ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

Statistics II – Lecture notes

Economics, Finance and Management

Nuno M. Brites

February, 2023

Contents

ISEG

Lisbon School
of Economics
& Management
Universidade de Lisboa

All course-related information can be found at

https://fenix.iseg.ulisboa.pt/courses/est2-e-847427112272198/pagina-inicial.

All errors and omissions are entirely my responsibility. Please notify me if you find any errors or typos. Suggestions and feedback are also welcome. Have a happy and successful year!

These notes should not be used in place of a thorough reading of the bibliography.

Thanks,

Nuno M. Brites

nbrites@iseg.ulisboa.pt

ISEG, February 2023

Chapter 1

Sampling

1.1 Introduction

- Descriptive statistics: "Organisation" of observations/information (almost always the first chapter of a Statistics book).
- Probability theory (Statistics I): starting from a certain model and calculating the probability of certain results or events.
- Statistical inference (Statistics II): starting from observations and trying to infer something about the model.

Example 1.1. Suppose a population following a normal distribution with an unknown mean μ. We intend to "estimate" (approximately) μ based on a sample. Can $\overline{X}$ be used to estimate μ?

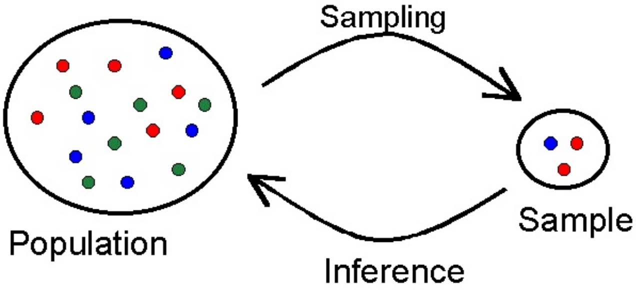


Figure 1.1: Image credit: http://testofhypothesis.blogspot.com/2014/09/the-sample.html

1.2 Random sampling

- Sampling process: the sample collection process must depend on chance.

Definition 1.1 (Random sampling)**.** When the n observed random variables, components of the vector $(X_1, \ldots, X_n)$ are **i.i.d.** (independent and identically distributed), it is said to be a **random sample** (r.s). Basically, we are considering that each X_i is, in terms of distribution, a "copy" of the r.v. X.

- Independence means that

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \times \ldots \times F_{X_n}(x_n).$$

- **Attention to the notation:** $X_i \neq x_i$.
- Random Sampling Process: The observed data are just one of many data sets that could have been obtained under the same circumstances. The observed sample of n observations, $(x_1, \ldots, x_n)$ is a realization of the n-dimensional random variable $(X_1, \ldots, X_n)$. Be careful with the notation:

  - $(X_1, \dots, X_n)$: random sample (variables)
  - $(x_1, \dots, x_n)$: observed sample (constants)
- The sample space, $\mathcal{X} \subset \mathbb{R}^n$ is the set of all selectable samples.
- The probabilistic model (representing the population/universe) is a family of distributions indexed by an unknown parameter (possibly a vector),

$$F_\theta = \{F(x \mid \theta) : \theta \in \Theta\}.$$

Example 1.2. Assume that $X \sim B(1, \theta)$, i.e., whether or not a person practises sports. The probabilistic model is, in this case,

$$F_\theta = \{f(x \mid \theta) : \theta^x (1-\theta)^{1-x} : x = 0, 1; \theta \in (0,1)\}.$$

A random sample with size 5 could be $x = (1, 1, 0, 1, 0)$.

1.3 Statistics

Definition 1.2 (Statistic)**.** A **Statistic** is a random variable or a random vector, $T = T(X_1, \dots, X_n)$, that depends on a random sample $(X_1, \dots, X_n)$ but not on any unknown parameter.

Remark. The main advantage of using Statistics is that it allows you to reduce information. Is it preferable to work with $\overline{X}$ or with the (entire) sample $(X_1, \dots, X_n)$?

Example 1.3. Let $X_1, \dots, X_n$ be a r.s. of a Bernoulli population. The statistic $T_1 = \sum_{i=1}^n X_i$ represents the number of successes in the sample. The statistic $T_2 = \frac{1}{n} \sum_{i=1}^n X_i = \overline{X}$ indicates the proportion of successes in the sample.

Example 1.4. Let $(X_1, \dots, X_n)$ represent a r.s. from a normal population $\mathcal{N}(\mu, \sigma^2)$ with unknown parameters. Then,

- $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\frac{1}{n}\sum_{i=1}^n X_i$ and $\frac{1}{n}\sum_{i=1}^n X_i^2$ are statistics.
- $\sum_{i=1}^n (X_i - \mu)/\sigma$ e $\sum_{i=1}^n (X_i/\sigma)^2$ are not statistics, as they depend on unknown parameters.

1.4 Sampling distributions

- Given a T statistic, which is a function of a r.s. $(X_1, \dots, X_n)$, we can get several T values depending on the observed samples:
  - Observed sample #1:
  $$(x_{11}, \dots, x_{1n}) \Longrightarrow t_1 = T(x_{11}, \dots, x_{1n}).$$
  - Observed sample #2:
  $$(x_{21}, \dots, x_{2n}) \Longrightarrow t_2 = T(x_{21}, \dots, x_{2n}).$$
  -
  $$\dots$$
  - Observed sample #k:
  $$(x_{k1}, \dots, x_{kn}) \Longrightarrow t_k = T(x_{k1}, \dots, x_{kn}).$$
- The distribution function (density function or probability function) determines the probabilistic behavior of the statistic $T = T(X_1, \dots, X_n)$.

Exercise 1.1. What is the distribution of $\overline{X}$ if $X \sim \mathcal{N}(\mu, \sigma^2)$?

```
#
```

```








#
```

- Distribution of the sample (joint probability function or probability density function):

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1}(x_1) \times \ldots \times f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \quad \text{(iid's)}$$

- Distribution of the statistic $T = T(X_1,\ldots,X_n)$:

$$P(T \le t) = \int\int \ldots \int_{A(t)} \left(\prod_{i=1}^{n} f_{X_i}(x_i) \right) dx_1 dx_2 \ldots dx_n. \quad \text{(continuous r.v.)}$$

$$P(T \le t) = \sum_{A(t)} \left(\prod_{i=1}^{n} f_{X_i}(x_i) \right). \quad \text{(discrete r.v.)}$$

$$A(t) = \{(x_1,\ldots,x_n) \in \mathbb{R}^n : T(x_1,\ldots,x_n) \le t\}.$$

- As we'll see later, there are easier ways to get the sampling distribution of a T statistic!

Example 1.5. Let $(X_1,\ldots,X_n)$ be a r.s. from a Poisson population, that is, $X \sim Po(\lambda)$ and $f(x) = P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0,1,2,\ldots,\lambda > 0$. From the Poisson distribution properties,

$$T = \sum_{i=1}^{n} X_i \sim Po(n\lambda).$$

Thus, the T statistic has a probability function given by

$$f_T(t) = P(T = t) = \frac{(n\lambda)^t e^{-n\lambda}}{t!}, \quad t = 0,1,2,\ldots,\lambda > 0.$$

1.4.1 Order statistics

- Let $(X_1,\ldots,X_n)$ be a r.s. and $X_i \sim F_X(x)$ with density/probability $f_X(x)$.
- Order statistics are obtained by ordering the sample:

$$(X_1,\ldots,X_n) \Longrightarrow (X_{(1)}, X_{(2)} \ldots, X_{(n)}), \text{ with } X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}.$$

- $X_{(1)}$ is the sample minimum and $X_{(n)}$ is the sample maximum.
- Sampling distribution of the **minimum**, $X_{(1)}$

$$G_1(x) \equiv P(X_{(1)} \le x) = 1 - (1 - F_X(x))^n.$$

- Sampling distribution of the **maximum**, $X_{(n)}$

$$G_n(x) \equiv P(X_{(n)} \le x) = (F_X(x))^n.$$

- If the random variable X is continuous, the sampling density of the **minimum/maximum**, are

$$g_1(x) \equiv G_1'(x) = \left(1 - (1 - F_X(x))^n\right)' = n\,(1 - F_X(x))^{n-1} f_X(x).$$

and

$$g_n(x) \equiv G_n'(x) = \left((F_X(x))^n\right)' = n\,(F_X(x))^{n-1} f_X(x).$$

Example 1.6. Let X be an exponentially distributed universe with parameter λ, from which a r.s. of size n was collected, i.e., $(X_1, \ldots, X_n)$. Determine the distribution function and density function of the sample minimum and maximum.

- $X \sim Exp(\lambda) \Longrightarrow F(x) = 1 - e^{-\lambda x}$.
- $G_n(x) = P(X_{(n)} \leq x) = (F_X(x))^n = \left(1 - e^{-\lambda x}\right)^n$.
- $g_n(x) = G_n'(x) = \left[\left(1 - e^{-\lambda x}\right)^n\right]' = n\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{n-1}$.
- $G_1(x) = P(X_{(1)} \leq x) = 1 - (1 - F_X(x))^n = 1 - \left(1 - \left(1 - e^{-\lambda x}\right)\right)^n = 1 - e^{-n\lambda x}$.
- $g_1(x) = G_1'(x) = \left[1 - e^{-n\lambda x}\right]' = n\lambda e^{-n\lambda x}$.

1.5 First results on the sample mean and variance

Definition 1.3. Let X be a universe, from which a r.s. of size n was collected, $(X_1, \ldots, X_n)$. The sample mean, $\overline{X}$, and the sample variance, S^2, are defined, respectively, by

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

Exercise 1.2. Show that

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

#

#

Theorem 1.1. *Let $(X_1, \ldots, X_n)$ be a r.s. from a population X with $\mu = E(X) < +\infty$ and $\sigma^2 = Var(X) < +\infty$. The distribution of the sample mean is:*

$$E(\overline{X}) = \mu, \qquad Var(\overline{X}) = \frac{\sigma^2}{n}.$$

Remark. On average, the sample mean takes the value of the population mean. When $n \to \infty$, the variance of the sample mean tends to zero.

Theorem 1.2. *Let $(X_1, \dots, X_n)$ be a r.s. from a population X with $\mu = E(X) < +\infty$ and $\sigma^2 = Var(X) < +\infty$. The expected value of the sample variance is:*

$$E(S^2) = \frac{n-1}{n}\sigma^2.$$

Remark. On average, $S^2 < \sigma^2$ because $\frac{n-1}{n} < 1$, which is illogical, especially in small sample sizes. A possible solution is to work with the corrected sample variance,

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

Thus,

$$E(S'^2) = E\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1}E(S^2) = \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2 = \sigma^2,$$

that is,

$$E(S'^2) = \sigma^2.$$

1.6 Asymptotic sampling distributions

Recall the Central Limit Theorem (CLT):

Theorem 1.3. *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with expected value $E(X_i) = \mu < \infty$ and variance $Var(X_i) = \sigma^2 < \infty$. Let Z_n be the standardized mean*

$$Z_n := \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Then, for n sufficiently large,

$$Z_n \overset{a}{\sim} \mathcal{N}(0,1).$$

Or equivalently,

$$\overline{X}_n \overset{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

In other words, for sufficiently large n, the sample mean $\overline{X}_n$ is close to be normal distributed with mean μ and variance σ^2/n.

- Exact distributions for the desired statistics are frequently unavailable.
- In general, it is possible to get asymptotic sampling distributions if the **population moments** exist up to a certain order.

Theorem 1.4. *Let $(X_1, \dots, X_n)$ be a r.s. from a population X with $\mu = E(X) < +\infty$ and $\sigma^2 = Var(X) < +\infty$. Then, for $n \to +\infty$, by the Central Limit Theorem (CLT)*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} \mathcal{N}(0,1).$$

We thus obtain the asymptotic distribution of the sample mean, that is,

$$\overline{X}_n \overset{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Remark: If $X \sim \mathcal{N}(\mu, \sigma^2)$, the distribution is obviously exact.

Exercise 1.3. Let $(X_1, \dots, X_{30})$ be a r.s. from a uniform population in the interval (0.10). Compute $P(\overline{X} < 5.5)$.

```
#

#
```

1.7 Bernoulli population sampling: case of one proportion

- The population consists of two types of individuals: those with and those without a particular characteristic.
- Let's consider a r.s. $(X_1, \ldots, X_n)$ from a population $X \sim B(1,p)$, that is,

$$f(x) = P(X = x) = p^x (1-p)^{1-x}, \quad x = 0, 1, \quad p \in (0,1).$$

- It is of general interest to establish the sampling distribution of two statistics:

$$Y = \sum_{i=1}^{n} X_i \qquad \text{(sum of successes)}$$

and

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{(proportion of successes).}$$

- Let's see first the exact distribution and then the asymptotic distribution ($n \geq 30$).

1.7.1 Exact distribution

- Distribution for the **sum of successes**:

$$X \sim B(1,p) \Longrightarrow X_i \sim B(1,p) \Longrightarrow Y = \sum_{i=1}^{n} X_i \sim B(n,p) \Longrightarrow E(Y) = np \text{ and } Var(Y) = np(1-p).$$

$$f_Y(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

- Distribution for the **proportion of successes**:

$$X \sim B(1,p) \Longrightarrow f_{\overline{X}}(u) = P(\overline{X} = u) = P\left(\frac{Y}{n} = u\right) = P(Y = nu) = \binom{n}{nu} p^{nu} (1-p)^{n-nu},$$

with $u = 0, 1/n, 2/n, \ldots, 1$.

1.7.2 Asymptotic distribution ($n \geq 30$)

- From Statistics I,

$$X_i \sim B(1,p) \Longrightarrow Y = \sum_{i=1}^{n} X_i \sim B(n,p) \Longrightarrow E(Y) = np \text{ and } Var(Y) = np(1-p).$$

By applying CLT, we get the **distribution for the sum of successes**, i.e.,

$$Z_n = \frac{Y - E(Y)}{\sqrt{Var(Y)}} = \frac{Y - np}{\sqrt{np(1-p)}} \overset{a}{\sim} \mathcal{N}(0.1).$$

- Applying **Theorem 1.1**,

$$E(\overline{X}) = \mu_X = p, \qquad Var(\overline{X}) = \frac{\sigma_X^2}{n} = \frac{p(1-p)}{n}.$$

By applying the CLT we get the **distribution for the proportion of successes**, i.e.,

$$Z_n = \frac{\overline{X} - E(\overline{X})}{\sqrt{Var(\overline{X})}} = \frac{\overline{X} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \overset{a}{\sim} \mathcal{N}(0,1).$$

1.8 Bernoulli population sampling: case of two proportions

- We now have two Bernoulli populations with parameters p_1 and p_2. Usually, the interest is to compare the two proportions p_1 and p_2 (e.g. proportion of students approved in classes 1 and 2). In sampling studies, the difference $p_1 - p_2$ is not known. The idea is to collect two independent samples (one from each population) and use the statistic $\overline{X}_1 - \overline{X}_2$ (the difference between observed proportions) to infer about $p_1 - p_2$.
- Suppose we have collected two r.s. (not necessarily of the same size):

$$X_1 \sim B(n_1, p_1) \Longrightarrow (X_{11}, \ldots, X_{1n}) \Longrightarrow \overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}.$$

$$X_2 \sim B(n_2, p_2) \Longrightarrow (X_{21}, \ldots, X_{2n}) \Longrightarrow \overline{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}.$$

- Goal: to determine the sampling distribution of $\overline{X}_1 - \overline{X}_2$.
- There is no exact result, so the asymptotic distribution is used.
- In the case of individual proportions, we have, by application of the CLT:

$$Z_{1n} = \frac{\overline{X}_1 - p_1}{\sqrt{\dfrac{p_1(1-p_1)}{n_1}}} \overset{a}{\sim} \mathcal{N}(0,1) \Longleftrightarrow \overline{X}_1 \overset{a}{\sim} \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right).$$

$$Z_{2n} = \frac{\overline{X}_2 - p_2}{\sqrt{\dfrac{p_2(1-p_2)}{n_2}}} \overset{a}{\sim} \mathcal{N}(0,1) \Longleftrightarrow \overline{X}_2 \overset{a}{\sim} \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

- So, for the difference in proportions, we have:

$$Z_n = \frac{\overline{X}_1 - \overline{X}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \overset{a}{\sim} \mathcal{N}(0,1).$$

$$\Longleftrightarrow$$

$$\overline{X}_1 - \overline{X}_2 \overset{a}{\sim} \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Exercise 1.4. The proportion of customers who opted for the TELELE mobile phone brand in the FENAQUE store was 0.35, and in the VORTENE store it was 0.29. Compute the probability that, taking a sample of 200 customers at the first store and 150 customers in the second, the sample proportion of customers who opted for the TELELE brand in the FENAQUE store is higher than that of the VORTENE store.

```
#

#
```

1.9 Normal population: distribution of the mean

- Let's consider a r.s. $(X_1, \ldots, X_n)$ from a population $X \sim \mathcal{N}(\mu, \sigma^2)$.
- The distribution of the sample mean (previously seen!) is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

or

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}\left(0, 1\right).$$

Exercise 1.5. Suppose that the duration (in minutes) of local telephone calls in a company is approximated by a normal distribution with $\mu = 17$ and $\sigma^2 = 25$. What is the probability that, in a random sample of 25 calls, the average duration is between 16 and 18 minutes?

```
#

#
```

1.10 Normal population: distribution of the variance

Theorem 1.5. *Consider a r.s.* $(X_1, \ldots, X_n)$ *from a population* $X \sim \mathcal{N}(\mu, \sigma^2)$. *The distribution of the variance is*

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1).$$

Exercise 1.6. Consider a normal population from which a sample of dimension 25 has been drawn. Compute the probability that the quotient between the corrected sample variance and the population variance is between 0.79 and 1.18.

```
#




















#
```

1.11 Normal population: Student's ratio

- Let's consider a r.s. $(X_1, \ldots, X_n)$ from a population $X \sim \mathcal{N}(\mu, \sigma^2)$.
- We have seen before that the exact distribution of the sample mean is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

or

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}\left(0, 1\right).$$

- What happens when σ is unknown? Since we have access to the sample, we can calculate the sample variance, S^2.
- Thus, when σ is **unknown**, the distribution of the sample mean is

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n-1}} = \frac{\overline{X} - \mu}{S'/\sqrt{n}} \sim t(n-1).$$

- The above ratio is called "Student's ratio" and follows a $t - Student$ (or simply t) distribution with $n-1$ degrees of freedom.

