameters. The main parameters of interest are β_0 and β_1, but also the variance of the error term, σ^2 needs to be estimated for inference purposes, as we will see.

7.1 *Estimation of the model*

The objective is to obtain estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the unknown parameters β_0 and β_1. This will give us, then, also an estimator of the conditional expectation $\mathbb{E}[Y|X]$ as

$$\widehat{Y} = \mathbb{E}[\widehat{Y|X = x}] = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

Some more terminology: We will call $\widehat{Y}$ the fitted values, or predicted values, of Y for given X. And we will call the difference between the observations of Y and the fitted values the residuals, i.e., $e := y - \widehat{\beta}_0 + \widehat{\beta}_1 x$. Note that errors and residuals are two different objects: Errors are unobserved differences between the observations Y and the true conditional expectations, whereas residuals are (observed) differences between observations Y and the fitted regression line. Errors are unobserved, residuals are observed or can be constructed.

We have seen in chapter 4 two estimation methods: the method of moments, and the method of maximum likelihood. For linear regression models, it turns out that a third estimation method is particulary suited: the method of least squares.

The idea of least squares is to minimize the sum of the squared residuals for all observations. For each observation $Y_i, i = 1, \ldots, n$, and for given parameters $\widehat{\beta_0}$ and $\widehat{\beta_1}$, we can calculate the residual $e_i = Y_i - \widehat{\beta}_0 + \widehat{\beta}_1 X_i$. If the regression line fits well the data, then this residual should be small in absolute value. One could minimize the sum of absolute values of the residuals, but that would not lead to simple analytical formulas, so that one prefers to work with the squared residuals, equally a measure for their size.

The sum of squared residuals will be called SSR and is defined as

$$SSR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right)^2$$

based on the observations (x_i, y_i), $i = 1, 2, \ldots, n$.

A first order condition for a minimum is that the estimators satisfy the following conditions:

$$\frac{\partial}{\partial \widehat{\beta}_0} SSR = 0 \text{ et } \frac{\partial}{\partial \widehat{\beta}_1} SSR = 0$$

This delivers the set of equations

$$\begin{aligned} \sum_{i=1}^{n} \underbrace{(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))}_{e_i} &= 0 \\ \sum_{i=1}^{n} \underbrace{(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))}_{e_i} x_i &= 0 \end{aligned}$$

Note that these equations imply that the mean of the residuals is equal to zero, and that

$$\sum_{i=1}^{n} \widehat{Y}_i e_i = 0$$

We now have a set of two equations that we can solve for the parameters to obtain the solution

$$\begin{aligned} \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \end{aligned}$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

and

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

We can find the following equivalent expressions, which is left as an exercise. It depends on the context which one will be the most useful.

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})y_i \\
&= \sum_{i=1}^{n}(y_i - \overline{y})x_i \\
&= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \\
&= \sum_{i=1}^{n} x_i y_i - n\overline{xy}
\end{aligned}$$

And similarly,

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})x_i \\
&= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2
\end{aligned}$$

7.2 *Properties of the least squares estimators*

As in chapter 4 on estimation, we would like to understand the statistical properties of the least squares estimators, in terms of bias, variance, MSE, etc. This is important for two reasons: First, to see whether the estimator is "good", compared perhaps with alternative estimators. Second, to be able to do inference, for example construct confidence intervals or test hypotheses.

If we can show that $\mathbb{E}[\widehat{\beta}_i|\mathbf{X}] = \beta_i$ then the estimator of β_i is unbiased. Here, $\mathbf{X}$ denotes the set of explanatory variables available in the sample, i.e. $\mathbf{X} = \{x_1, \ldots, x_n\}$.

Let us first calculate the bias of $\widehat{\beta}_1$. From the equivalent expressions for the estimator we know that

$$\begin{aligned}
\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{S_{xx}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{S_{xx}}
\end{aligned}$$

from which we get

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_1|\mathbf{X}] &= \frac{\sum_{i=1}^n (x_i-\overline{x})\mathbb{E}[Y_i|x_i]}{S_{xx}} = \frac{\sum_{i=1}^n (x_i-\overline{x})(\beta_0+\beta_1 x_i)}{S_{xx}} \\ &= \beta_0\frac{\sum_{i=1}^n (x_i-\overline{x})}{S_{xx}} + \beta_1\frac{\sum_{i=1}^n (x_i-\overline{x})x_i}{S_{xx}} = \beta_1.\end{aligned}$$

Therefore, $\widehat{\beta}_1$ is unbiased.

The variance of $\widehat{\beta}_1$ can be calculated as follows.

$$\begin{aligned}\text{Var}[\widehat{\beta}_1|\mathbf{X}] &= \text{Var}\left[\frac{\sum_{i=1}^n (x_i-\overline{x})Y_i}{S_{xx}}|\mathbf{X}\right] \\ &= \frac{1}{S_{xx}^2}\sum_{i=1}^n (x_i-\overline{x})^2\underbrace{\text{Var}[Y_i|x_i]}_{\sigma^2} \\ &= \frac{\sigma^2 S_{xx}}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}}.\end{aligned}$$

Note that this variance depends on two things. First, the variance of the error term, σ^2. The larger the error variance, the larger also the variance of our estimator, which makes intuitive sense because there is a higher level of noise in the regression. Second, the variance also depends on the variation of X, measured by S_{xx}. But this is in the denominator, so a higher variation of the observed X values leads to a smaller variance of our estimator. This may seem surprising at first sight, but the intuitive reason is that with a higher variation of X, more information can be gained about the slope of the regression line. Imagine the extreme case of all X observations concentrated close to one point, then the variation S_{xx} will be very small, but it will be very difficult to estimate the slope of the regression line. This explains why the variance is inversely proportional to S_{xx}.