Definition 1.4 (t-Student distribution)**.** Consider U and V two random variables such that $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(n)$. Then,

$$T = \frac{U}{\sqrt{V/n}} \sim t(n),$$

that is, the r.v. T follows a $t - Student$ distribution with n degrees of freedom.

$t - Student's$ distribution properties:

- $E(T) = 0$ and $Var(T) = \dfrac{n}{n-2}$.
- $t(n) \to \mathcal{N}(0,1)$ as $n \to +\infty$.
- The density shape of the $t - student$ distribution is similar to the density shape of the normal distribution.
- The distribution of t-students is tabulated (see Table 8 – Newbold). Attention: in this table, the probabilities are calculated in the right-side tab!

1.12 Normal populations: difference between two means

Assume two normal populations $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ from which two random samples (not necessarily of equal size) were collected:

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \Longrightarrow (X_{11}, \ldots, X_{1n}) \Longrightarrow \overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}.$$

$$X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2) \Longrightarrow (X_{21}, \ldots, X_{2n}) \Longrightarrow \overline{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}.$$

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

- The previous result only applies when the variances of the two populations are known (a problem similar to the one that led to the introduction of the Student's ratio).
- When the variances are unknown but we assume they are equal, we have

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} \sqrt{\dfrac{(n_1 - 1)S_1'^2 + (n_2 - 1)S_2'^2}{n_1 + n_2 - 2}}} \sim t(n_1 + n_2 - 2).$$

- For large samples, when the variances are unknown and possibly different, we have

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1'^2}{n_1} + \dfrac{S_2'^2}{n_2}}} \overset{a}{\sim} \mathcal{N}(0,1).$$

- If the sample sizes are rather small, the approximation obtained using the CLT can be slightly improved by using the so-called Welch approximation:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1'^2}{n_1} + \dfrac{S_2'^2}{n_2}}} \overset{a}{\sim} t(r),$$

where r is the largest integer contained in

$$\frac{\left(\frac{s_1'^2}{n_1}+\frac{s_2'^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1'^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_1'^2}{n_2}\right)^2}.$$

Exercise 1.7. A certain pharmaceutical company has launched a new sleeping drug on the market that has been used in hospitals. It was found that patients who were not taking this drug slept for an average of 7.5 hours, with a standard deviation of 1.4 hours, while patients who received this drug slept for an average of 8 hours, with a standard deviation of 2 hours. At a certain hospital, 31 patients who were not taking the referred medication and 61 patients who were taking it were counted. What is the probability that patients in the first group slept more on average than those in the second group? Assume the normality of the distributions.

```
#














#
```

1.13 Normal populations: ratio of variances

- Take two random samples from two gaussian populations, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
- It is natural to think of the statistic $S_1'^2/S_2'^2$ when inferring the ratio of variances σ_1^2/σ_2^2 of two independent Gaussian populations.
- By **Theorem 1.6** we know that

$$(n-1)\frac{S'^2}{\sigma^2} \sim \chi^2(n-1).$$

- When considering two populations,

$$U = (n_1-1)\frac{S_1'^2}{\sigma_1^2} \sim \chi^2(n_1-1) \text{ and } V = (n_2-1)\frac{S_2'^2}{\sigma_2^2} \sim \chi^2(n_2-1), \text{ one get}$$

$$F = \frac{U}{V} \cdot \frac{n_2 - 1}{n_1 - 1} = \frac{(n_1 - 1)\frac{S_1'^2}{\sigma_1^2}}{(n_2 - 1)\frac{S_1'^2}{\sigma_1^2}} \cdot \frac{n_2 - 1}{n_1 - 1} = \frac{S_1'^2}{S_2'^2} \cdot \frac{\sigma_2^2}{\sigma_1^1} \sim F(n_1 - 1, n_2 - 1).$$

$F(n_1 - 1, n_2 - 1)$ represents the $F - Snedecor$ distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

Definition 1.5 (F-Snedecor distribution)**.** Consider two independent random variables U and V, such that $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$. Then,

$$F = \frac{U/m}{V/n} \sim F(m, n),$$

that is, the r.v. F follows a $F - Snedecor$ distribution with m and n degrees of freedom.

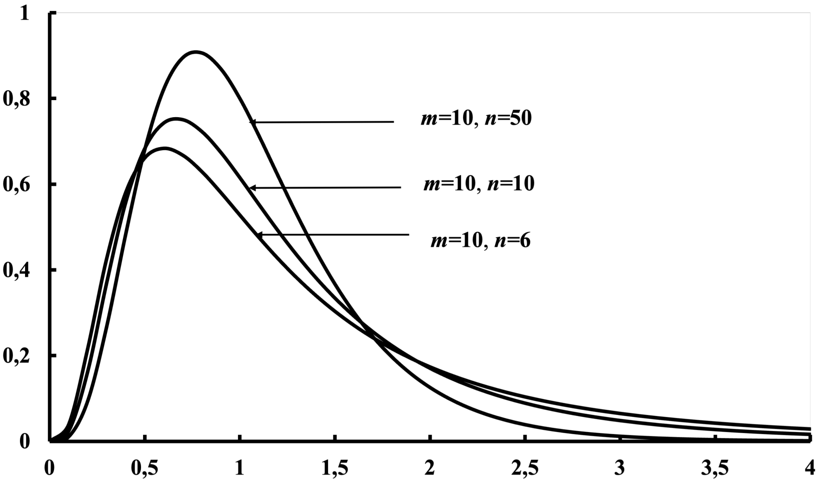


Figure 1.2: F distribution examples

$F - Snedecor$ distribution properties:

- $E(F) = \frac{n}{n-2}$, $n > 2$ and $Var(F) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$, $n > 4$.
- The distribution of F-Snedecor is tabulated (see Tables 9a and 9b – Newbold). Attention: in this table, the probabilities are calculated in the right-side tab!
- To obtain values in the left tab, use the property:

$$F \sim F(m, n) \Longrightarrow \frac{1}{F} \sim F(n, m).$$

Exercise 1.8. Assume that the IQ test results in countries A and B are well modelled by normal distributions of means of 100, and that a sample of size 16 is collected in country A and another of size 10 in country B. Assuming that the variances in the two populations are the same, what is the probability that the quotient between the corrected variances of the two samples is greater than 3.77?

```
#
```

```
#
```

Chapter 2

Point estimation

2.1 Introduction

- **Goal**: To say what we have learned about the unknown quantities (parameters) after observing some data (a random sample) that we believe contains relevant information.
- **Examples**:
  - What would we say if the probability that a future patient will respond successfully to treatment after we observe the results from a collection of other patients?
  - What can we say about whether a machine is functioning properly after we observe some of its output?
- **Background**: The methods of statistical inference, which we shall develop to address these questions, are built upon the theory of probability.
- Let's consider a r.s. $(X_1, \ldots, X_n)$ from a population with p.d.f./p.m.f. belonging to the familiy

$$F_\theta = \{f(x|\theta) \,:\, \theta \in \Theta\}.$$

- The functional form $f(\cdot)$ is known but the parameters θ are unknown.
- **Problem**: How to use the information contained in the sample to "guess" (estimate) the value of the unknown parameter(s) θ?
- **Important Idea**: given the size of the sample n, the more accurate the response (estimation), the less confidence there is.
- Parametric Estimation:
  - prioritizing accuracy $\Longrightarrow$ point estimation $\Longrightarrow$ estimates
  - prioritizing confidence $\Longrightarrow$ interval estimation $\Longrightarrow$ confidence interval
- Parameters:
  - Multidimensional parameter - Assume that the return on a financial asset follows a normal distribution with a mean of μ and a variance of σ^2. Then, we would like to estimate two unknown parameters: μ (mean return) and σ (volatility).
  - Parameter function - Assume that the number of claims per year on a given motor insurance policy follows a Poisson distribution with parameter λ. Instead of λ (the average number of claims per year), we might be interested in $P(X = 0 \mid \lambda) = e^{-\lambda}$, which is the probability that there are no claims. As a result, we'd like to calculate the function $h(\lambda) = e^{-\lambda}$ that represents such a probability.
- **Starting point**:
  - random sample $(X_1, X_2, \ldots, X_n)$ from a population $F_\theta = \{f(x|\theta) \,:\, \theta \in \Theta\}$.
  - $f(\cdot)$ is known and only θ is unknown.
  - the parametric space is Θ and $\theta \in \Theta$.

Definition 2.1 (Estimator)**.** An estimator is a Statistic, $T(X_1, X_2, \dots, X_n)$, that estimates some fact about the population. You can also think of an estimator as the rule that creates an estimate.

Example 2.1. $T(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$.

Definition 2.2 (Estimate)**.** Observed value of the estimator for a given observed sample.

Example 2.2. $t = T(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$.

Remark. How do we find an estimator for a given unknown parameter? Given two estimators, how do we assess their quality?

2.2 Method of Moments

- **Idea**: if we know that the parameter θ that we want to estimate is the mean of the population distribution (the first raw moment), then we can use the sample average (the first raw sample-moment) to estimate it: the larger the sample, the more similar the two will be. We will consider a random sample $(X_1, X_2, \dots, X_n)$ from a population with p.d.f./p.m.f. $f(x \mid \theta_1, \theta_2, \dots, \theta_k)$ with k unknown parameters.

Definition 2.3 (Method of Moments)**.** The Method of Moments estimator(s) can be obtained as follows. Let $\mu'_k = E(X^k) = f(\theta_1, \dots, \theta_L)$ denote the k^{th} raw population moment and let $\mu_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$ denote the k^{th} raw sample moment.

- Step 1: Determine L the number of parameters $(\theta_1, \dots, \theta_L)$ to estimate.
- Step 2: Find μ'_k and equate to m'_k for $k = 1, \dots, L$.
- Step 3: Solve this system of L equations for $\theta_1, \dots, \theta_L$.

The solutions are the Method of Moments Estimators $(\tilde{\theta}_1, \dots, \tilde{\theta}_L)$.

Exercise 2.1. Consider $X \sim B(1, \theta)$ of which a sample of size n was drawn with the purpose of estimating θ. Obtain $\tilde{\theta}$.

```
#









#
```

Exercise 2.2. Consider $X \sim \mathcal{N}(\mu, \sigma^2)$ of which a sample of size n was drawn with the purpose of estimating μ and σ^2. Obtain $\tilde{\mu}$ and $\tilde{\sigma^2}$.

```
#
```

```
#
```

Exercise 2.3. Consider $X \sim U(-\theta, \theta)$ of which a sample of size n was drawn with the purpose of estimating θ. Obtain $\tilde{\theta}$.

```
#



#
```

Exercise 2.4. Consider $X \sim Exp(\lambda)$ of which a sample of size n was drawn with the purpose of estimating λ. Obtain $\tilde{\lambda}$.

```
#



#
```

2.3 Maximum Likelihood

Definition 2.4 (Likelihood function)**.** Consider the joint probability function of the observations in a random sample, regarded as a function of the unknown parameter θ:

- $(X_1, X_2, \ldots, X_n)$ r.s. of a population with p.d.f./p.m.f. $f(x \mid \theta)$.
- joint p.d.f./p.m.f. of the random sample

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n,$$

representing the probability associated to the specific sample that was observed $(x_1, x_2, \ldots, x_n)$.

- for a given **fixed** observed random sample $(x_1, x_2, \ldots, x_n)$, this probability interpreted as function of the parameter θ defines the **likelihood function**:

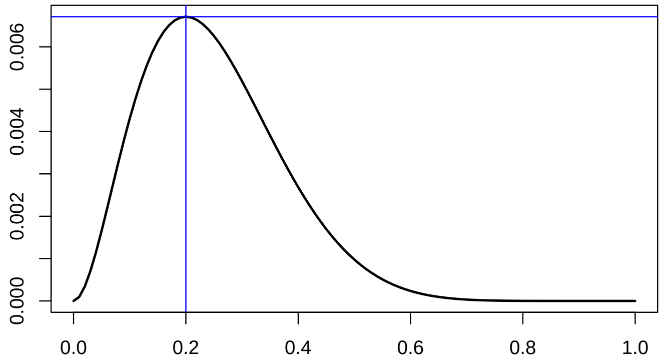
$$L(\theta) := L(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad \theta \in \Theta.$$

- for each value of θ in the parameter space, $L(\theta)$ gives the likelihood of observing the specific sample $(x_1, x_2, \ldots, x_n)$.

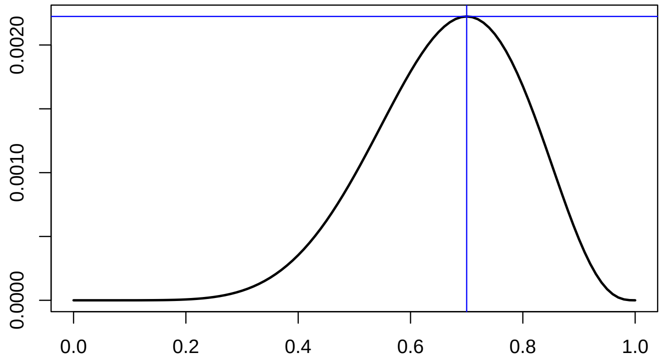
Exercise 2.5. Consider $X \sim B(1, \theta)$ of which a sample of size n was drawn with the purpose of estimating θ. Obtain the likelihood function and take a look at the following figures:

```
#









#
```

- $n = 10$ and $\sum_{i=1}^{10} x_i = 2 \Longrightarrow L(\theta) = \theta^2 (1-\theta)^{10-2} = \theta^2 (1-\theta)^8$. **The more likely values for θ are around** 0.2.



- $n = 10$ and $\sum_{i=1}^{10} x_i = 7 \Longrightarrow L(\theta) = \theta^7 (1-\theta)^{10-7} = \theta^7 (1-\theta)^3$. **The more likely values for θ are around** 0.7.



Definition 2.5 (Maximum Likelihood Estimator)**.** Given the observed sample $(x_1, \ldots, x_n)$ search for an estimate $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ such that

$$L(\hat{\theta} \mid x_1, x_2, \ldots, x_n) \geq L(\theta \mid x_1, x_2, \ldots, x_n), \qquad \theta \in \Theta.$$

This estimate corresponds to the estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$.