Let us now calculate the bias of $\widehat{\beta}_0$

We examine $\widehat{\beta_0} = \overline{Y} - \widehat{\beta}_1\overline{x}$; and show that $\widehat{\beta}_0$ is unbiased:

$$\mathbb{E}[\widehat{\beta}_0|\mathbf{X}] = \mathbb{E}[\overline{Y}|\mathbf{X}] - \mathbb{E}[\widehat{\beta}_1|\mathbf{X}]\overline{x} = \mathbb{E}[\overline{Y}|\mathbf{X}] - \beta_1\overline{x}$$

But,

$$\begin{aligned}\mathbb{E}[\overline{Y}|\mathbf{X}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^n(\beta_0+\beta_1 x_i+\varepsilon_i|\mathbf{X}] \\ &= \mathbb{E}[\beta_0+\beta_1\overline{x}+\frac{1}{n}\sum_{i=1}^n\varepsilon_i|\mathbf{X}] \\ &= \beta_0+\beta_1\overline{x}\end{aligned}$$

from which we finally obtain

$$\mathbb{E}[\widehat{\beta}_0|\mathbf{X}] = \beta_0+\beta_1\overline{x}-\beta_1\overline{x} = \beta_0.$$

Thus, the estimator of the intercept is also unbiased. We then calculate the variance of $\widehat{\beta}_0$:

$$\operatorname{Var}[\widehat{\beta}_0|\mathbf{X}] = \underbrace{\operatorname{Var}[\overline{Y}|\mathbf{X}]}_{\frac{\sigma^2}{n}} + \overline{x}^2 \underbrace{\operatorname{Var}[\widehat{\beta}_1|\mathbf{X}]}_{\frac{\sigma^2}{S_{xx}}} - 2\overline{x} \underbrace{\operatorname{Cov}[\overline{Y}, \widehat{\beta}_1|\mathbf{X}]}_{0}$$

because

$$\begin{aligned}\operatorname{Cov}[\overline{Y}, \widehat{\beta}_1|\mathbf{X}] &= \operatorname{Cov}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}} Y_i|\mathbf{X}\right] \\ &= \sum_{i=1}^{n} \frac{x_i - \overline{x}}{nS_{xx}} \underbrace{\operatorname{Var}[Y_i|x_i]}_{=\sigma^2} = 0\end{aligned}$$

We obtain

$$\operatorname{Var}[\widehat{\beta}_0|\mathbf{X}] = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}.$$

Finally, because the two estimators are random variables, we can also calculate their covariance:

$$\operatorname{Cov}[\widehat{\beta}_0, \widehat{\beta}_1|\mathbf{X}] = -\frac{\overline{x}\sigma^2}{S_{xx}} \neq 0 \text{ si } \overline{x} \neq 0$$

which implies that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are, in general, dependent.

We have provided estimators of the parameters of interest β and discussed their properties in terms of bias and variance. For inference, it will also be necessary to estimate the error variance $\sigma^2 = \operatorname{Var}[\varepsilon|X]$ which is unknown. We will use the following natural estimator:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n} SSR$$

So the the estimator of the error variance takes the sum of the squared residuals, SSR, and divides it by the number of observations, n. It is like an average of the squared residuals. Note again the difference between residuals and errors: errors are unknown while residuals are obtained from the fitted regression and therefore known. It makes intuitive sense to estimate the unknown error variance by the mean squared residuals, because the mean of the residuals is equal to zero. Similar to S^2, our variance estimator in the case of a single variable Y, $\widehat{\sigma^2}$ is a biased estimator, which we accept here without proof. But this bias tends to zero as $n \to \infty$, so that the estimator is asymptotically unbiased.

7.3 *Parameter inference*

We now want to do inference about the unknown parameters β_0 and β_1, i.e. construct confidence intervals or test hypotheses about these parameters. Let us first define the following constants, for given explanatory variables:

$$c_{00} = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}, \quad c_{11} = \frac{1}{S_{xx}},$$

By the CLT and Slutsky's theorem, we know that

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}\sqrt{c_{ii}}} \to_d N(0,1) \tag{7.1}$$

For a test of β_i, $i = 0,1$, consider the hypothesis that it takes a specific value β_{i0}:

$$\begin{cases} H_0 : \beta_i = \beta_{i0} \\ H_a : \begin{cases} \beta_i > \beta_{i0} \\ \beta_i < \beta_{i0} \\ \beta_i \neq \beta_{i0} \end{cases} \end{cases}$$

So here β_{i0} is a presumed value under H_0, it is known for us, a value of interest. For example, if $\beta_{i0} = 0$, then H_0 says that the term associated with β_{i0} is simply absent from the model.

To test this hypothesis we replace in (7.1) the unknown coefficient β_i by the presumed value under H_0, i.e. β_{i0}. We do as if H_0 was correct. The obtained test statistics is:

$$T = \frac{\widehat{\beta}_i - \beta_{i0}}{\widehat{\sigma}\sqrt{c_{ii}}} \approx_d N(0,1) \text{ sous } H_0$$

It is very important that the test statistic is not dependent on any unknown coefficients, and that it is possible to calculate T from sample information only.

Depending on the type of alternative hypothesis, the rejection region takes the following form:

$$\begin{cases} t > z_\alpha \\ t < -z_\alpha \\ |t| > z_{\alpha/2} \end{cases}$$

A confidence interval for β_i at level $1 - \alpha$ is given by

$$\left[\widehat{\beta}_i \pm z_{\alpha/2}\widehat{\sigma}\sqrt{c_{ii}}\right].$$

Recall the equivalence of confidence intervals and hypothesis tests. If we want to test, for example, $H_0 : \beta_i = 0$, it suffices to see whether 0 is contained in the $1 - \alpha$ confidence interval for β_i. If it does, we do not reject H_0, if it does not, we reject.

7.4 *Confidence interval for* $\mathbb{E}[Y|X = x^*]$

Rather than testing parameters individually, we can also test any linear combination of them. One particular linear combination is given by the regression line, $\beta_0 + \beta_1 x$. We can formulate hypotheses about the regression line, or alternatively construct confidence intervals.

Suppose that we want to estimate $\mathbb{E}[Y|X]$ for a fixed value x^* of the explanatory variable X, and associate a confidence interval. The point estimator is, of course, $\widehat{Y} := \hat{\beta_0} + \hat{\beta_1} x^*$, i.e. we replace the unknown parameters β_i by their point estimators. Note first that this is an unbiased estimator:

$$\begin{aligned}
\mathbb{E}[\widehat{Y}|\mathbf{X}] &= \mathbb{E}[\hat{\beta_0}|\mathbf{X}] + \mathbb{E}[\hat{\beta_1}|\mathbf{X}]x^* \\
&= \beta_0 + \beta_1 x^* \\
&= \mathbb{E}[Y|X = x^*]
\end{aligned}$$

The error margin of $\widehat{Y}$ depends on the variance:

$$\begin{aligned}
& \mathbb{V}\mathrm{ar}(\hat{\beta_0} + \hat{\beta_1} x^* |\mathbf{X}) \\
= \; & \underbrace{\mathbb{V}\mathrm{ar}(\hat{\beta_0}|\mathbf{X})}_{\frac{\sigma^2}{S_{xx}} \frac{\Sigma_{i=1}^n x_i^2}{n}} + (x^*)^2 \underbrace{\mathbb{V}\mathrm{ar}(\hat{\beta_1}|\mathbf{X})}_{\frac{\sigma^2}{S_{xx}}} + 2x^* \underbrace{\mathrm{Cov}[\widehat{\beta}_0, \widehat{\beta}_1|\mathbf{X}]}_{-\overline{x}\frac{\sigma^2}{S_{xx}}} \\
= \; & \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right).
\end{aligned}$$

This variance is minimal at the value x^* that corresponds to the sample mean of X, because at that value, the second term inside the brackets is zero. In other words, if we want to construct a confidence interval at the point of the sample mean of X, the confidence interval for the regression will be smallest. On the other hand, the variance increases as X^* moves away from $\bar{X}$, and therefore the corresponding confidence intervals become wider.

The $1 - \alpha$ confidence interval for $\mathbb{E}[Y|X = x^*] = \beta_0 + \beta_1 x^*$ is given by

$$\left[\widehat{\beta}_0 + \widehat{\beta}_1 x^* \pm z_{\alpha/2} \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \right].$$

Finally, we can also construct prediction intervals for Y. Imagine that, after estimation of the model, we obtain a new observation of X, say X^*, for which we wish to predict Y. Our estimator $\widehat{Y}$ of $\mathbb{E}[Y|X = x^*]$ can be viewed as a predictor of Y at the value $X = x^*$. It has some optimality properties, for example it the predictor that minimizes the mean squared prediction error.

We can construct a prediction interval for the new variable Y, but with two modifications: First, the error margin will be larger than for the confidence interval for $\mathbb{E}[Y|X = x^*]$ because Y additionally depends on the error term ε, which has a positive variance. In other words, the prediction interval for Y will be wider than the confidence interval for $\mathbb{E}[Y|X = x^*]$. Second, we would need to add an assumption about the distribution of the error term. This is because Y depends on ε, and so any prediction interval for Y will depend on the distribution of ε.

7.5 Goodness of fit

After having estimated the linear model, one would like to assess whether the fitted model explains well the data. In other words, the question is how good is the fitted model, or the so-called "goodness-of-fit". A perfect fit would be obtained if all observations lie on a straight line. In that case, knowing X would give a perfect predictor of Y, there is no variability whatsoever around that predictor. If, on the other hand, are very dispersed around the regression line, then the quality of fit is inferior and a prediction would be highly uncertain.

A measure for the variability of Y_i is given by the sum of squares of the differences between the observations Y_i and their sample mean, $\overline{Y}$. We can write

$$\begin{aligned}
\text{Variation of } Y_i &= \sum_{i=1}^{n} \left[Y_i - \overline{Y}\right]^2 \\
&= \sum_{i=1}^{n} \left[(\hat{Y}_i - \overline{Y}) + e_i\right]^2 \\
&= \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} e_i^2 + 2 \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}) e_i}_{0}
\end{aligned}$$

where the third term is zero because $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ and $\sum_{i=1}^{n} e_i = 0$ by the system of equations that led to the least squares estimator.

This decomposition can be interpreted in the following way:

- The sum $\sum_{i=1}^{n} (Y_i - \overline{Y})^2$ measures the total variation of Y_i around its sample mean $\overline{Y}$.

- The sum $\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$ represents the variation of Y_i that is explained by the regression.

- The $\sum_{i=1}^{n} e_i^2$ is the variation of Y_i that is not explained by the regression. It is the residual variation, which has been minimized to obtain the least squares estimators.

This decomposition of the total sum of squares into the explained and the residual sum of squares allows us to introduce a measure for the quality of fit. It is a relative measure, given by the following ratio

$$\begin{aligned}
R^2 &:= \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \\
&= \frac{\text{Variation explained by the fitted model}}{\text{Variation of the sample } Y_i}
\end{aligned}$$

that is called the coefficient of determination (or simply R^2). The R^2 is therefore the proportion of the total variation of Y that is explained by the model. Because it is a proportion of two positive valued quantities, it must be nonnegative and is bounded by one:

$$0 \leqslant R^2 \leqslant 1$$

We have two extreme cases: First, if $\sum_{i=1}^n e_i^2 = 0$, then $R^2 = 1$, which is the ideal situation in which all observations are on a straight line, and the model perfectly explains the variation of Y. Second, if $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 = 0$, then $R^2 = 0$, meaning that the model does not explain anything of the variation of Y.

Note also that, using the decomposition of the total variation of Y, the R^2 can be written equivalently in the following way:

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2 - \sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

So the R^2 can also be viewed as 1 minus the unexplained, or residual variation of the fitted model.

Finally, it is interesting to note that the coefficient of determination, R^2, is equal to the squared correlation coefficient between X and Y, r^2. This explains the used symbol R^2. The proof of this property is left as an exercise.