- Maximum: root of the first derivative (if it exists) and negative second derivative (if it exists).
- Usually, it is easier to consider the logarithm of the likelihood function (log-likelihood function):

$$l(\theta) = \ln(L(\theta)).$$

- Since the logarithm is a monotonic increasing function, $l(\theta)$ and $L(\theta)$ have the same maximizer.
- In general, the maximizer is given by:

$$\frac{dL(\theta)}{d\theta} = 0, \quad \frac{d^2L(\theta)}{d\theta^2} < 0 \quad \text{or} \quad \frac{dl(\theta)}{d\theta} = 0, \quad \frac{d^2l(\theta)}{d\theta^2} < 0.$$

Remark. Be careful!

- The maximising point may not be an interior point of the parametric space.
- The MLE does not need to be unique.
- $L(\theta)$ may have **local** maxima.
- The likelihood (or log-likelihood) function may not be differentiable.

Exercise 2.6. Consider $X \sim B(1, \theta)$ of which a sample of size n was drawn with the purpose of estimating θ. Obtain $\hat{\theta}$.

```
#








#
```

Exercise 2.7. $(X_1, \dots X_n)$ r.s. from X with p.d.f. $f(x \mid \theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$, and $E(X) = \dfrac{\theta}{1+\theta}$. Find $\tilde{\theta}$ and $\hat{\theta}$.

#

#

Exercise 2.8. $X \sim U(0, \theta)$. Find $\hat{\theta}$.

#

```
#
```

2.3.0.1 Invariance property of MLE

Theorem 2.1. *Let $\hat{\theta}$ be the MLE for θ and $h(x)$ a function of x. Then $h(\hat{\theta})$ is the MLE for $h(\theta)$.*

Exercise 2.9. $X \sim Po(\lambda)$. Find the MLE for $P(X < 0.1)$.

```
#



#
```

2.4 Properties of estimators

2.4.1 Unbiased estimators

Definition 2.6 (Unbiased estimator)**.** An estimator $T = T(X_1, X_2, \ldots, X_n)$ for θ is centered or *unbiased* if

$$E(T) = \theta, \quad \forall \theta \in \Theta.$$

Remark. The expected value is equal to the true value of the parameter θ, whatever the value of θ in the parametric space. The definition only makes sense if $E(T)$ exists.

Definition 2.7 (Bias)**.** If $E(T) \neq \theta$, then the estimator is biased and the bias is given by:

$$bias(T) = E(T) - \theta.$$

Exercise 2.10. $X \sim B(1, \theta)$. Is $\hat{\theta} = \overline{X}$ unbiased for θ?

```
#



#
```

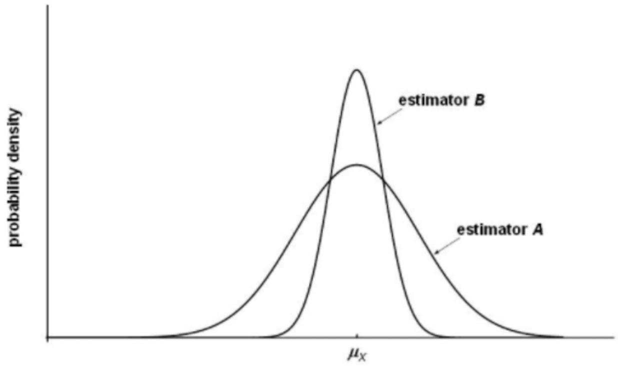


Figure 2.1: Two unbiased estimators.

- The concept of unbiasedness does not allow one to distinguish between two unbiased estimators with different sampling distributions (namely regarding variance):

Definition 2.8 (Efficiency)**.** Let T_1 and T_2 be two unbiased estimators for θ. The estimator T_1 is more efficient than T_2 when

$$Var(T_1) < Var(T_2), \quad \forall \theta \in \Theta.$$

The estimator T^* is the most efficient for θ when it is more efficient than any other unbiased estimator T for θ.

Remark. Regarding efficiency:

- Efficiency requires the existence of second-order moments in the estimator.
- The efficiency definition incorporates two different concepts:
  - relative efficiency: relation between two unbiased estimators for θ.
  - absolute efficiency: regarding all unbiased estimators for θ.
- In order to obtain the most efficient estimator, we rely on the **Fréchet-Cramér-Rao** lower bound.

Theorem 2.2 (Fréchet-Cramér-Rao lower bound)**.** *Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a population with p.d.f./p.m.f. $f(x \mid \theta)$, satisfying certain regularity conditions, and let $T = T(X_1, X_2, \ldots, X_n)$ be an unbiased estimator for θ. Then,*

$$Var(T) \geq \frac{1}{n\mathcal{I}(\theta)},$$

where

$$\mathcal{I}(\theta) = E\left[\left(\frac{d\ln(f(X|\theta))}{d\theta}\right)^2\right] = -E\left[\frac{d^2\ln(f(X|\theta))}{d\theta^2}\right]$$

*is the **Fisher information**.*

Remark. Generally, the second equality is easier to calculate. For certain distributions, the expression $\mathcal{I}(\theta)$ is known:

Exercise 2.11. Given $f(x \mid \theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0$, show that $\mathcal{I}(\theta) = \theta^{-2}$.

#

Distribution	Fisher information
$X \sim B(n;\theta)$ (n known)	$\mathcal{I}(\theta) = n/[\theta(1-\theta)]$
$X \sim Po(\lambda)$	$\mathcal{I}(\theta) = 1/\lambda$
$X \sim N(\mu,\sigma^2)$ (σ^2 known)	$\mathcal{I}(\theta) = 1/\sigma^2$
$X \sim N(\mu,\sigma^2)$ (μ known)	$\mathcal{I}(\theta) = 1/(2\sigma^4)$
$X \sim G(\alpha,\lambda)$ (α known)	$\mathcal{I}(\theta) = \alpha/\lambda^2$

Figure 2.2: Fisher information.

#

Remark. With knowledge of the Fréchet-Cramér-Rao lower bound, we compare the estimator's variance to it:

- If they are equal, there is no other unbiased estimator with a lower variance, so the estimator is the most efficient.
- If they are different: the ratio

$$\frac{[n\mathcal{I}(\theta)]^{-1}}{Var(T)}$$

provides an indication of the relative efficiency of our estimator relative to the hypothetical estimator with variance equal to the lower bound.

- The efficiency concept is linked to unbiasedness.
- What if we want to compare estimators that are biased?

Definition 2.9 (Mean Squared Error)**.** Let $T = T(X_1, X_2, \ldots, X_n)$ be an estimator for θ. Then, the mean squared error (MSE) is

$$MSE(T) = E\left[(T-\theta)^2\right] = Var(T) + (\underbrace{E(T)-\theta}_{bias(T)})^2.$$

Remark. Regarding MSE:

- The MSE balances variance and bias.
- For unbiased estimators, MSE means variance.
- The estimator T_1 is "better" than T_2 if

$$MSE(T_1) < MSE(T_2),\ \forall\theta \in \Theta.$$

- The estimator T_1 is the "best" estimator if its MSE is lower or equal to the MSE of any other estimator for θ.

- In general, the MSE depends on θ.

2.4.2 Consitency

Definition 2.10 (Consistent estimator)**.** The estimator $T_n = T(X_1, X_2, \dots, X_n)$ is said to be mean square consistent if

$$\lim_{n \to +\infty} E[(T - \theta)^2] = 0, \quad \forall \theta \in \Theta.$$

Remark. **Necessary and Sufficient** condition for the estimator T_n to be mean square consistent:

$$\lim_{n \to +\infty} E(T_n) = \theta \quad \textit{and} \quad \lim_{n \to +\infty} Var(T_n) = 0.$$

Definition 2.11 (Weak consistent estimator)**.** The estimator $T_n = T(X_1, X_2, \dots, X_n)$ is said to be weak consistent if

$$\forall \epsilon > 0, \quad \lim_{n \to +\infty} P(\theta - \epsilon < T_n < \theta + \epsilon) = 1, \quad \forall \theta \in \Theta.$$

Remark. Mean squared consistency $\Longrightarrow$ weak consistency.

Remark. **Properties of moment estimators**:

- Are not unique, since they can be obtained from moments of different orders.
- Can lead to estimators that are not admissible.
- Do not enjoy the invariance property.
- Under general conditions:
  - are consistent;
  - for large samples, their distribution is approximately normal.

Remark. **Properties of maximum likelihood estimators**:

- Under general conditions are consistent.
- In general, they are not necessarily unbiased.
- If there is an estimator that attains the Fréchet-Cramér-Rao lower bound, then it is a maximum likelihood estimator.
- Under general conditions are asymptotically Normal. In case there is only one parameter θ, we have:

$$\sqrt{n\mathcal{I}(\theta)}(\hat{\theta} - \theta) \overset{a}{\sim} \mathcal{N}(0, 1).$$

Exercise 2.12. Let

$$T = \frac{(n-1)X_1 + X_n}{n}$$

be an estimator of μ, where $(X_1, \dots, X_n)$ represents a random sample of size n from a normal population with mean μ and variance σ^2. Check if T is unbiased to μ and if it is a consistent mean-square estimator for μ?

```
#
```

```
#
```

Chapter 3

Interval estimation

3.1 Confidence intervals

- It is usually of interest to obtain a measure of the distance between the point estimator and the unknown parameter.
- Instead of considering an isolated estimate $\hat{\theta}$ for θ, we propose an interval (t_1, t_2) to which a certain "level of confidence" is associated.
- In several cases, the interval is of the form

$$\left(\hat{\theta} - \delta, \hat{\theta} + \delta\right),$$

where δ can be seen as a precision measure or error measure of the point estimate $\hat{\theta}$.

Definition 3.1 (Random interval)**.** Let $T_1 = T_1(X_1, X_2, \ldots, X_n)$ and $T_2 = T_2(X_1, X_2, \ldots, X_n)$, $T_1 < T_2$, be two statistics such that

$$P(T_1 < \theta < T_2) = 1 - \alpha, \quad 0 < \alpha < 1, \quad \forall \theta \in \Theta.$$

Then, (T_1, T_2) is a **random interval** for θ with probability $1 - \alpha$.

Definition 3.2 (Confidence interval)**.** Let $(x_1, x_2, \ldots, x_n)$ be a particular observed sample of $(X_1, X_2, \ldots, X_n)$ and $t_1 = T_1(x_1, x_2, \ldots, x_n)$, $t_2 = T_2(x_1, x_2, \ldots, x_n)$ be the observed values of T_1 and T_2 for that realized sample.

Then (t_1, t_2) is **a confidence interval** at the $(1 - \alpha)100\%$ level of confidence for θ.

3.2 Pivotal quantities

Definition 3.3 (Pivotal Quantity)**.** A **pivotal quantity**, $Z(X_1, X_2, \ldots, X_n, \theta)$:

- is a function of the random sample;
- is a function of the parameter θ;
- has known p.d.f./p.m.f. $g(z)$, independent from θ;
- is independent from any other unknown parameter.

Definition 3.4 (Obtaining a confidence interval)**.**

1. Find the appropriate pivotal quantity Z.
2. Given the confidence level $(1 - \alpha)100\%$, find two values in the domain of Z, $z_1(\alpha)$ and $z_2(\alpha)$, such that

$$P\left(z_1(\alpha) < Z < z_2(\alpha)\right) = 1 - \alpha.$$

3. From $z_1(\alpha) < Z < z_2(\alpha)$ obtain $T_1(X_1, X_2, \ldots, X_n) < \theta < T_2(X_1, X_2, \ldots, X_n)$, that is,

$$P\left(z_1(\alpha) < Z < z_2(\alpha)\right) = P\left(T_1(X_1, X_2, \dots, X_n) < \theta < T_2(X_1, X_2, \dots, X_n)\right) = 1 - \alpha.$$

4. Random interval with probability $1 - \alpha$: $(T_1(X_1, X_2, \dots, X_n), T_2(X_1, X_2, \dots, X_n)) = (T_1, T_2)$.
5. Finally, **a confidence interval** at the $(1 - \alpha)100\%$ level for θ is given by the realization of the random interval to an observed sample:

$$CI_{(1-\alpha)100\%}(\theta) = (T_1(x_1, x_2, \dots, x_n), T_2(x_1, x_2, \dots, x_n)) = (t_1, t_2).$$

Remark. In general, we only apply points 1 and 5. It is also important to notice that:

- The definitions were presented for θ, but the generalization for $h(\theta)$ is straightforward.
- A confidence interval is simply the particular realization of the random interval, in the very same way as going from the estimator to the estimate in the context of point estimation.
- Thus, we only assign probability to the random interval, not to the confidence interval.
- The concept of confidence interval can be generalized to more dimensions $(\theta_1, \theta_2, \dots, \theta_k)$, $k > 1$, in which case we obtain confidence regions.

Exercise 3.1. With the pivotal quantity $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, show that a confidence interval at 95% level for the mean of a population with known σ is

$$\mathrm{CI}_{95\%}(\mu) = \left(\overline{x} \mp z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\right) = \left(\overline{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\,,\ \overline{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\right).$$

#

#

3.3 Confidence intervals for normal populations

Definition 3.5. Confidence interval for the **mean** with **known variance**:

- Pivotal quantity:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

- Confidence interval:

$$CI_{(1-\alpha)100\%}(\mu) = \left(\overline{x} \mp z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right), \qquad \Phi(z_{\alpha/2}) = 1 - \alpha/2.$$

Definition 3.6. Confidence interval for the **mean** with **unknown variance**:

- Pivotal quantity:
$$T = \frac{\overline{X} - \mu}{S'/\sqrt{n}} \sim t(n-1).$$
- Confidence interval:
$$CI_{(1-\alpha)100\%}(\mu) = \left(\overline{x} \mp t_{\alpha/2} \frac{s'}{\sqrt{n}} \right), \qquad P(T > t_{\alpha/2}) = \alpha/2.$$

Definition 3.7. Confidence interval for the **variance**:

- Pivotal quantity:
$$Q = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2(n-1).$$
- Confidence interval:
$$CI_{(1-\alpha)100\%}(\sigma^2) = \left(\frac{(n-1)s'^2}{q_2}, \frac{(n-1)s'^2}{q_1} \right), \quad P(Q < q_1) = P(Q > q_2) = \alpha/2.$$

We now consider two normal populations $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and two independent random samples $(X_{11}, X_{12}, \ldots, X_{1m})$ and $(X_{21}, X_{22}, \ldots, X_{2n})$.

Definition 3.8. Confidence interval for the difference of **means** with **known variances**:

- Pivotal quantity:
$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} \sim \mathcal{N}(0,1).$$
- Confidence interval:
$$CI_{(1-\alpha)100\%}(\mu_1 - \mu_2) = \left(\overline{x}_1 - \overline{x}_2 \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right).$$

Definition 3.9. Confidence interval for the difference of **means** with **unknown (but equal) variances**:

- Pivotal quantity:
$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{1}{m} + \dfrac{1}{n}} \sqrt{\frac{(m-1)S_1'^2 + (n-1)S_2'^2}{m+n-2}}} \sim t(m+n-2).$$
- Confidence interval:
$$CI_{(1-\alpha)100\%}(\mu_1 - \mu_2) = \left(\overline{x}_1 - \overline{x}_2 \mp t_{\alpha/2} \sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)s_1'^2 + (n-1)s_2'^2}{m+n-2}} \right).$$

Definition 3.10. Confidence interval for the difference of **means** with **unknown (possibly different) variances**:

- Pivotal quantity:
$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1'^2}{m} + \dfrac{S_2'^2}{n}}} \overset{\text{a}}{\sim} t(r),$$
where r is the largest integer contained in

$$\frac{\left(\frac{s_1'^2}{m}+\frac{s_2'^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1'^2}{m}\right)^2+\frac{1}{n-1}\left(\frac{s_1'^2}{n}\right)^2}.$$

- Confidence interval:

$$CI_{(1-\alpha)100\%}(\mu_1-\mu_2)=\left(\overline{x}_1-\overline{x}_2\mp t_{\alpha/2}\sqrt{\frac{s_1'^2}{m}+\frac{s_2'^2}{n}}\right).$$

Definition 3.11. Confidence interval for the **ratio of variances** σ_2^2/σ_1^2:

- Pivotal quantity:

$$F=\frac{S_1'^2}{S_2'^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F(m-1,n-1).$$

- Confidence interval:

$$CI_{(1-\alpha)100\%}(\sigma_2^2/\sigma_1^2)=\left(f_1\frac{s_2'^2}{s_1'^2},f_2\frac{s_2'^2}{s_1'^2}\right),\quad P(F<f_1)=P(F>f_2)=\alpha/2.$$

Remark. Using the tables, f_2 is "directly" obtained, while f_1 is obtained as follows:

- go to table $F(n-1,m-1)$ and find the value f with associated right-tail probability $\alpha/2$;
- do $f_1=\frac{1}{f}$.

Exercise 3.2. Consider a population with a normal distribution of unknown parameters. From this population, a random sample of dimension 25 was taken. Suppose that the sample provided the following results:

$$\sum_{i=1}^{25}x_i=75\text{ and }\sum_{i=1}^{25}x_i^2=321.$$

(a) Construct a 95% confidence interval for the mean.

```
#
```

```
#
```

(b) Construct a 95% confidence interval for the standard deviation.

```
#

#
```

Exercise 3.3. Based on a casual sample of 16 observations taken from a normal population, the following confidence interval for the expected value was constructed according to the usual process: $(7.398,\ 12.602)$.

(a) Knowing that, with the sample information, $s = 3.872$ was obtained, what is the degree of confidence that you can assign to the above confidence interval?

```
#

#
```

(b) Based on the same sample, construct a confidence interval of 95% for the population variance.

```
#
```

```







#
```

3.4 Confidence intervals for large samples

When n is large ($n > 30$), use the CLT to obtain asymptotic intervals, which are approximate and valid. Consider that all of the aforementioned pivot quantities follow a standard normal distribution asymptotically.

Exercise 3.4. Suppose that the annual expenditure on consumer goods of the inhabitants of cities A and B is associated with random variables with a normal distribution of unknown parameters. 100 people from the city A were surveyed, resulting in an average annual expenditure of 5100 monetary units (m.u.) and a corrected standard deviation of 1020 m.u. 125 people from the city B were also surveyed, resulting in an average of 6150 m.u. with a corrected standard deviation of 1300 m.u. Using a confidence interval of 95%, check whether it is possible to state that the average annual expenditure on consumer goods is higher in either of the two cities.

```
#




















#
```

Chapter 4

Parametric hypothesis testing

4.1 Introduction

Idea

- To establish a conjecture on the unknown aspects of the (population) distribution.
- Check if the existing information in the observed sample $(x_1, \dots, x_n)$ supports or does not support this conjecture.

Definition 4.1 (Statistical hypothesis)**.** Any conjecture about unknown aspects of the distribution of X.

Example 4.1. $X \sim exp(\lambda = 2)$?, $X \sim \mathcal{N}(\mu, \sigma^2)$?

Definition 4.2 (Non parametric hypothesis)**.** The hypothesis is about the distribution of the population X.

Example 4.2. $X \sim F(\cdot)$?, $X \sim \mathcal{N}(\cdot, \cdot)$?

Definition 4.3 (Parametric hypothesis)**.** The hypothesis is about parameters of the distribution of X. In this case, the functional form of the distribution function of X is known.

Example 4.3. $X \sim N(\mu = 1, \sigma^2)$?, $X \sim exp(\lambda = 2)$?

Remark. In this curricular unit, we only study parametric hypothesis tests.

Exercise 4.1. For each of the following propositions, indicate whether or not it is a statistical hypothesis:

(a) $\mu = 3$.

(b) $\overline{x} = 4$.

(c) $P(X < 2.5) = 0.4$.

(d) $2 < \sigma < 3$.

(e) $\overline{X} < 3$;

```
#









#
```

4.2 Hypothesis testing

Definition 4.4 (Null and alternative hypothesis)**.** Assume $X \sim f(x \mid \theta)$, $\theta \in \Theta$, θ unknown.

- Null hypothesis (**typically, it corresponds to what we suspect to be true**):

$$H_0 : \theta \in \Theta_0.$$

- Alternative hypothesis:

$$H_1 : \theta \in \Theta_1.$$

Every parametric hypothesis divides the parametric space Θ into Θ_0 and Θ_1, i.e.,

$$\Theta = \Theta_0 \cup \Theta_1 \qquad \text{and} \qquad \Theta_0 \cap \Theta_1 = \varnothing.$$

Definition 4.5 (Simple statistical hypothesis)**.** When the parametric subspace contains only one element.

Definition 4.6 (Composite statistical hypothesis)**.** When the parametric subspace contains more than one element.

Example 4.4. We wish to assess whether a given coin is fair: $X \sim B(1, \theta)$ and $\theta = P(\text{“}success\text{”})$

$$X = \begin{cases} 1, & \text{heads} \\ 0, & \text{tails} \end{cases}$$

- Parametric space: $\theta \in \Theta = [0, 1]$.
- Null hypothesis: $H_0 : \ \theta = 0.5$ (simple hypothesis).
- Alternative hypothesis: $H_1 : \ \theta \neq 0.5$ (composite hypothesis).

Definition 4.7 (Statistical hypothesis test)**.** It is a **rule** allowing to specify a subset of the parametric space (sample results space) $W \subset \mathbb{R}^n$ such that:

- if $(x_1, \ldots, x_n) \in W \Longrightarrow$ we reject H_0;
- if $(x_1, \ldots, x_n) \notin W \Longrightarrow$ we do not reject H_0.

The final decision always refers to H_0 (reject H_0 or do not reject H_0).

The statistical hypothesis test defines a partition of the sample space into two regions, W and $\overline{W}$

$$W \cup \overline{W} = \mathbb{R}^n \qquad \text{and} \qquad W \cap \overline{W} = \varnothing,$$

where W denotes the **rejection region** (RR) or **critical region** (CR).

Definition 4.8 (Test statistic)**.** Alternatively, and in almost all cases of practical interest, we work with a **test statistic** (TS):

$$T = T(X_1, \ldots, X_n) \Longrightarrow t_{obs} = T(x_1, \ldots, x_n).$$

In this case, the RR, W, is defined by means of the TS:

- if $t_{obs} \in W_T \Longrightarrow$ we reject H_0;
- if $t_{obs} \notin W_T \Longrightarrow$ we do not reject H_0.

In summary, the **components of a statistical hypothesis test** are:

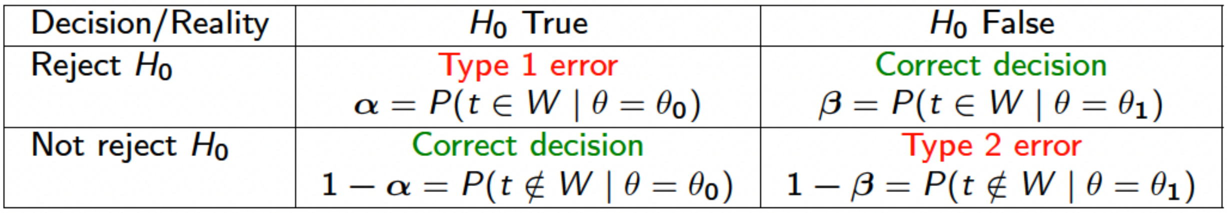
- Null hypothesis, H_0: kept unless evidence shows otherwise;
- Alternative hypothesis, H_1: adopted if H_0 is rejected;
- Test statistic, $T = T(X_1, \ldots, X_n)$: based on which the decision rule will be made;
- Rejection region (RR), W_T: the decision rule.

Types of error

- The hypothesis test is carried out based on a random sample (we do not have access to the whole population).
- Thus, the decision of rejecting, or not, the null hypothesis may be wrong!
- We must consider two types of error:
  - Type 1 error: rejecting H_0 when H_0 is true.
  - Type 2 error: not rejecting H_0 when H_0 is false.
- We will split our study into 3 cases: simple hypothesis *vs* simple hypothesis, simple hypothesis *vs* composite hypothesis and bilateral tests.

4.3 Simple hypothesis *vs* simple hypothesis

$$\boldsymbol{H_0 : \theta = \theta_0} \quad \text{versus} \quad \boldsymbol{H_1 : \theta = \theta_1}$$

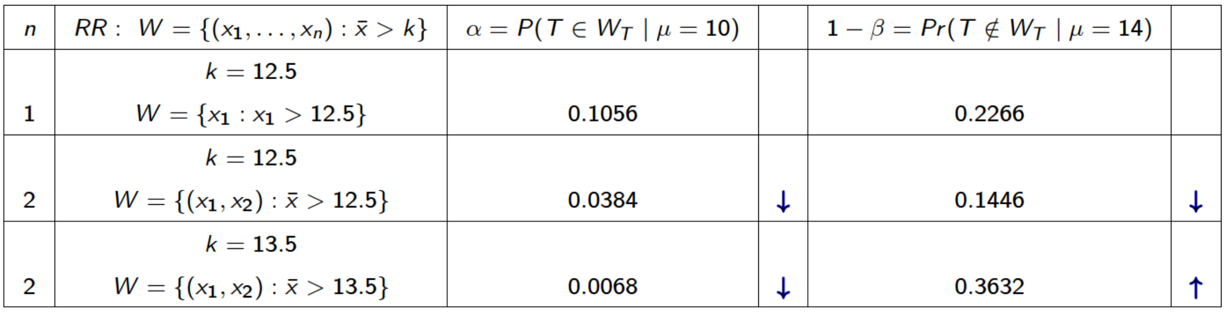


Decision/Reality	H_0 True	H_0 False
Reject H_0	Type 1 error $\alpha = P(t \in W \mid \theta = \theta_0)$	Correct decision $\beta = P(t \in W \mid \theta = \theta_1)$
Not reject H_0	Correct decision $1 - \alpha = P(t \notin W \mid \theta = \theta_0)$	Type 2 error $1 - \beta = P(t \notin W \mid \theta = \theta_1)$

- **Test dimension**: $\alpha = P(\text{reject } H_0 \mid H_0 \text{ true}) = P(\text{reject } H_0 \mid \theta = \theta_0)$.
- **Test power**: $\beta = P(\text{reject } H_0 \mid H_0 \text{ false}) = P(\text{reject } H_0 \mid \theta = \theta_1)$.
- Ideal: **lowest value of α and highest value of β.**
- The reduction of the two error probabilities (or of one of them fixing the other) can only be achieved by increasing the sample size (see eg. ahead).
- By changing the RR, other values for α and β are obtained (see eg. ahead).
- Since it is impossible to minimize both types of error simultaneously, we resort to the Neyman-Pearson lemma in order to obtain the most powerful test (omitted topic).

Example 4.5.

$$X \sim \mathcal{N}(\mu, \sigma^2 = 4), \qquad H_0 : \mu = 10 \quad vs \quad H_1 : \mu = 14 :$$



n	$RR: \ W = \{(x_1, \ldots, x_n) : \bar{x} > k\}$	$\alpha = P(T \in W_T \mid \mu = 10)$		$1 - \beta = Pr(T \notin W_T \mid \mu = 14)$	
1	$k = 12.5$ $W = \{x_1 : x_1 > 12.5\}$	0.1056		0.2266	
2	$k = 12.5$ $W = \{(x_1, x_2) : \bar{x} > 12.5\}$	0.0384	↓	0.1446	↓
2	$k = 13.5$ $W = \{(x_1, x_2) : \bar{x} > 13.5\}$	0.0068	↓	0.3632	↑

Exercise 4.2. The duration, in hours, X, of a certain type of component has a normal distribution with a standard deviation equal to 50. To test:

$$H_0 : \ \mu = 250 \ \text{ vs } \ H_1 : \mu = 200,$$

the rule is: reject H_0 if $\overline{x} < 230$.

(a) If the decision is made based on a random sample of 16 components, calculate the dimension and power associated with this test.

```
#






























#
```

(b) What is the minimum sample size so that the probability of committing the first type of error is less than 0.025?

```
#

















#
```

4.4 Simple hypothesis *vs* composite hypothesis

$$\boldsymbol{H_0 : \theta = \theta_0} \quad \text{versus} \quad \boldsymbol{H_1 : \theta > \theta_1} \quad \text{or} \quad \boldsymbol{H_1 : \theta < \theta_1}$$

- Type I error probability does not change (the significance level depends only on H_0).
- Type II error probability β is now a **function** (of all possible values of the parameter in the alternative).

Definition 4.9 (Power function)**.** Test $H_0 : \ \theta = \theta_0$ *vs* $H_1 : \ \theta > \theta_0$, with RR W.

$$\beta(\theta) = P(\text{rej. } H_0 \mid H_0 \text{ false}) = P(\text{rej.} H_0 \mid \theta > \theta_0) = P((X_1, \ldots, X_n) \in W \mid \theta),$$

for $\theta \in \Theta_1 = \{\theta : \theta > \theta_0\}$.

Note: the case $H_1 : \theta < \theta_0$ is similar with the necessary adaptations.

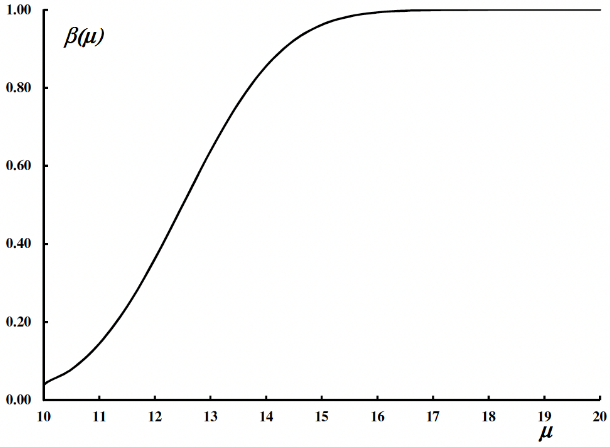


Figure 4.1: Example of a power function.

Remark. Important remarks:

- When $H_0 : \theta \leqslant \theta_0$ *vs* $H_1 : \theta > \theta_0$ (composite hyp. against unilateral composite alternative) we should proceed as if we had $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta > \theta_0$.
- Similarly, for $H_0 : \theta \geqslant \theta_0$ *vs* $H_1 : \theta < \theta_0$, we should proceed as if we had $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta < \theta_0$.
- In both cases, we are choosing the worst-case scenario.

Exercise 4.3. Let X be a random variable that represents the amount of wine in a 75-centiliter bottle. Assume that X has a normal distribution with a standard deviation of 2. To test:

$$H_0 : \mu = 75 \text{ vs } H_1 : \mu < 75,$$

a random sample of 10 bottles was selected, rejecting the null hypothesis if $\overline{x} < 74.1$, where $\overline{x}$ is the average amount of wine per bottle in the observed sample.

(a) Compute the dimension of this test.

#

#

(b) Determine the power function and calculate its value when $\mu = 74$ and $\mu = 72.5$.

#

```
#
```

4.5 Bilateral tests

$$\boldsymbol{H_0 : \theta = \theta_0} \quad \text{versus} \quad \boldsymbol{H_1 : \theta \neq \theta_0}$$

- Define the RR on both tails of the test statistic distribution, assigning equal probability $\alpha/2$ to each side.

4.6 The p-value

- If α is fixed, the test result is whether H_0 is rejected or not.
- In that case, we do not take into account whether the observed value of the test statistic is close or not to the critical value at that significance level.
- The p-value is an **alternative** way of reporting the test result that overcomes this “limitation”.

Definition 4.10. Let $T(x_1, \ldots, x_n) = t_{obs}$ be the observed value of the test statistic. The $p-$value:

- is a tool to check if the test statistic is in the rejection region. It is also a measure of the evidence for rejecting H_0.
- the smaller its value, the smaller the consistency of the data with the hypothesis (“the more it is rejected” H_0);
- Rejection rule: $p - value < \alpha \implies$ reject H_0.