With these interpretations, we see that the R^2 measures the quality of the model to fit the data. The higher the R^2, the better the fit of the model. Keep in mind, however, that this depends on the data and context of the application. In some situations, an R^2 of 20% might be fully satisfactory, while in others one might be more ambitious and require an R^2 of at least 80%. So there is no general guideline that says that the R^2 should attain a certain minimum value, this very much depends on the context and the data.

7.6 Summary

To summarize this chapter, the idea of a linear regression model is to model the conditional expectation of a response variable Y given an explanatory variable X by a linear function, i.e. a straight line. An estimator of the unknown parameters is given by the least quares estimator, which minimizes the sum of squared residuals, and which is unbiased. Using the expressions for the variances of the estimators, we can construct confidence intervals and do hypotheses tests about the unknown regression coefficients. To assess the quality of fit of the model, we have introduced the R^2 which is the proportion of the variation of Y that is explained by the regression. The higher the R^2, the better the quality of fit.

Finally, note that, of course, the regression can be extended to contain multiple explanatory variables, $X_1, \ldots, X_K$. It is also possible to allow for nonlinear regression functions, i.e. a model where the conditional expectation of Y given X is a nonlinear function of X. These issues are left for future lectures.

7.7 *Exercises*

7.1: For the linear model (without intercept) $Y_i = \beta X_i + \varepsilon_i$, $i = 1, \dots, n$, derive the least squares estimator of the coefficient β.

7.2: A linear model is given by $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \dots, n$. When considering the inverse causal direction, a model could be written as $X_i = \alpha_0 + \alpha_1 Y_i + u_i$, where u_i is another error term. Under what condition are the least squares estimators of β_1 and α_1 identical?

7.3: A log-linear model is defined by $\log Y_i = \beta_0 + \beta_1 \log X_i + \varepsilon_i$, $i = 1, \dots, n$. Show that the coefficient β_1 is equal to the elasticity of Y with respect to X.

7.4: A linear model is given by $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \dots, n$. Show that the least squares residuals, $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$, have a mean of zero.

7.5: A linear model is given by $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \dots, n$, with $Var(\varepsilon_i) = \sigma^2$. Calculate the correlation of the two least squares estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$.

7.6: Show that the coefficient of determination, R^2, is equal to the squared correlation coefficient between X and Y, r^2.

7.7: A linear model is given by $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \dots, n$, with $Var(\varepsilon_i) = \sigma^2$. Show that the statistic used for testing $H_0 : \beta_1 = 0$ can be written in two equivalent ways, i.e.

$$\frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{S_{xx}}} = \frac{r\sqrt{n}}{\sqrt{1-r^2}}$$

where r is the sample correlation coefficient between X and Y. (Hint: Use that $\widehat{\beta}_i = r\sqrt{S_{yy}/S_{xx}}$ and that $r^2 = 1 - (n\widehat{\sigma^2})/S_{yy}$ by the definition of the r^2.) Try to interpret the obtained expression.

8
Solutions to the Exercises

8.1 Solutions Chapter 1

1.1: The kurtosis of a r.v. Y is defined as $K(Y) = \mathbb{E}[Z^4]$, where

$$Z := \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathbb{V}\text{ar}(y)}}$$

By Jensen's inequality, we know that

$$\mathbb{E}[Z^4] \geq \mathbb{E}(Z^2)^2$$

because the square is a convex function. However, $\mathbb{E}(Z^2) = 1$ by construction, because Z is the centered and standardized Y, i.e. mean zero and variance one. Therefore,

$$K(Y) = \mathbb{E}[Z^4] \geq 1.$$

1.2: By the Cauchy-Schwarz inequality, we know that

$$Cov(X,Y)^2 \leq \mathbb{V}\text{ar}(X)\mathbb{V}\text{ar}(Y)$$

and thus, taking the square root on both sides,

$$|Cov(X,Y)| \leq \sqrt{\mathbb{V}\text{ar}(X)\mathbb{V}\text{ar}(Y)}$$

and

$$-1 \leq \frac{Cov(X,Y)}{\sqrt{\mathbb{V}\text{ar}(X)\mathbb{V}\text{ar}(Y)}} \leq 1$$

1.3:

$$\begin{aligned}
\mathbb{E}(\overline{Y}) &= \mathbb{E}\left(\frac{\sum_{i=1}^n Y_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) \text{ (par linéarité de l'espérance)} \\
&= \frac{1}{n} n\mathbb{E}(Y_1) \text{ (car les } Y_i \text{ sont iid)} \\
&= \mathbb{E}(Y_1) = \mu.
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\overline{Y}) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)\\
&= \frac{1}{n^2}\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{n} Y_i\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}(Y_i) \text{ (car les variables sont indépendantes)}\\
&= \frac{1}{n^2} n\mathbb{V}\mathrm{ar}(Y_1) \text{ (car les variables sont identiquement distribuées)}\\
&= \frac{\mathbb{V}\mathrm{ar}(Y_1)}{n} = \frac{\sigma^2}{n}.
\end{aligned}$$

1.4: Le théorème de Chebyshev dit que, pour tout $k > 0$,

$$P(|Z_n - \theta| > k\sigma_n) \leq \frac{1}{k^2}$$

où $\sigma_n^2 := \mathbb{V}\mathrm{ar}(Z_n)$. Pour tout n et $\epsilon > 0$,

$$k = \frac{\epsilon}{\sigma_n}$$

est un nombre positif. Dès lors,

$$P(|Z_n - \theta| > \epsilon) = P(|Z_n - \theta| > \frac{\epsilon}{\sigma_n}\sigma_n) \leq \frac{1}{(\epsilon/\sigma_n)^2} = \frac{\sigma_n^2}{\epsilon^2}.$$

Maintenant on laisse n tendre vers infinie. Si $\lim_{n\to\infty}\sigma_n^2 = 0$, alors

$$\lim_{n\to\infty} P(|Z_n - \theta| > \epsilon) = \lim_{n\to\infty}\frac{\sigma_n^2}{\epsilon^2} = 0$$

ce qui démontre que Z_n tend vers θ en probabilité.