4.7 Normal populations: testing mean and variance

Definition 4.11. Testing the mean with known variance

- **Bilateral test**

$$H_0: \ \mu = \mu_0 \quad \text{vs} \quad H_1: \ \mu \neq \mu_0$$

$$TS: \ Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$p - value = 2P(Z \geq |z_{obs}| \mid H_0)$$

$$RR = \{z : |z| > z_{\alpha/2}\} \text{ or } RR = \left\{\overline{x} : \overline{x} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} \cup \left\{\overline{x} : \overline{x} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

- **Right tailed test**

$$H_0: \ \mu = \mu_0 \quad \text{vs} \quad H_1: \ \mu > \mu_0$$

$$TS:\ Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \geq z_{obs} \mid H_0)$$

$$RR = \{z : z > z_\alpha\} \text{ or } RR = \left\{\overline{x} : \overline{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$$

- **Left tailed test**

$$H_0:\ \mu = \mu_0 \quad \text{vs} \quad H_1:\ \mu < \mu_0$$

$$TS:\ Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \leq z_{obs} \mid H_0)$$

$$RR = \{z : z < -z_\alpha\} \text{ or } RR = \left\{\overline{x} : \overline{x} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\right\}$$

Definition 4.12. Testing the mean with unknown variance

- **Bilateral test**

$$H_0:\ \mu = \mu_0 \quad \text{vs} \quad H_1:\ \mu \neq \mu_0$$

$$TS:\ T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

$$p - value = 2P(T \geq |t_{obs}| \mid H_0)$$

$$RR = \{t : |t| > t_{\alpha/2}\} \text{ or } RR = \left\{\overline{x} : \overline{x} > \mu_0 + t_{\alpha/2} \frac{s'}{\sqrt{n}}\right\} \cup \left\{\overline{x} : \overline{x} < \mu_0 - t_{\alpha/2} \frac{s'}{\sqrt{n}}\right\}$$

- **Right tailed test**

$$H_0:\ \mu = \mu_0 \quad \text{vs} \quad H_1:\ \mu > \mu_0$$

$$TS:\ T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

$$p - value = P(T \geq t_{obs} \mid H_0)$$

$$RR = \{t : t > t_\alpha\} \text{ or } RR = \left\{\overline{x} : \overline{x} > \mu_0 + t_\alpha \frac{s'}{\sqrt{n}}\right\}$$

- **Left tailed test**

$$H_0:\ \mu = \mu_0 \quad \text{vs} \quad H_1:\ \mu < \mu_0$$

$$TS:\ T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} \sim t(n-1)$$

$$p - value = P(T \leq t_{obs} \mid H_0)$$

$$RR = \{t : t < -t_\alpha\} \text{ or } RR = \left\{\overline{x} : \overline{x} < \mu_0 - t_\alpha \frac{s'}{\sqrt{n}}\right\}$$

Definition 4.13. Testing the variance

- **Bilateral test**

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

$$TS : Q = \frac{(n-1)S'^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$p - value = 2\min\{p_1, p_2\} \quad \text{(see next definition)}$$

$$RR = \{q : q < q_{1-\alpha/2}\} \cup \{q : q > q_{\alpha/2}\} \text{ or } RR = \left\{s'^2 : s'^2 < \frac{q_{1-\alpha/2}\sigma_0^2}{n-1}\right\} \cup \left\{s'^2 : s'^2 > \frac{q_{\alpha/2}\sigma_0^2}{n-1}\right\}$$

- **Right tailed test**

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 > \sigma_0^2$$

$$TS : Q = \frac{(n-1)S'^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$p - value = p_1 = P(Q \geq q_{obs} \mid H_0)$$

$$RR = \{q : q > q_\alpha\} \cup \text{ or } RR = \left\{s'^2 : s'^2 > \frac{q_\alpha \sigma_0^2}{n-1}\right\}$$

- **Left tailed test**

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 < \sigma_0^2$$

$$TS : Q = \frac{(n-1)S'^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$p - value = p_2 = P(Q \leq q_{obs} \mid H_0)$$

$$RR = \{q : q < q_{1-\alpha}\} \cup \text{ or } RR = \left\{s'^2 : s'^2 < \frac{q_{1-\alpha}\sigma_0^2}{n-1}\right\}$$

We now consider two normal populations $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, and two independent random samples $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$.

Definition 4.14. Testing the equality of means with known variances

- **Bilateral test**

$$H_0: \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0 \quad \text{vs} \quad H_1: \mu_X \neq \mu_Y \iff \mu_X - \mu_Y \neq 0$$

$$TS: Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1)$$
$$p - value = 2P(Z \geq |z_{obs}| \mid H_0)$$

$$RR = \{z : |z| > z_{\alpha/2}\} \text{ ou } RR = \left\{ |\overline{x} - \overline{y}| : |\overline{x} - \overline{y}| > z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \right\}$$

- **Right tailed test**

$$H_0: \mu_X = \mu_Y \quad \text{vs} \quad H_1: \mu_X > \mu_Y$$

$$TS: Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \geq z_{obs} \mid H_0)$$

$$RR = \{z : z > z_\alpha\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} > z_\alpha\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \right\}$$

- **Left tailed test**

$$H_0: \mu_X = \mu_Y \quad \text{vs} \quad H_1: \mu_X < \mu_Y$$

$$TS: Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim \mathcal{N}(0,1)$$

$$p - value = P(Z \leq z_{obs} \mid H_0)$$

$$RR = \{z : z < -z_\alpha\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} < -z_\alpha\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}} \right\}$$

Definition 4.15. Testing the equality of means with unknown (but equal) variance

- **Bilateral test**

$$H_0: \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0 \quad \text{vs} \quad H_1: \mu_X \neq \mu_Y \iff \mu_X - \mu_Y \neq 0$$

$$TS: T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2)$$
$$p - value = 2P(T \geq |t_{obs}| \mid H_0)$$

$$RR = \{t : |t| > t_{\alpha/2}\} \text{ ou } RR = \left\{ |\overline{x} - \overline{y}| : |\overline{x} - \overline{y}| > t_{\alpha/2}\sqrt{\frac{1}{m} + \frac{1}{n}}\sqrt{\frac{(m-1)s_X'^2 + (n-1)s_Y'^2}{m+n-2}} \right\}$$

- **Right tailed test**

$$H_0: \ \mu_X = \mu_Y \quad \text{vs} \quad H_1: \ \mu_X > \mu_Y$$

$$TS: \ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2)$$

$$p - value = P(T \geq t_{obs} \mid H_0)$$

$$RR = \{t : t > t_\alpha\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} > t_\alpha \sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)s_X'^2 + (n-1)s_Y'^2}{m+n-2}} \right\}$$

- **Left tailed test**

$$H_0: \ \mu_X = \mu_Y \quad \text{vs} \quad H_1: \ \mu_X < \mu_Y$$

$$TS: \ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2)$$

$$p - value = P(T \leq t_{obs} \mid H_0)$$

$$RR = \{t : t < -t_\alpha\} \text{ ou } RR = \left\{ \overline{x} - \overline{y} : \overline{x} - \overline{y} < -t_\alpha \sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)s_X'^2 + (n-1)s_Y'^2}{m+n-2}} \right\}$$

Definition 4.16. Testing the equality of means with unknown (possibly different) variance

- **Bilateral test**

$$H_0: \ \mu_X = \mu_Y \iff \mu_X - \mu_Y = 0 \quad \text{vs} \quad H_1: \ \mu_X \neq \mu_Y \iff \mu_X - \mu_Y \neq 0$$

$$TS: \ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X'^2}{m} + \frac{S_Y'^2}{n}}} \sim t(r),$$

where r is the largest integer contained in

$$\frac{\left(\frac{s_X'^2}{m} + \frac{s_Y'^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_X'^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s_Y'^2}{n}\right)^2}.$$

$$p - value = 2P(T \geq |t_{obs}| \mid H_0)$$

$$RR = \{t : |t| > t_{\alpha/2}\} \text{ ou } RR = \left\{ |\overline{x} - \overline{y}| : |\overline{x} - \overline{y}| > t_{\alpha/2} \sqrt{\frac{s_X'^2}{m} + \frac{s_Y'^2}{n}} \right\}$$

- **Right tailed test**

$$H_0:\ \mu_X = \mu_Y \quad \text{vs} \quad H_1:\ \mu_X > \mu_Y$$

$$TS:\ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X'^2}{m} + \dfrac{S_Y'^2}{n}}} \sim t(r),$$

where r is the largest integer contained in

$$\frac{\left(\dfrac{s_X'^2}{m} + \dfrac{s_Y'^2}{n}\right)^2}{\dfrac{1}{m-1}\left(\dfrac{s_X'^2}{m}\right)^2 + \dfrac{1}{n-1}\left(\dfrac{s_Y'^2}{n}\right)^2}.$$

$$p - value = P(T \geq t_{obs} \mid H_0)$$

$$RR = \{t : t > t_\alpha\} \text{ ou } RR = \left\{\overline{x} - \overline{y} : \overline{x} - \overline{y} > t_\alpha \sqrt{\frac{s_X'^2}{m} + \frac{s_Y'^2}{n}}\right\}$$

- **Left tailed test**

$$H_0:\ \mu_X = \mu_Y \quad \text{vs} \quad H_1:\ \mu_X < \mu_Y$$

$$TS:\ T = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_X'^2}{m} + \dfrac{S_Y'^2}{n}}} \sim t(r),$$

where r is the largest integer contained in

$$\frac{\left(\dfrac{s_X'^2}{m} + \dfrac{s_Y'^2}{n}\right)^2}{\dfrac{1}{m-1}\left(\dfrac{s_X'^2}{m}\right)^2 + \dfrac{1}{n-1}\left(\dfrac{s_Y'^2}{n}\right)^2}.$$

$$p - value = P(T \leq t_{obs} \mid H_0)$$

$$RR = \{t : t < -t_\alpha\} \text{ ou } RR = \left\{\overline{x} - \overline{y} : \overline{x} - \overline{y} < -t_\alpha \sqrt{\frac{s_X'^2}{m} + \frac{s_Y'^2}{n}}\right\}$$

Definition 4.17. Testing the ratio of variances

- **Bilateral test**

$$H_0:\ \sigma_X^2 = \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 = 1 \text{ vs } H_1:\ \sigma_X^2 \neq \sigma_Y^2 \iff \sigma_X^2/\sigma_Y^2 \neq 1$$

$$TS:\ F = \frac{S_X'^2}{S_Y'^2} \sim F(m-1, n-1)$$

$$p - value = 2\min\{p_1, p_2\} \quad \text{(see next definition)}$$

$$RR = \left\{F : F < 1/F^*_{\alpha/2}\right\} \cup \left\{F : F > F_{\alpha/2}\right\} \text{ or}$$

$$RR = \left\{s'^2_X/s'^2_Y : s'^2_X/s'^2_Y < 1/F^*_{\alpha/2}\right\} \cup \left\{s'^2_X/s'^2_Y : s'^2_X/s'^2_Y > F_{\alpha/2}\right\} \text{ where}$$

$$F_{\alpha/2} : P(F > F_{\alpha/2}) = \alpha/2,\ F^*_{\alpha/2} : P(1/F > F^*_{\alpha/2}) = \alpha/2$$

- **Right tailed test**

$$H_0 :\ \sigma^2_X = \sigma^2_Y \iff \sigma^2_X/\sigma^2_Y = 1 \text{ vs } H_1 :\ \sigma^2_X > \sigma^2_Y \iff \sigma^2_X/\sigma^2_Y > 1$$

$$TS :\ F = \frac{S'^2_X}{S'^2_Y} \sim F(m-1, n-1)$$

$$p - value = p_1 = P(F \geq F_{obs} \mid H_0)$$

$$RR = \{F : F > F_\alpha\} \text{ or } RR = \left\{s'^2_X/s'^2_Y : s'^2_X/s'^2_Y > F_\alpha\right\}$$

- **Left tailed test**

$$H_0 :\ \sigma^2_X = \sigma^2_Y \iff \sigma^2_X/\sigma^2_Y = 1 \text{ vs } H_1 :\ \sigma^2_X < \sigma^2_Y \iff \sigma^2_X/\sigma^2_Y < 1$$

$$TS :\ F = \frac{S'^2_X}{S'^2_Y} \sim F(m-1, n-1)$$

$$p - value = p_2 = P(F \leq F_{obs} \mid H_0)$$

$$RR = \{F : F < 1/F^*_\alpha\} \text{ or } RR = \left\{s'^2_X/s'^2_Y : s'^2_X/s'^2_Y < 1/F^*_\alpha\right\}$$

Exercise 4.4. A certain wine producer assures the inspection authorities that his wine has an average acidity content that does not exceed 0.5 g/l. It is assumed that the acidity content is a random variable with a normal distribution of unknown parameters.

(a) Based on a sample of dimension n, formalize a statistical test that allows you to analyse the veracity of the producer's statement.

```
#















#
```

(b) Observing a sample of 20 bottles, an average of 0.7 g/l and a corrected standard deviation of 0.08 were obtained. Should the supervisory authorities act against the producer? Justify your answer using a suitable hypothesis test.

```
#



#
```

Exercise 4.5. A tax office has two employees who receive tax returns. Assume that the time it takes each employee to serve a person has a normal distribution, with standard deviations equal to 2 minutes. Mr. Diogo Costa, arriving to hand in his statement, notices that the queue next to counter A has 20 people, while the queue next to counter B has 15 people, and he naturally opts for this one. When he starts to be served (an hour and fifteen minutes later), he notices that the twenty-first person in the next row has just been served. Can it be said that the average time spent by two employees serving a person is identical? (Consider dimensions 0.05 and 0.1.)

```
#



#
```

Exercise 4.6. To assess the quality of the environment in the two largest Portuguese cities, two random variables are considered: X and Y, which represent the number of particles suspended in the air (micrograms/m^3) in Lisbon and in Porto, respectively (the more particles in suspension, the worse the air quality). Assume that the two random variables have a normal distribution. The Ministry of the Environment collected two random samples: one of size 16 in Lisbon and another of size 13 in Porto. The observed results are as follows: $\overline{x} = 92.9$, $s'_X = 25.4$, $\overline{y} = 86.1$, $s'_Y = 28.1$.

(a) Based on an adequate statistical test, for $\alpha = 0.05$, show that the equality of variances of the two random variables is not rejected.

```
#











#
```

(b) Assuming equal variances in Lisbon and Porto, what is the p-value of the appropriate statistical test to assess whether the air quality is worse in the center of Lisbon than in the center of Porto?

```
#











#
```

4.8 Large populations

When n is large ($n > 30$), use the CLT to obtain asymptotic distributions to the TS, which are approximate and valid. Consider that all of the aforementioned pivot quantities follow a standard normal distribution asymptotically.

Exercise 4.7. In a car rental company, the main parameter for establishing the daily rental rate for light passenger vehicles in the unlimited mileage category is the average number of kilometres travelled daily, which, at the current rate, is assumed not to exceed 275. To assess whether it is necessary to revise this tariff, a random sample of 500 rentals in this category was collected, with an average of 278.5 and a corrected variance of 6430.5. Use a hypothesis test of size $\alpha = 0.05$ to assess whether it is necessary to revise the daily rate for this type of rental.

```
#









#
```

Chapter 5

Multiple variable regression analysis

5.1 Introduction

- Greene (Econometric Analysis, 1997) defines **Econometrics** as "the field of economics that concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory."
- Starting point: study, based on data, a certain phenomenon of economic nature. Examples: evolution of household consumption, GDP, debt, …

Definition 5.1. Theoretical model:

- the theory (common sense and/or intuition) leads to the construction of a theoretical model that is always an abstract representation (an approximation) of reality.
- this model establishes a relationship between variables.
- the models that will be studied consist of the analysis of the behavior of a dependent variable, z, as a function of other variables $w_1, w_2, \ldots, w_p$, named independent or explanatory variables $z = h(w_1, \ldots, w_p)$, where the relationship usually involves a set of parameters $(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$.
- is a function in the mathematical sense of the term: each value of the argument corresponds to a single value of the function.

Example 5.1. Consumption function, $consumption = f(income)$:

$$consumption = \alpha_1 + \alpha_2 income.$$

Example 5.2. Production function of a good (Cobb-Douglas), $Q = f(K, L)$, where K, L are factors of production, for example $K = capital\ invested$, $L = labor$:

$$Q = \alpha_1\, K^{\alpha_2}\, L^{\alpha_3}.$$

Example 5.3. Constant elasticity substitution production function (a much more complicated model):

$$Q = \beta\left[(1-\delta)L^{-\rho} + \delta K^{-\rho}\right]^{-\gamma/\rho}.$$

- We will only study models that involve a linear (or are able to be linearized) relationship in relation to the parameters because they cover a significant variety of situations and are easier to handle.
- A linear relation with respect to parameters $\beta_1, \ldots, \beta_k$ is defined as $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, regardless of whether the linear functional form is direct or could be linearized.
- Linearity with respect to parameters:

$$z = \alpha_1 + \alpha_2 w + \alpha_3 w^2$$

is linear with respect to the parameters α_i, but not with respect to the variable w.

- Linearity with respect to the variables:

$$z = \alpha_1 + \alpha_2 w_2 + \alpha_3^2 w_3$$

it is linear with respect to variables w_2, w_3, but it is not linear (nor linearizable) with respect to parameters α_i.

Example 5.4. Linear and linearizable relations:

- Linear relation with respect to parameters

$$consumption = \beta_1 + \beta_2 income.$$

- The relation

$$Q = \alpha_1 K^{\alpha_2} L^{\alpha_3}$$

allows a linearization by taking the logarithm

$$\ln Q = \ln \alpha_1 + \alpha_2 \ln K + \alpha_3 \ln L \iff \ln Q = A + \alpha_2 \ln K + \alpha_3 \ln L,$$

where $A = \ln \alpha_1$ is a constant.

- This is not a linear relationship, nor is it linearizable (we do not study these cases here)

$$Q = \beta \left[(1-\delta) L^{-\rho} + \delta K^{-\rho} \right]^{-\gamma/\rho}.$$

We will work with models defined as

$$z = h(w_1, w_2, \ldots, w_p)$$

or, after possible linearization (we always assume $x_1 = 1$)

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k.$$

This theoretical model is not a statistical model. Why?

Remark. Before converting the above model to a statistical model, it is important to address two questions:

1. What is the nature (random or deterministic) of the variables involved in the model? We will postulate that the model variables, as well as their observations, are random in nature.
2. Relational flexibility of the theoretical model: when considering

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k,$$

it is implied that the only explanatory factors of y are $x_1, x_2, \ldots, x_k$, which is generally an absurd assumption! The flexibility is obtained by introducing an additional variable u that **covers all factors that were not considered** and that can affect the behavior of y.

- Incorporating u in the model $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$ allows us to obtain

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u.$$

Note that u is not observable.

Definition 5.2 (Residual variable)**.** The variable u introduced above is called **residual variable** and represents everything that needs to be added to $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k$ in order to get y.

Example 5.5. The goal is to investigate the impact of education on wages. It is common knowledge that, in addition to education, other factors such as professional experience, sector of activity, and so on influence salary. The following variables are taken into account:

- the worker's annual average monthly salary: $salar$;
- the worker's number of years of education: $educ$;
- number of years of professional experience after finishing school: $exper$;
- number of years in current position: $empc$;
- binary variable that takes the value 1 if the individual is a woman and 0 if the individual is a man: $woman$.

One can propose an exponential function for the original model:

$$salar = e^{\alpha_1 + \alpha_2 educ + \alpha_3 exper + \alpha_4 empc + \alpha_5 woman}.$$

Linearizing, we get

$$\overbrace{\underbrace{\ln(salar)}_{y}}^{lsalar} = \underbrace{\alpha_1}_{\beta_1} + \underbrace{\alpha_2}_{\beta_2}\underbrace{educ}_{x_2} + \underbrace{\alpha_3}_{\beta_3}\underbrace{exper}_{x_3} + \underbrace{\alpha_4}_{\beta_4}\underbrace{empc}_{x_4} + \underbrace{\alpha_5}_{\beta_5}\underbrace{woman}_{x5}.$$

Introducing the residual variable:

$$lsalar = \alpha_1 + \alpha_2 educ + \alpha_3 exper + \alpha_4 empc + \alpha_5 woman + u.$$

5.2 Types of economic data

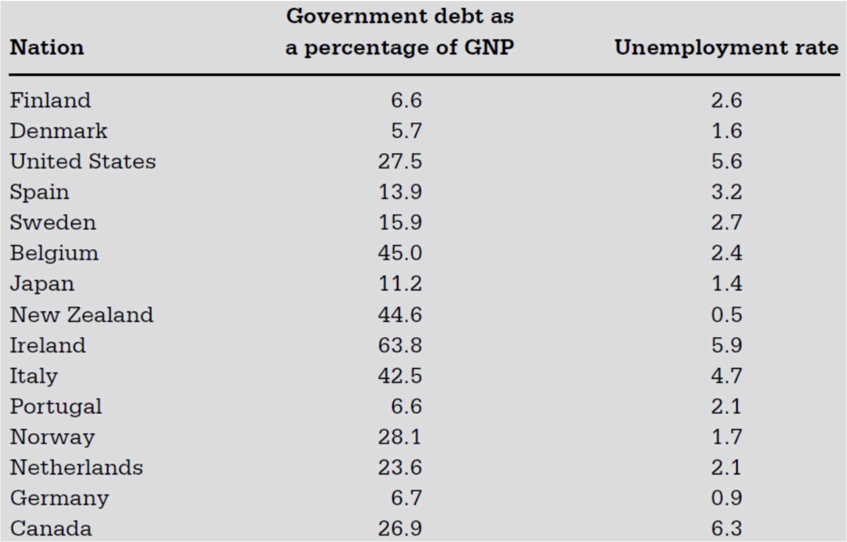
- In order to perform an econometric analysis, we need data.
- Econometrics, as mentioned earlier, has developed as an independent statistical tool mainly due to the specificity of the data under study.
- In general, the available data are observational, that is, non experimental (eg. interest rates).
- The type of available data is a crucial issue, as it determines both the kinds of questions that can actually be answered as well as the types of techniques that should be used.
- There are basically three types of data: **cross-sectional**, time series and panel.

Definition 5.3 (Cross-sectional data)**.** A sample of individuals (or firms, countries, etc.) collected at a specific point in time. Examples: surveys of familiar income, surveys of unemployment, ...

Definition 5.4 (Time-series data)**.** Observations on a set of variables over time for one statistical unit (individual, industry, sector, country, etc.). Example: GDP, interest rates.

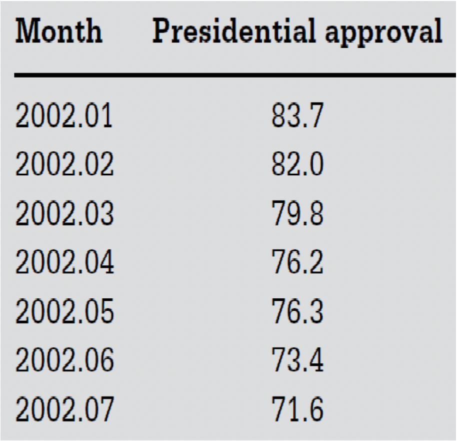
- The frequency of observation of the data is important.
- Data are naturally ordered chronologically.
- It is unusual to assume independence of the observations over time, since the past in general is relevant.

Definition 5.5 (Panel data)**.** Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities are observed across time. These entities could be states, companies, individuals, countries, etc.



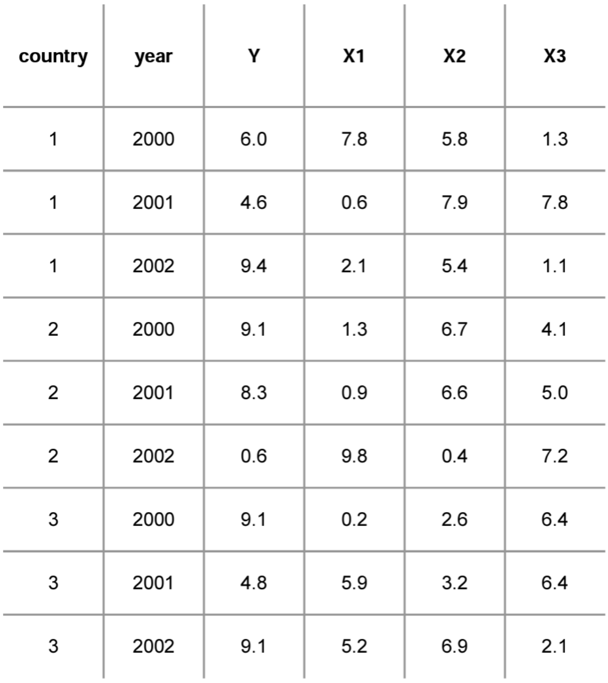
Nation	Government debt as a percentage of GNP	Unemployment rate
Finland	6.6	2.6
Denmark	5.7	1.6
United States	27.5	5.6
Spain	13.9	3.2
Sweden	15.9	2.7
Belgium	45.0	2.4
Japan	11.2	1.4
New Zealand	44.6	0.5
Ireland	63.8	5.9
Italy	42.5	4.7
Portugal	6.6	2.1
Norway	28.1	1.7
Netherlands	23.6	2.1
Germany	6.7	0.9
Canada	26.9	6.3

Figure 5.1: Cross-sectional data example.



Month	Presidential approval
2002.01	83.7
2002.02	82.0
2002.03	79.8
2002.04	76.2
2002.05	76.3
2002.06	73.4
2002.07	71.6

Figure 5.2: Time-series data example.



country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Figure 5.3: Panel data example.

5.3 The linear regression model

Definition 5.6 (Linear regression model)**.** The linear regression model is defined as

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u,$$

where:

- y: dependent or response variable;
- x_j, $j = 2, \ldots, k$: regressor (independent variable or explanatory variable);
- β_j, $j = 1, \ldots, k$: slope or regression coefficients (constants) for each regressor (β_1 is called y-intercept);
- u: residual variable (always a non-observable variable).

To estimate the parameters (regression coefficients) of the theoretical model, it is necessary to start from a sample, with dimension n:

$$\{(y_t, x_{t2}, \ldots, x_{tk}) : t = 1, 2, \ldots, n\}$$

which gives rise to n sample relations

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} + u_t,$$

where u_t represents the residual variable associated with the observation t.

The n inequalities can be presented using matrix notation:

$$\left\{\begin{array}{lcl} y_1 & = & \beta_1 + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_k x_{1k} + u_1 \\ y_2 & = & \beta_1 + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_k x_{2k} + u_2 \\ \vdots & \vdots & \vdots \\ y_n & = & \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_k x_{nk} + u_n \end{array}\right. \iff \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}.$$

5.3.1 Model's hypothesis

H1 - Linearity:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}.$$

- Only linear models will be treated: models that are easier to treat.

H2 - Exogeneity (regressors are exogenous):

$$E(u_t \mid X) = 0,\ t = 1, 2, \ldots, n.$$

- None of the information contained in X can be used to calculate $E(u_t)$. Important consequences of this assumption:
  - The unconditioned expected value of the residual variable is equal to zero (see prop. of the iterated expected value):
  $$E(u_t) = E(E(u_t \mid X)) = E(0) = 0, \quad t = 1, \ldots, n.$$
  - $E(y_t \mid X) = E(\beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t \mid X) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + E(u_t \mid X) =$
  $= \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + 0 = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} \iff E(y_t|X) = y_t - u_t \iff$
  $$u_t = y_t - E(y_t \mid X)$$
  (remains unknown as parameters are unknown)

- There is no linear association between the regressors and the residual variable (**model keypoint**)

$$Cov(x_{sj}, u_t) = E(x_{sj}u_t) - E(x_{sj})E(u_t) = E(E(x_{sj}u_t \mid X)) = E(x_{sj}E(u_t \mid X)) = E(0) = 0,$$

$t, s = 1, \dots, n;\ j = 2, 3, \dots, k.$

H3 - Conditioned homoscedasticity:

$$Var(u_t \mid X) = \sigma^2 > 0,\ t = 1, 2, \dots, n.$$

- The (conditioned) variance of the residual variable is constant $\forall t$, which implies 2 important consequences:
  - The unconditioned variance of the residual variables is constant:

$$Var(u_t) = Var[E(u_t \mid X)] + E[Var(u_t \mid X)] = Var[0] + E[\sigma^2] = \sigma^2.$$

  - The variance of y_t conditioned by X is constant:

$$Var(y_t \mid X) = Var(\beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t \mid X) = Var(u_t \mid X) = \sigma^2.$$

H4 - Absence of autocorrelation:

$$Cov(u_t, u_s \mid X) = 0,\ t = 1, 2, \dots, n,\ t \neq s.$$

- Important hypothesis for models applied to time-series data:
  - there is an "order" among the observations, which does not happen in cross-sectional models;
  - with time-series data it is frequent to specify models in which H4 does not hold, i.e., $Cov(u_t, u_s \mid X) \neq 0$ for $t \neq s$.
- Consequences of this hypothesis:
  - H2 + H4 $\Longrightarrow E(u_t u_s \mid X) = Cov(u_t, u_s \mid X) = 0, t \neq s,\ t, s = 1, \dots, n.$
  - Lack of correlation (unconditioned)

$$Cov(u_t, u_s) = E[E(u_t u_s \mid X)] - E(u_t)E(u_s) = 0 - 0 = 0.$$

  - Conditional covariances between the observations of the regression do not depend on the observations of the regressors:

$$Cov(y_t, y_s \mid X) = Cov(u_t, u_s \mid X) = 0, \quad t \neq s.$$

H5 - Non-existence of exact multicollinearity. The rank of matrix X is equal to k (number of regression coefficients) and $k < n$:

- More technical hypothesis that is intended to guarantee that the matrix $X^T X$ admits inverse $(X^T X)^{-1}$, that is, the columns of the matrix X are linearly independent.
- Non-existence of multicollinearity: x_j is not a linear combination of the remaining regressors $j = 1, \dots, k$:

$$x_j \neq \gamma_1 + \gamma_2 x_2 + \cdots \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_k x_k.$$

- Example (of multicollinearity): x_2 : yield in euros and x_3 : income in thousands of euros $\Longrightarrow x_2 = 1000x_3$.

H6 - Normal distribution of the residual variable conditioned by X (useful for statistical inference):

$$u_t \mid X \sim \mathcal{N}(0, \sigma^2).$$

Remark. Covariance matrix, U:

- Hypotheses H3 and H4 make it possible to determine the covariance matrix of U conditioned by X:

$$Cov(U \mid X) = \begin{bmatrix} Var(u_1 \mid X) & Cov(u_1, u_2 \mid X) & \cdots & Cov(u_1, u_n \mid X) \\ Cov(u_2, u_1 \mid X) & Var(u_2 \mid X) & \cdots & Cov(u_2, u_n \mid X) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(u_n, u_1 \mid X) & Cov(u_n, u_2 \mid X) & \cdots & Var(u_n \mid X) \end{bmatrix}$$

- Under H3: diagonal elements are all σ^2.
- Under H4: remaining elements are equal to zero:

$$Cov(U \mid X) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

The only unknown parameter of this matrix is σ^2.

5.4 Ordinary least square estimation

- Parameters (unknown, i.e., to be estimated) of the model:

$$\beta_1, \beta_2, \ldots, \beta_k \text{ and } \sigma^2.$$

- Estimation of regression parameters: Ordinary Least Squares (OLS) method.
- We will obtain the estimates (and other statistical information) through software outputs (MS Excel and STATA).

Definition 5.7. Estimator for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k).$$

To estimate $\boldsymbol{\beta}$ start with:

- a sample of n observations of k variables and y;
- $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_k)$: "approximation" for $\boldsymbol{\beta}$.

OLS method:

Obtain the vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_k)$ that minimizes the sum of squared residuals:

$$\begin{aligned} \varphi(\tilde{\beta}) &= \sum_{t=1}^{n} \tilde{u}_t^2 = \sum_{t=1}^{n} (y_t - \boldsymbol{x}_t \cdot \tilde{\boldsymbol{\beta}})^2 = \sum_{t=1}^{n} (\text{obs.value}_t - \text{appr.value}_t)^2 = \sum_{t=1}^{n} (\text{residual}_t)^2 \\ &= \sum_{t=1}^{n} (y_t - (\tilde{\beta}_1 + \tilde{\beta}_2 x_{t2} + \cdots + \tilde{\beta}_k x_{tk}))^2. \end{aligned}$$

How to get b?