1.5: Par la loi des grands nombres, on sait que : $\frac{S_n}{n} \xrightarrow{p} E(X_i)$. Par ailleurs,

$$E(X_i) = \int_0^1 x\, 2x\, dx = \frac{2 \cdot 1^3}{3} - 0 = \frac{2}{3}.$$

Définissons $g : \mathbb{R} \to \mathbb{R} : x \mapsto \frac{1}{x}$. Puisque $\frac{S_n}{n} \xrightarrow{p} \frac{2}{3}$, on a (par l'indice) que $g\left(\frac{S_n}{n}\right) = \frac{n}{S_n} \xrightarrow{p} g\left(\frac{2}{3}\right) = \frac{3}{2}$.

1.6: On doit calculer $P(|X_{(n)} - \theta| < \epsilon)$ où ϵ est une constante positive et vérifier si cette probabilité tend vers 1 lorsque n tend vers l'infini. Pour ce faire, on a :

$$\begin{aligned}
P(|X_{(n)} - \theta| < \epsilon) &= P(-\epsilon < X_{(n)} - \theta < \epsilon)\\
&= P(\theta - \epsilon < X_{(n)} < \theta + \epsilon)\\
&= P(X_{(n)} < \theta + \epsilon) - P(X_{(n)} < \theta - \epsilon).
\end{aligned}$$

Puisque $X_{(n)}$ ne peut pas excéder θ (par définition de la distribution uniforme)[1], on peut écrire que:

$$P(|X_{(n)} - \theta| < \epsilon) = 1 - P(X_{(n)} < \theta - \epsilon).$$

Notons également que le maximum $X_{(n)}$ est inférieur à une constante, si chacune des variables aléatoires $X_1, \cdots, X_n$ est inférieure à cette constante. De plus, les variables aléatoires X_i étant $i.i.d.$, on obtient :

$$P(X_{(n)} < \theta - \epsilon) = [P(X_1 < \theta - \epsilon)]^n.$$

En d'autres mots, on voudrait pouvoir calculer $P(X_{(n)} < \theta - \epsilon)$. Pour plus de facilité, nous allons nous intéresser à trouver l'expression de $P(X_{(n)} < u)$ où u est une constante arbitraire :

$$\begin{aligned} P(X_{(n)} < u) &= P(X_1 < u \text{ et } X_2 < u \cdots \text{ et } X_n < u) \text{ (par définition de } X_{(n)}) \\ &= P(X_1 < u)P(X_2 < u) \cdots P(X_n < u) \text{ (car échantillon indépendant)} \\ &= P(X_1 < u)^n \text{ (car les } X_i \text{ sont iid.)} \end{aligned}$$

Enfin en utilisant la fonction de répartition de la loi uniforme on a :

$$\begin{aligned} P(X_{(n)} < \theta - \epsilon) &= [P(X_1 < \theta - \epsilon)]^n \\ &= \left(\int_0^{\theta - \epsilon} \frac{1}{\theta} dx \right)^n \\ &= \left(\frac{\theta - \epsilon}{\theta} \right)^n. \end{aligned}$$

Ainsi:

$$P(|X_{(n)} - \theta| < \epsilon) = 1 - \left(\frac{\theta - \epsilon}{\theta} \right)^n.$$

Le quotient $\frac{\theta - \epsilon}{\theta}$ étant compris dans l'intervalle $[0; 1)$, quelque soit la valeur positive que prend ϵ, le dernier terme de l'équation précédente tend vers 0 lorsque n tend vers l'infini. En d'autres termes, on a :

$$\lim_{n \to +\infty} P(|X_{(n)} - \theta| < \epsilon) = 1 - \lim_{n \to +\infty} \left(\frac{\theta - \epsilon}{\theta} \right)^n = 1 - 0 = 1.$$

1.7: Posons $W_i = X_i - Y_i$, alors on a que

$$\overline{W} = \frac{1}{n} \sum_{i=1}^n W_i = \frac{1}{n} \left(\sum_{i=1}^n X_i - \sum_{i=1}^n Y_i \right) = \overline{X} - \overline{Y},$$

[1] En effet, si on a une variable aléatoire Y qui suit une loi uniforme $\mathcal{U}[0, \theta]$, alors on sait que pour tout $\epsilon > 0$:

$$P(Y < \theta + \epsilon) = P(Y < \theta) + P(\theta \leq Y < \theta + \epsilon) = 1 + 0.$$

On a que $P(Y < \theta) = 1$ et $P(\theta \leq Y < \theta + \epsilon) = 0$ par définition de la fonction de densité de la loi uniforme, que l'on note f_Y. En effet, $P(Y < \theta) = P(0 < Y < \theta) = \theta \frac{1}{\theta}$ et $f_Y(y) = 0$ pour $\theta < y < \theta + \epsilon$, ainsi on a bien que $P(\theta \leq Y < \theta + \epsilon) = 0$.

où $\overline{W}$ est la moyenne de n variables iid. En appliquant le TCL[2], et en augmentant la taille de l'échantillon n, la distribution de probabilité de $\overline{W}$ se rapproche d'une loi normale de moyenne

$$E(\overline{W}) = \mu_1 - \mu_2$$

et de variance

$$V(\overline{W}) = V(\overline{X}) + V(\overline{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} = \frac{\sigma_1^2 + \sigma_2^2}{n}.$$

Il apparaît assez clairement que U_n n'est que la version centrée et réduite de $\overline{W}$. Comme cette dernière variable se rapproche d'une loi normale, alors U_n se rapproche d'une loi normale centrée et réduite à mesure que $n \to +\infty$.