- The vector $\boldsymbol{b}$ minimizes the sum of the squared residuals and gives us the adjusted model ("line"). It is possible to show that:

$$(X^T X)\boldsymbol{b} = X^T Y \qquad \underbrace{\Leftrightarrow}_{H5:\ X^T X \text{ is invertible}} \qquad \boldsymbol{b} = (X^T X)^{-1} X^T Y.$$

- The main consequence of choosing to minimize the square of residuals is to give greater weight to large residuals to the detriment of small ones.

Definition 5.8 (Adjusted linear regression function)**.** Once the OLS estimator of the regression coefficients is determined

$$\boldsymbol{b} = (X^T X)^{-1} X^T Y = (b_1, \ldots, b_k)$$

we obtain the adjusted linear regression function (for the data):

$$\hat{y}_t = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}.$$

$\boldsymbol{b} - \boldsymbol{\beta}$ represents the deviation between the estimator $\boldsymbol{b}$ and the true value of the vector of regression coefficients, $\boldsymbol{\beta}$:

$$\begin{aligned} b &= (X^T X)^{-1} X^T \underbrace{Y}_{X\beta + U} = (X^T X)^{-1} X^T (X\beta + U) = \beta + (X^T X)^{-1} X^T U \iff \\ \iff b - \beta &= (X^T X)^{-1} X^T U \quad \iff \quad (X^T X)(b - \beta) = X^T U. \end{aligned}$$

This deviation can never be determined exactly, because U is not observable.

Definition 5.9 (OLS residuals)**.** OLS Residuals related to the observation t:

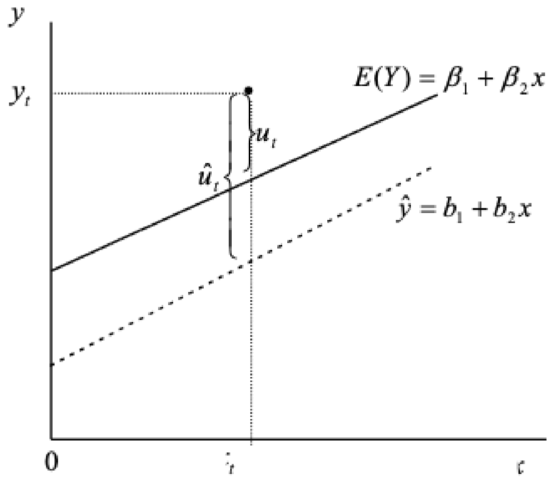
$$\hat{u}_t = y_t - \hat{y}_t, \qquad t = 1, 2, \ldots, n$$

Matrix notation:

$$\hat{U} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}, \qquad \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = Xb \qquad \Longrightarrow \hat{U} = Y - Xb = Y - \hat{Y}.$$

Remark. It is critical not to mix up:

- MS residuals $\neq$ residual variables, $\hat{u}_t \neq u_t$.
- theoretical linear regression function $\neq$ adjusted linear regression function:
    - theoretical linear regression: $E(y_t \mid X = x) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk}$.
    - adjusted linear regression: $\hat{y}_t = \widehat{E}(y_t \mid X = x) = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}$.
    - the value b_j represents the OLS estimative of β_j, $j = 2, \ldots, k$.



5.5 Interpretation of the estimation results

Interpretations are made in terms of the conditional expected value of Y

$$E(Y \mid X) = E(y_t \mid x_{t2}, \ldots, x_{tk}) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk},$$

which, for each set $(x_{t2}, \ldots, x_{tk})$ is estimated by

$$\hat{y}_t = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}.$$

To exemplify the most common situations, consider two examples:

1. Variable y is the variable of interest

$$E(y_t \mid x_{t2}, \ldots, x_{tk}) = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} \Longrightarrow \hat{y}_t = b_1 + b_2 x_{t2} + \cdots + b_k x_{tk}.$$

· β_j: represents a marginal change, that is, the change in $E(y \mid X)$ when x_j changes by one unit, everything else being constant (*ceteris paribus*), $j = 2, \cdots, k$;

· β_1 : independent term (in general, it does not have its own interpretation).

2. Variable of interest z with $y = \ln(z)$, $x_2 = \ln w_2$ and $x_3, \ldots, x_k$ do not result from transformation.

$$E(\ln z_t \mid w_{t2}, x_{t_3} \ldots, x_{tk}) = \beta_1 + \beta_2 \ln w_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk}$$

$$\Longleftrightarrow$$

$$E(y_t \mid x_{t2}, \cdots, x_{tk}) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} \Longrightarrow \hat{y}_t = b_1 + b_2 x_{t2} + b_3 x_{t3} + \cdots + b_k x_{tk}.$$

The interpretation of the parameters must now be done more carefully:

· β_2: represents a marginal change, that is, the change in $E(\ln z \mid X)$ when $\ln w_2$ changes by one unit, everything else being constant (*ceteris paribus*); this interpretation is meaningless (see the following point);

· β_2: represents a point elasticity of z (or better, the conditional expected value of z) with respect to w_2, that is, when w_2 varies by 1% the conditional expected value of z will vary by approximately β_2%, *ceteris paribus*;

· β_j, $j = 3, \cdots, k$: represents a point semi-elasticity of z (or better, the conditional expected value of z) with respect to x_j. In concrete terms, when x_j changes by one unit, the conditional expected value of z will change by approximately $100\beta_j$%, $j = 3, \cdots, k$, c.p.;

· The exact percentage change in z is given by

$$100(e^{b_j \Delta x_j} - 1).$$

It is recommended that you carefully read Chapter 12 of the reference book (Newbold) where these aspects are described in depth for a better understanding.

Summary

y-scale	x-scale	Interpretation
Y	X_j	$\Delta X_j = 1 \Longrightarrow \Delta E(Y \mid X) = \beta_j$
$\ln(Y)$	X_j	$\Delta X_j = 1 \Longrightarrow \Delta E(Y \mid X) = 100\beta_j\%$
Y	$\ln(X_j)$	$\Delta X_j = 1\% \Longrightarrow \Delta E(Y \mid X) = \beta_j/100$
$\ln(Y)$	$\ln(X_j)$	$\Delta X_j = 1\% \Longrightarrow \Delta E(Y \mid X) = \beta_j\%$

Example 5.6.

$$\widehat{logsalar}_t = 5.81505 + 0.0554 yeduc_t + 0.0230 yexper_t + 0.0040 yempl_t$$

Interpretation of estimates:

- The OLS estimate of the semi-elasticity of expected salary in relation to the number of years of education (return to education) is 0.0554, which means that if the number of education years increases by one, salary will increase (*c.p.*) by approximately $100 \times 0.0554\% = 5.54\%$ (the exact value is $e^{0.0554\times 1} - 1 \approx 5.69\%$).
- Similar interpretation is given to the other coefficients;
- The signs of the three estimates coincide with the expected signs for the respective parameters.

5.6 Properties of the OLS residuals

1. The sum of the residuals equals zero:

$$\sum_{t=1}^{n} \hat{u}_t = 0.$$

2. The sum of the products of the observations of each regressor by the residuals is zero:

$$\sum_{t=1}^{n} x_{tj}\hat{u}_t = 0, \quad j = 1, \ldots, k.$$

3. The sum of the products of the values adjusted by the residuals is equal to zero:

$$\sum_{t=1}^{n} \hat{y}_t\hat{u}_t = 0.$$

4. The sum of squares of the regression observations is equal to the sum of squares of the respective adjusted values plus the sum of squares of the residuals:

$$\sum_{t=1}^{n} y_t^2 = \sum_{t=1}^{n} \hat{y}_t^2 + \sum_{t=1}^{n} \hat{u}_t^2.$$

5. The OLS estimator of β, represented by b, whether or not conditioned by X, is centered (unbiased):

$$E(b \mid X) = E(b) = \beta.$$

6. The OLS estimator b, conditional on X, is linear in Y:

$$b = (X^TX)^{-1}X^TY = \phi(Y), \quad \phi \text{ linear function of } X.$$

7. The covariance matrix of the OLS estimator b, conditional on X, is

$$Cov(b \mid X) = \sigma^2(X^TX)^{-1},$$

thus,

$$Var(b_j \mid X) = \sigma^2_{b_j} = \sigma^2 m^{jj}, \qquad j = 1, \ldots, k,$$

where m^{jj} = diagonal element of order j of matrix $(X^TX)^{-1}$.

8. The OLS estimator of β is consistent.
9. (Gauss-Markov Theorem)

Theorem 5.1 (Gauss-Markov). *Whatever the estimator $\hat{\beta}$ of β is, linear and unbiased, the estimator b conditioned by X is more efficient than $\hat{\beta}$ (has smaller variance). Also, b is said to be BLUE (Best Linear Unbiased Estimator) for β (complicated proof).*

This property can be extended to a linear combination of the regression coefficients

$$\delta = c_1\beta_1 + c_2\beta_2 + \cdots + c_k\beta_k,$$

and it is possible to show that $\hat{\delta} = c_1b_1 + \cdots + c_kb_k$ is BLUE for δ.

Definition 5.10 (Unbiased estimator of the variance of the residual variables)**.** Now we want to estimate

$$\sigma^2 = Var(u_t) = E(u_t^2) - E^2(u_t) \underbrace{=}_{E(u_t)=0} E(u_t).$$

We can't use $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} u_t^2$, because u_t isn't observable. As a result, we can make use of

$$s^2 = \frac{1}{n-k}\sum_{t=1}^{n} \hat{u}_t^2$$

which is a centered estimator for σ^2.

We define the **standard error** of the regression as

$$s = \sqrt{s^2}.$$

Once σ^2 has been estimated, the covariance matrix of b, conditioned by X, can be estimated by:

$$\widehat{Cov}(b \mid X) = s^2(X^T X)^{-1}, \quad \text{particularly} \quad \widehat{Var}(b_j \mid X) = s^2 m^{jj} = s_{b_j}^2.$$

We define the **standard error** of b_j by the quantity

$$s_{b_j} = \sqrt{\widehat{Var}(b_j \mid X)} = s\sqrt{m^{jj}}.$$

How to assess the goodness of fit after estimating the model?

Definition 5.11 (Coefficient of determination)**.** To evaluate the "goodness of fit" (of the regression to the data) one can resort to the **coefficient of determination**:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}, \qquad 0 \leqslant R^2 \leqslant 1,$$

where

Total **S**um of **S**quares: total variation in the dependent variable

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Explained **S**um of **S**quares: variation (of the dependent variable) explained by the regression

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Residual **S**um of **S**quares: variation (of the dependent variable) not explained by the regression

$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{u_i}^2.$$

Remark: $SST = SSE + SSR$.

The closer to one the coefficient of determination is, the better the goodness of fit.

Remark. We should only use R^2 to compare models that have the **same dependent variable**.

The coefficient of determination, R^2, has two major drawbacks:

- Being a summary measure of interpretation that is not always easy: what is a low/high value?
- When one more regressor is added to the model, whatever it may be, the value of R^2 never decreases (for the same sample), because $\sum_{t=1}^{n} \hat{u}_t^2$ does not grow.

Definition 5.12 (Adjusted coefficient of determination)**.** To overcome the shortcomings of using R^2, the **adjusted coefficient of determination** is used, which penalizes the introduction of more variables (regressors):

$$\overline{R}^2 = 1 - \frac{SSR/(n-k)}{SST/(n-1)} = R^2 - (1-R^2)\frac{k-1}{n-k}.$$

This coefficient has the inconvenience of being negative and never reaching the value of 1.

It should only be used to compare models that have the same dependent variable.

5.7 Statistical inference on the MLR model

Starting from a sample, we could be interested in:

- Estimating the coefficients and their variances through OLS;
- Construct confidence intervals for the coefficients;
- Test hypotheses on the parameters.

The objective now is to make inference on the model's parameters β_j.

Example: $\ln(sal) = \beta_0 + \beta_1 educ + u$.

Could it be that $\beta_1 = 0$? Or that $\beta_1 < 0$? Or that $\beta_1 = 1$?

We need to find a test statistic with a fully known distribution.

In order to do this, we need to choose a distribution for the population.

REMARK: closely follow the Hypothesis testing chapter!

Statistical inference on the variance of residual variables

Reduced interest but is important for Econometrics...

Hypothesis:

$$H_0 : \sigma^2 = \sigma_0^2 \quad vs \quad H_1 : \sigma^2 \neq \sigma_0^2 \quad (or \;\; H_1 : \sigma^2 < \sigma_0^2, \; or \;\; H_1 : \sigma^2 > \sigma_0^2).$$

TS:

$$Q = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k), \quad \text{with} \quad s^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{u}_i^2.$$

Statistical inference for one regression coefficient

Hypothesis:

$$H_0 : \beta_j = \beta_{0j} \quad vs \quad H_1 : \beta_j \neq \beta_{0j} \quad (or \;\; H_1 : \beta_j < \beta_{0j}, \; or \;\; H_1 : \beta_j > \beta_{0j}).$$

TS:

$$T_j = \frac{b_j - \beta_j}{s_{b_j}} = \frac{b_j - \beta_j}{s\sqrt{m^{jj}}} \sim t(n-k).$$

Particular (and very important) case: $\beta_{0j} = 0$ (statistical significance of the regressor).

Statistical inference for one regression coefficient (signal test)

Hypothesis:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j > 0 \; (or \; < 0).$$

TS:

$$T_j = \frac{b_j}{s_{b_j}} \sim t(n-k).$$

Statistical inference on one linear combination of regression coefficients

The goal now is to test

$$\delta = c_1\beta_1 + c_2\beta_2 + \cdots + c_k\beta_k = c\beta^T.$$

The MQ estimator of δ is

$$\hat{\delta} = c_1 b_1 + c_2 b_2 + \cdots + c_k b_k = cb^T.$$

It can be shown that

$$t_{\hat{\delta}} = \frac{\hat{\delta} - \delta}{s_{\hat{\delta}}} \sim t(n-k).$$

$s_{\hat{\delta}} = s\sqrt{c(X^TX)^{-1}c^T}$ is the standard error of $\hat{\delta}$ and can be obtained from the covariance matrix of b.

A practical solution will be seen later when the matrix of covariances of b is not available (in MSExcel it is not easy to obtain this matrix).

Joint nullity test of regression coefficients (nullity of a subset)

Find out if some of the regression coefficients are jointly equal to zero. Suppose we want to test whether the last $m = k - p$ coefficients are equal to zero. Assumptions:

$$H_0 : \beta_{p+1} = 0, \beta_{p+2} = 0, \ldots, \beta_k = 0$$

$$vs$$

$$H_1 : \exists \, \beta_j \neq 0, \quad j = p+1, \ldots, k.$$

Perform a test with 3 steps:

Step 1 - Estimate the model **without restrictions**, i.e., with all the regressors, and obtain $SSR_1 = \sum_{t=1}^{n} \hat{u}_t^2$.

Step 2 - Estimate the model **with restrictions**, i.e., eliminating the regressors that are assumed to have a zero coefficient:

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_p x_{tp} + \underbrace{\beta_{p+1}}_{=0} x_{tp+1} + \cdots + \underbrace{\beta_k}_{=0} x_{tk} \Longrightarrow y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_p x_{tp},$$

and get $SSR_0 = \sum_{t=1}^{n} \hat{u}_t^2$.

Step 3 - Compare the models, using the ST

$$F = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(n-k)} = \frac{SSR_0 - SSR_1}{ms^2} \sim F(m, n-k),$$

where $m = k - p$ is the number of constraints.

$s^2 = \dfrac{SSR_1}{n-k}$ is the estimate of σ^2 based on the unconstrained model.

REMARK

Instead of using SSR_0 and SSR_1, we can use the respective determination coefficients:

$$F = \frac{(R^2 - R_0^2)/m}{(1-R^2)/(n-k)} \sim F(m, n-k).$$

If the individual test of each of the coefficients included in H_0 does not reject nullity and the joint test does, then one should suspect possible multicollinearity.

The reverse situation (one does not reject the joint nullity of some regressors, with the F-test, but one rejects the nullity for a particular coefficient by the t-test) is also possible, but in this case it is generally preferable to trust in the t-test (more powerful than the F-test to detect whether a given regression coefficient is different from zero).

Global significance test of the regression

Test the nullity of all coefficients with the exception of the independent term:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0 \quad \text{vs} \quad H_1 : \exists \beta_j \neq 0, \;\; j = 2, \ldots, k.$$

Not rejecting the null hypothesis corresponds to verifying that the proposed model is not globally adequate to describe the behavior of the regression.