8.2 Solutions Chapter 2

2.1: will be discussed in class

2.2: will be discussed in class

2.3: Puisque $\widehat{p}_1$ et $\widehat{p}_2$ sont des estimateurs non-biaisés des vraies proportions, p_1 et p_2, nous avons aussi

$$\mathbb{E}[\widehat{p}_1 - \widehat{p}_2] = \mathbb{E}[\widehat{p}_1] - \mathbb{E}[\widehat{p}_2] = p_1 - p_2.$$

De même, par l'indépendance des deux échantillons, nous savons que $\mathbb{V}\mathrm{ar}(\widehat{p}_1 - \widehat{p}_2) = \mathbb{V}\mathrm{ar}(\widehat{p}_1) + \mathbb{V}\mathrm{ar}(\widehat{p}_2)$. Notons que $\widehat{p}_1 = \frac{v_1}{n_1}$, où n_1 est le nombre total d'hommes dans l'échantillon, et v_1 le nombre d'hommes dans l'échantillon qui sont en faveur d'une vaccination. La v.a. v_1 est une variable binomiale avec paramètres p_1 et n_1, elle possède donc une variance de $\mathbb{V}\mathrm{ar}(v_1) = n_1 p_1 (1 - p_1)$. Dès lors,

$$\mathbb{V}\mathrm{ar}(\widehat{p}_1) = \frac{n_1 p_1 (1 - p_1)}{n_1^2} = \frac{p_1 (1 - p_1)}{n_1}.$$

De manière similaire, nous obtenons

$$\mathbb{V}\mathrm{ar}(\widehat{p}_2) = \frac{n_2 p_2 (1 - p_2)}{n_2^2} = \frac{p_2 (1 - p_2)}{n_2}$$

et finalement

$$\mathbb{V}\mathrm{ar}(\widehat{p}_1 - \widehat{p}_2) = \frac{p_1 (1 - p_1)}{n_1} + \frac{p_2 (1 - p_2)}{n_2}.$$

2.4: Pour $\hat{\mu}_1$: $\mathbb{E}(\hat{\mu}_1) = \frac{1}{3} \sum_{i=1}^{3} \underbrace{\mathbb{E}(Y_i)}_{\mu} = \frac{1}{3} 3\mu = \mu$, et $B(\hat{\mu}_1) = \mu - \mu = 0$, c'est donc un estimateur non-biaisé. Il possède une variance

$$\mathbb{V}\mathrm{ar}(\hat{\mu}_1) = \frac{1}{3^2} \sum_{i=1}^{3} \underbrace{\mathbb{V}\mathrm{ar}(Y_i)}_{\sigma^2} = \frac{\sigma^2}{3}$$

[2] Théorème central limite.

De manière similaire, nous obtenons pour $\hat{\mu}_2$:

$$\mathbb{E}(\hat{\mu}_2) = \frac{1}{4}\underbrace{\mathbb{E}(Y_1)}_{\mu} + \frac{1}{2}\underbrace{\mathbb{E}(Y_2)}_{\mu} + \frac{1}{4}\underbrace{\mathbb{E}(Y_3)}_{\mu} = \mu$$

c'est donc également un estimateur non-biaisé. Par contre, sa variance vaut

$$\mathbb{V}\mathrm{ar}(\hat{\mu}_2) = \frac{1}{4^2}\underbrace{\mathbb{V}\mathrm{ar}(Y_1)}_{\sigma^2} + \frac{1}{2^2}\underbrace{\mathbb{V}\mathrm{ar}(Y_2)}_{\sigma^2} + \frac{1}{4^2}\underbrace{\mathbb{V}\mathrm{ar}(Y_3)}_{\sigma^2} = \frac{3\sigma^2}{8}.$$

L'efficacité relative de $\hat{\mu}_1$ par rapport à $\hat{\mu}_2$ est donnée par

$$\text{eff}(\hat{\mu}_1, \hat{\mu}_2) = \frac{\mathbb{V}\mathrm{ar}(\hat{\mu}_2)}{\mathbb{V}\mathrm{ar}(\hat{\mu}_1)} = \frac{\frac{3\sigma^2}{8}}{\frac{\sigma^2}{3}} = \frac{9}{8} = 1.125$$

L'estimateur $\hat{\mu}_2$ possède une variance 12.5% plus élevée que celle de $\hat{\mu}_1$, il est 12.5 % moins efficace.

8.3 Solutions Chapter 3

3.1: La marge d'erreur pour une moyenne à un niveau de confiance $1-\alpha$ vaut $z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$. Ici, $\alpha = 2\%$, et donc, $z_{\alpha/2} = z_{0.01} = 2.32$. De plus, $\hat{\sigma} = 15$. Nous avons donc l'inégalité

$$2.32\frac{15}{\sqrt{n}} < 5$$

ou bien $n > 2.32^2 \times 9 = 48.44$.

3.2: La marge d'erreur pour une variance est $z_{\alpha/2}\frac{S^2\sqrt{\hat{K}-1}}{\sqrt{n}}$. Ici, $n = 20$, $\hat{K} = 3$, et $S^2 = 2$. L'intervalle de confiance est donc

$$[2 \pm z_{\alpha/2}\frac{2\sqrt{2}}{\sqrt{20}}] = [2 \pm z_{\alpha/2}\sqrt{2/5}]$$

La borne inférieure n'est pas négative si

$$z_{\alpha/2}\sqrt{2/5} \leq 2$$

ou bien $z_{\alpha/2} \leq \sqrt{5} \approx 2.236$, et on trouve $\alpha \geq 0.018$. Le niveau de confiance maximal pour assurer que l'intervalle ne couvre pas de valeurs négatives maximal est donc $1 - \alpha = 0.982$.

3.3: L'estimateur ponctuel de la différence entre les temps de réaction des hommes et des femmes est $\bar{X} - \bar{Y} = 3.6 - 3.8 = -0.2$. La variance estimée de cette différence est

$$\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y} = \frac{0.18}{50} + \frac{0.14}{50} = 0.0064.$$

(a) Avec un niveau de confiance de 95%, nous avons $z_{\alpha/2} = z_{0.025} = 1.96$, ce qui donne l'intervalle de confiance

$$[-0.2 \pm 1.96 \times \sqrt{0.0064}] = [-0.3568, -0.04].$$

Puisque cet intervalle ne couvre que des valeurs négatives, nous pourrions dire que, avec un niveau de confiance de 95%, le temps de réaction des hommes est plus petite que celui des femmes.

(b) Avec un niveau de confiance de 99%, nous avons $z_{\alpha/2} = z_{0.005} = 2.58$, ce qui donne l'intervalle de confiance

$$[-0.2 \pm 2.58 \times \sqrt{0.0064}] = [-0.4064, 0.0064].$$

Maintenant, puisque cet intervalle couvre aussi des valeurs positives, nous ne pouvons pas prétendre que, avec un niveau de confiance de 99%, le temps de réaction des hommes est plus petite que celui des femmes.

8.4 Solutions Chapter 4

4.1: Ici nous avons le cas d'une v.a. normale avec espérance $\mu = 0$ et variance σ^2. C'est-à-dire on connaît l'espérance μ et sa valeur de zéro. Il s'agit d'estimer le paramètre σ^2 avec la méthode de moments. Le premier moment ne donne ici aucune information, puisque

$$\mathbb{E}[Y] = \mu = 0.$$

Passons au deuxième moment (non-centré):

$$\mathbb{E}(Y^2) = \sigma^2 + \mu^2$$

puisque la variance est égale à $\sigma^2 = \mathbb{E}(Y^2) - \mu^2$ par définition. Mais $\mu = 0$, donc:

$$\mathbb{E}(Y^2) = \sigma^2$$

Ceci est égalisé avec le deuxième moment de l'échantillon, i.e.

$$m_2' = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

et nous obtenons l'estimateur de σ^2 selon la méthode des moments:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

4.2: Maintenant, μ est aussi un paramètre inconnu, à estimer, c'est-à-dire $Y \sim N(\mu, \sigma^2)$. Le premier moment théorique est $\mathbb{E}(Y) = \mu$ qui est égalisé avec le premier moment échantillon, i.e. $\bar{Y}$, ce qui donne directement l'estimateur de μ:

$$\widehat{\mu} = \bar{Y}.$$

Le deuxième moment théorique est égale à

$$\mathbb{E}(Y^2) = \sigma^2 + \mu^2$$

ce qui est égalisé avec le deuxième moment d'échantillon, i.e.

$$m_2' = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

En remplaçant μ par $\widehat{\mu}$, nous obtenons l'estimateur de σ^2 selon la méthode de moments:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2.$$

Notons que cet estimateur est égale à notre estimateur appelé S^2.

4.3: Tout d'abord, nous devons calculer le premier moment théorique de Y:

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^1 f(y|\theta)y dy \\ &= (\theta+1)\int_0^1 y^{\theta+1} dy \\ &= \frac{\theta+1}{\theta+2}[y^{\theta+2}]_0^1 \\ &= \frac{\theta+1}{\theta+2}\end{aligned}$$

ce qui est à égaliser avec le premier moment de l'échantillon, c'est-à-dire, $\bar{Y}$. Nous obtenons donc l'équation

$$\frac{\widehat{\theta}+1}{\widehat{\theta}+2} = \bar{Y}$$

qui est à résoudre pour $\widehat{\theta}$, ce qui donne la solution

$$\widehat{\theta} = \frac{2\bar{Y}-1}{1-\bar{Y}}$$

C'est un estimateur consistant par la propriété P4 puisque $\widehat{\theta}$ est une fonction continue de $\bar{Y}$ qui lui est un estimateur consistant de $\mathbb{E}(Y)$ (loi des grands nombres).

4.4: Il est supposé ici que l'espérance d'une v.a. Poisson est donnée par son paramètre λ. Nous pouvons donc simplement procéder à l'estimer en égalisant l'espérance théorique, λ, avec la moyenne de l'échantillon, $\bar{Y}$, ce qui donne l'estimateur selon la méthode de moments:

$$\widehat{\lambda} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

4.5: L'estimateur selon la méthode de maximum vraisemblance maximise la fonction de vraisemblance,

$$L(\lambda) = \prod_{i=1}^{n} f(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

ou bien, de manière équivalente, maximiser le logarighme de cette fonction,

$$\begin{aligned}\log L(\lambda) &= \sum_{i=1}^{n} \log f(y_i|\lambda) \\ &= \log\lambda \sum_{i=1}^{n} y_i - n\lambda - \sum_{i=1}^{n} \log(y_i!)\end{aligned}$$

Pour maximiser cette fonction, calculons la dérivée et égalisons la à zéro pour la condition de premier ordre:

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} y_i - n = 0$$

Cette équation sera résolue par rapport à λ pour donner l'estimateur de maximum vraisemblance:

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{Y}$$

Observons qu'il n' y a pas de différence dans ce cas entre l'estimateur de moments et celui de maximum vraisemblance.

8.5 Solutions Chapter 5

5.1: Pour tout $\mu \in \mathbb{R}$, on a :

$$\sqrt{n}(\bar{X} - \mu_0) = \sqrt{n}(\bar{X} - \mu) + \sqrt{n}(\mu - \mu_0) \sim \mathcal{N}(\sqrt{n}(\mu - \mu_0), 1).$$

Or, on sait que : $\sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0, 1)$. Alors, $\sqrt{n}(\bar{X} - \mu) + \sqrt{n}(\mu - \mu_0) \sim \mathcal{N}(\sqrt{n}(\mu - \mu_0), 1)$.

Nous rejetons H_0 si $T > z_\alpha$. Il s'en suit que :

$$\beta(\mu) = P(T < z_\alpha|\mu) = P(N(0,1) + \sqrt{n}(\mu - \mu_0) < z_\alpha) = \Phi(z_\alpha - \sqrt{n}(\mu - \mu_0)),$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée réduite. La fonction puissance notée $\pi(\mu)$, est simplement $\pi(\mu) = 1 - \beta(\mu)$.

On constate que si n s'accroît de plus en plus jusqu'à tendre vers l'infini, $\pi(\mu) \to 1$ puisque $\mu - \mu_0 > 0$ lorsque H_a est vraie. Ainsi $\beta(\mu) \to 0$ lorsque $n \to +\infty$.

5.2: Lorsque $\mu = \mu_0$ (H_0 est vraie), β prend la valeur $\beta(\mu_0) = \Phi(z_\alpha) = 1 - \alpha$. Il devrait en être ainsi puisque la probabilité de rejeter H_0 à tort est α. La probabilité de ne pas rejeter H_0 lorsqu'elle est vraie serait donc $1 - \alpha$.

5.3: Nous avons :

$$0.9 = 1 - \Phi(z_\alpha - \sqrt{n}).$$

Donc, $\Phi(z_\alpha - \sqrt{n}) = 0.1$, ou $\sqrt{n} = z_\alpha - \Phi^{-1}(0.1)$, ou[3] $n = (z_\alpha - \Phi^{-1}(0.1))^2$. Pour $\alpha = 0.05$, $z_\alpha = 1.645$, et $\Phi^{-1}(0.1)$ est le quantile 10% de la distribution normale centrée réduite, qui vaut -1.2815. n doit donc être plus large que 8.56. Il suffit alors que $n \geq 9$ pour avoir la puissance requise.

[3] La fonction Φ^{-1} est une *fonction quantile* : soit $p \in [0, 1]$ alors $\Phi^{-1}(p) = x$ où x est tel que $\Phi(x) = P(Z < x) = p$ et ainsi $x = z_{1-p}$.

8.6 Solutions Chapter 6

1. will be discussed in class, see also Chapter 14.2 of Wackerly et al.

2. (a) En remplaçant $\widehat{E}[n_{ij}]$ par $\frac{r_i c_j}{n}$ dans la formule de X^2, on obtient l'expression après en développant le carré.

   (b) Dès lors, il suit que X^2 ne change pas quand on multiplie chaque entrée du tableau par une même constante entière non nul.

8.7 Solutions Chapter 7

7.1: Minimiser $\sum_i (Y_i - \beta X_i)^2$ par rapport à β permet de respecter la condition de première ordre:

$$\frac{\partial \left(\sum_i (Y_i - \hat{\beta} X_i)^2\right)}{\partial \hat{\beta}} = -2 \sum_i (Y_i - \hat{\beta} X_i) X_i \overset{!}{=} 0$$

ce qui donne la solution :

$$\hat{\beta} = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}.$$

7.2: Les estimateurs des moindres carrés sont donnés par[4] :

$$\hat{\beta_1} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}, \quad \hat{\alpha_1} = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (Y_i - \bar{Y})^2}.$$

Les deux estimateurs sont identiques si les variances d'échantillons de X et Y sont égales puisque cela implique que les deux dénominateurs sont égaux

7.3: L'élasticité de Y par rapport à X est définie comme $\left(\frac{dY}{dX}\right) \frac{X}{Y}$, ou de manière équivalente (puisque $dY = Y d \log Y$) comme :

$$\frac{d \log Y}{d \log X}$$

qui dans le modèle log-linéaire est tout simplement égal à β_1. En effet, on a :

$$\frac{d \log Y}{d \log X} = \frac{d\,(\beta_0 + \beta_1 \log X + \epsilon)}{d \log X} = \beta_1 \text{ (puisque l'on dérive par rapport à } \log X).$$

7.4: La somme des résidus des moindres carrés est :

$$\begin{aligned}
\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= \sum_i (Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) - \hat{\beta}_1 X_i) \text{ (par "définition" de } \hat{\beta}_0) \\
&= n\bar{Y} - n\bar{Y} - \hat{\beta}_1 \sum_i (X_i - \bar{X}) \\
&= 0
\end{aligned}$$

où on a bien que $\sum_i (X_i - \bar{X})$ (voir TP 2, Exercices théoriques 3).

[4] Voir la formule slide $6 - 6$.

7.5: La covariance et les variances sont données par (voir notes de cours) :

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{X}\frac{\sigma^2}{S_{xx}}, \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad Var(\hat{\beta}_0) = \sigma^2 \frac{n^{-1}\sum_i X_i^2}{S_{xx}},$$

Donc,

$$Corr(\hat{\beta}_0, \hat{\beta}_1) = \frac{Cov(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_0)Var(\hat{\beta}_1)}} = \frac{-\bar{X}}{\sqrt{n^{-1}\sum_i X_i^2}}.$$

On peut aisément montrer que cette corrélation est toujours une valeur comprise entre -1 and 1 (mais ceci ne fait pas l'objet de la question).

7.6: Par définition,

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}.$$

En utilisant les calculs de la page 76 du syllabus, on a aussi

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}}.$$

Il suit de la défintion de SSE que l'on peut réécrire cette quantité de la manière suivante

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy}.$$

On déduit

$$R^2 = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2,$$

par définition de $\hat{\beta}_1$ et r^2.

7.7: Puisque $\hat{\beta}_1 = S_{xy}/S_{xx}$ et $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, on a d'abord : $\frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{S_{xx}}} = r\frac{\sqrt{S_{yy}}}{\widehat{\sigma}}$. Puis, parce que $\widehat{\sigma}^2 = n^{-1}SSE$ et $r^2 = 1 - SSE/S_{yy}$ (voir notes de cours), on obtient :

$$\frac{\widehat{\beta}_1}{\widehat{\sigma}/\sqrt{S_{xx}}} = \frac{r\sqrt{n}}{\sqrt{1-r^2}}.$$

Le carré du coefficient de corrélation est également appelé coefficient de détermination.

Bibliography

Baltagi, B. (2011). *Econometrics*. 5th ed. Springer.

Hafner, C. M. (2021). "Teaching statistical inference without normality". *LIDAM discussion paper, UCLouvain*.

Linton, O. (2017). *Probability, Statistics and Econometrics*. Academic Press.

Stock, J. and Watson, M. (2012). *Introduction to Econometrics*. 3rd ed. Pearson.

Wackerly, D., Mendenhall, W., and Sheaffer, R. (2008). *Mathematical Statistics with Applications*. 7th ed.