The test statistic is obtained by the previous system, now being the number of constraints $m = k - 1$:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{SSE/(k-1)}{SSR/(n-k)} \sim F(k-1, n-k).$$

Final remark

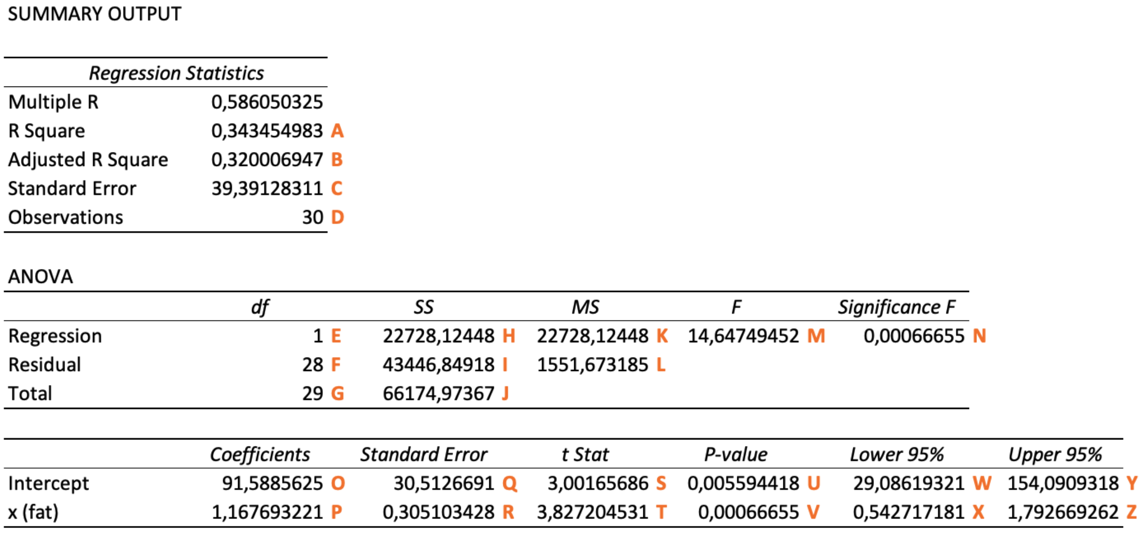
When the residual variable does not have a normal distribution (violation of hypothesis H6), but the sample is large, the CLT can be applied:

$$T_j = \frac{b_j - \beta_j}{s_{b_j}} \overset{a}{\sim} \mathcal{N}(0,1).$$

Exercise 5.1. In a study on cholesterol in a certain risk group, a random sample of 30 subjects was collected, and the following variables were observed:

- $x =$ number of grams of fat consumed per day;
- $y =$ amount of cholesterol in the blood (in milligrams per deciliter).

The estimation results with EXCEL are shown below:



SUMMARY OUTPUT

Regression Statistics		
Multiple R	0,586050325	
R Square	0,343454983	A
Adjusted R Square	0,320006947	B
Standard Error	39,39128311	C
Observations	30	D

ANOVA

	df		SS		MS		F		Significance F	
Regression	1	E	22728,12448	H	22728,12448	K	14,64749452	M	0,00066655	N
Residual	28	F	43446,84918	I	1551,673185	L				
Total	29	G	66174,97367	J						

	Coefficients		Standard Error		t Stat		P-value		Lower 95%		Upper 95%	
Intercept	91,5885625	O	30,5126691	Q	3,00165686	S	0,005594418	U	29,08619321	W	154,0909318	Y
x (fat)	1,167693221	P	0,305103428	R	3,827204531	T	0,00066655	V	0,542717181	X	1,792669262	Z

(a) Write down the meaning of all the orange letters from A to Z.

```
#



#
```

(b) Write the theoretical and the estimated models.

```
#



#
```

(c) Interpret the value of the letter P.

```
#

#
```

(d) Test the hypothesis that the slope of the line is equal to 1 against the alternative that it is greater than 1 (dimension 0.05).

```
#

#
```

(e) Construct a 95% confidence interval for the slope of the regression line.

```
#

#
```

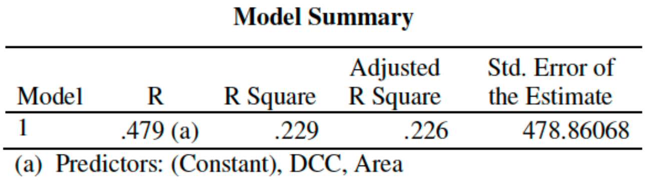
(f) Based on the confidence interval constructed in the previous question, what can you conclude about the relationship between cholesterol and ingested fat?

```
#

#
```

Exercise 5.2. Consider the following regression model (checking the basic assumptions):

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \quad t = 1, 2, \cdots, 500,$$

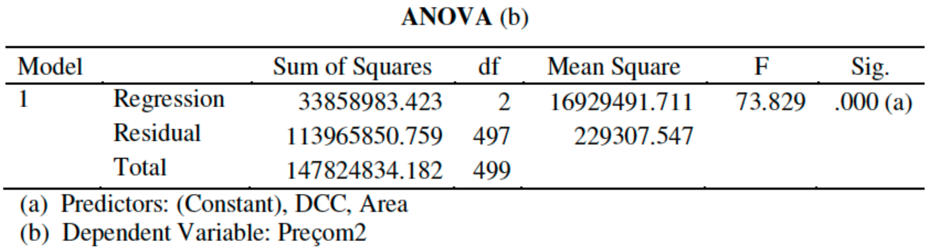
where y is the price (in monetary units) per m^2 of an apartment in a given city, x_2 is the area of the apartment, and x_3 is the distance to the city centre in kilometers. The model was estimated with the SPSS software, and the results are:



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.479 (a)	.229	.226	478.86068

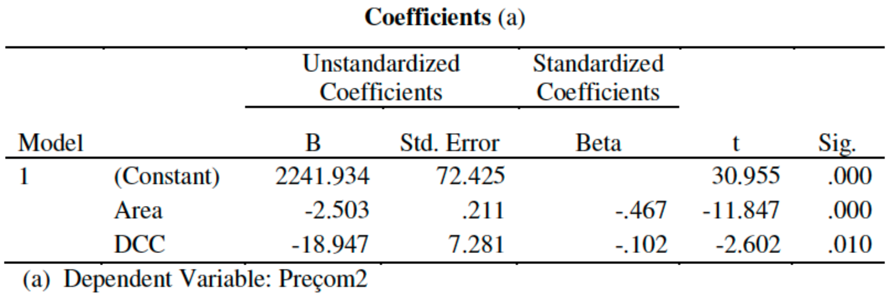
(a) Predictors: (Constant), DCC, Area



ANOVA (b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33858983.423	2	16929491.711	73.829	.000 (a)
	Residual	113965850.759	497	229307.547		
	Total	147824834.182	499			

(a) Predictors: (Constant), DCC, Area
(b) Dependent Variable: Preçom2



Coefficients (a)

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	2241.934	72.425		30.955	.000
	Area	-2.503	.211	-.467	-11.847	.000
	DCC	-18.947	7.281	-.102	-2.602	.010

(a) Dependent Variable: Preçom2

Remark: The column "Standardized coefficients" is not important at this stage.

(a) Briefly analyse the results obtained in statistical terms (consider $\alpha = 0.05$) and in economic terms, interpreting the estimates obtained for the regression coefficients.

```
#




#
```

(b) Perform the test

$$H_0: \ \beta_2 = -2 \ \text{ vs } \ H_1: \beta_2 < -2.$$

In the case of not rejecting the null hypothesis, would it be reasonable to exclude the regressor x_2 from the model?

```
#

```

```








#
```

Exercise 5.3. The company Electrik intends to build an explanatory model of family consumption (in monetary units) of electricity, y, as a function of family income, x_2, the number of individuals in each family, x_3, and the area of the respective house in square meters, x_4. The estimation results with EXCEL are shown below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932985993
R Square	0.870462864
Adjusted R Square	0.805694295
Standard Error	4.571741129
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	3	842.6950983	280.8983661	13.43959
Residual	6	125.4049017	20.90081695	
Total	9	968,1		

	Coefficients	*Standard Error*	*t Stat*	*p-value*
Intercept	12.47998124	9.282257233	1.344498534	0.227386
x_2	0.060527128	0.024073045	2.514311292	0.045637
x_3	-2.81451648	2.679535456	-1.050374786	0.334
x_4	0.020535759	0.057861717	0.354910983	0.734798

(a) Taken as a whole, do you think that the regressors included in this model are useful for explaining consumption? Test at 0.05.

```
#











#
```

(b) Looking only at the p-value in the table, say which of the regressors appears to be statistically significant.

```
#


```

```
#
```

(c) Criticize the adopted model, taking the number of observations into account.

```
#

#
```

(d) Test at 0.05 the hypothesis that the increase in income implies, on average, an increase in consumption.

```
#

#
```

(e) The company Electrik decided to estimate another model in which it only included family income as a variable. The estimation results are shown below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.90358399
R Square	0.816464028
Adjusted R Square	0.793522031
Standard Error	4.712764247
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	790.4188252	790.4188252	35.58819
Residual	8	177.6811748	22.21014685	
Total	9	968.1		

	Coefficients	*Standard Error*	*t Stat*	*p-value*
Intercept	11.02028658	2.963330537	3.718885369	0.005881
Rendimento	0.050015429	0.008383996	5.965583414	0.000336

Test at 0.05 the simultaneous nullity of the coefficients concerning the variables “number of individuals” and “area of the house.” Which of the two models seems preferable?

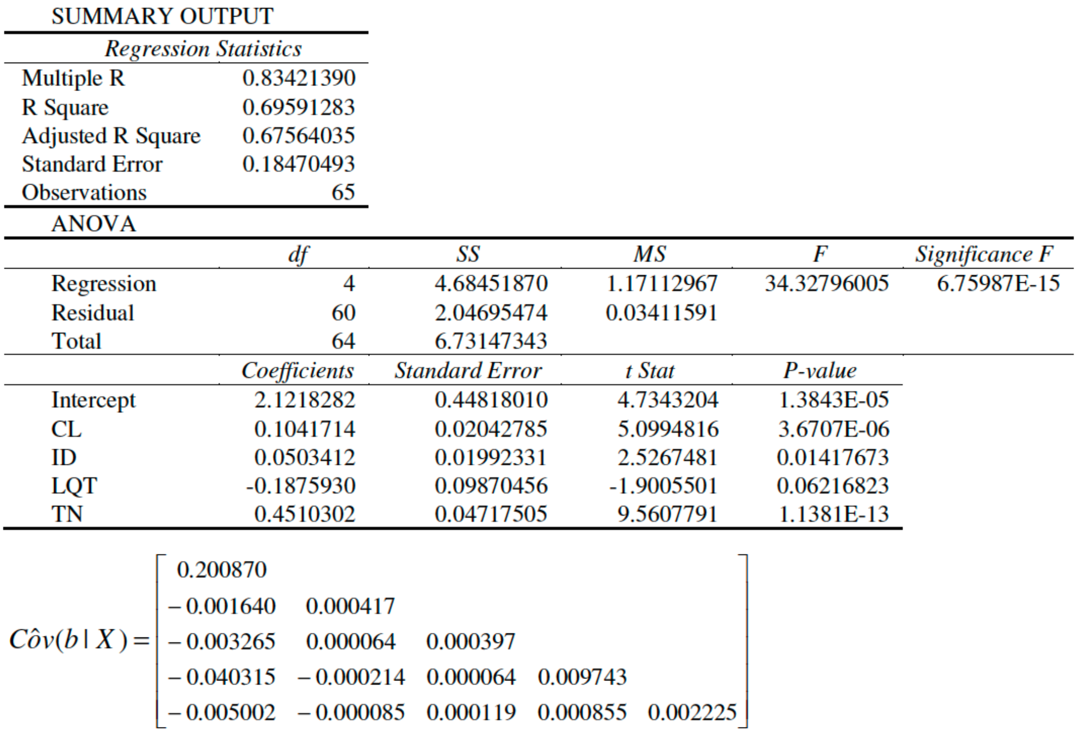
```
#
```

```
#
```

Exercise 5.4. An economist from an association of wine producers decided to build an explanatory model of the price (in euros/bottle) of wines from a certain demarcated region, PR. For this purpose, he selected as explanatory variables the classification, from 1 to 10, given by a magazine (CL), the age of the harvest in years (ID), the quantity produced in thousands of bottles (QT), and a variable (TN), which takes the value 1 if the wine has mostly grapes of the "Touriga Nacional" variety (a variety of red wine grape) and 0 otherwise, and proposed the following specification:

$$LPR = \beta_1 + \beta_2 CL + \beta_3 ID + \beta_4 LQT + \beta_5 TN + u,$$

where LPR represents the natural logarithm of the price of the bottle LQT is the natural logarithm of the quantity produced. Assuming that the hypotheses of the multiple linear regression model were satisfied and that a random sample of 65 producers was observed (one bottle per producer), the model was estimated, obtaining the following results:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83421390
R Square	0.69591283
Adjusted R Square	0.67564035
Standard Error	0.18470493
Observations	65

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	4.68451870	1.17112967	34.32796005	6.75987E-15
Residual	60	2.04695474	0.03411591		
Total	64	6.73147343			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.1218282	0.44818010	4.7343204	1.3843E-05
CL	0.1041714	0.02042785	5.0994816	3.6707E-06
ID	0.0503412	0.01992331	2.5267481	0.01417673
LQT	-0.1875930	0.09870456	-1.9005501	0.06216823
TN	0.4510302	0.04717505	9.5607791	1.1381E-13

$$\hat{Cov}(b \mid X) = \begin{bmatrix} 0.200870 & & & & \\ -0.001640 & 0.000417 & & & \\ -0.003265 & 0.000064 & 0.000397 & & \\ -0.040315 & -0.000214 & 0.000064 & 0.009743 & \\ -0.005002 & -0.000085 & 0.000119 & 0.000855 & 0.002225 \end{bmatrix}$$

When answering the following questions, use a dimension of 5% for all the tests you have to perform:

(a) Both as a whole and individually, do you think that the regressors included in this model are useful in explaining the logarithm of the selling price of wine in that demarcated region?

```
#
```

```
#
```

(b) Interpret the estimates obtained for the coefficients β_2 and β_4.

```
#

#
```

(c) Construct the confidence interval at 95% for β_3.

```
#

#
```

(d) Comment, justifying the following sentence: "For a dimension of 5% there is statistical evidence that on average, and under equal circumstances regarding the remaining explanatory variables, the greater the quantity produced, the lower the price of wine."

```
#

#
```

(e) Perform the test

$$H_0: \ \beta_2 = 2\beta_3 \ \text{ vs } \ H_1: \ \beta_2 \neq 2\beta_3.$$

```
#









#
```

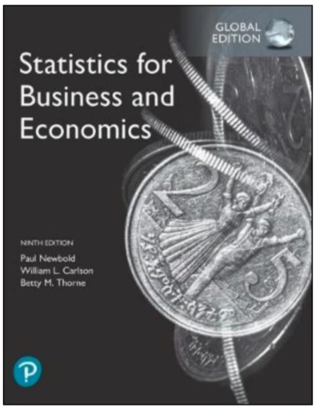
Acknowledgements

Some exercises and examples were based on exercises in the book:

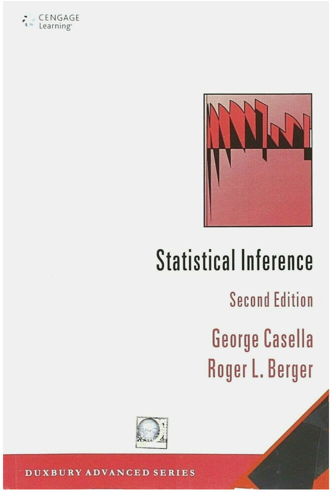
- B. Murteira, C. Ribeiro, J. Andrade e Silva, C. Pimenta, F. Pimenta, (2015) Introdução à Estatística (3rd Edition). Escolar Editora.

Bibliography

- P. Newbold, W. Carlson, B. Thorne, (2020) Statistics for Business and Economics (9th Edition), Pearson Education. [English]



- G. Casella, R. Berger, (2002) Statistical Inference (2nd Edition), Thomson Learning. [English]



- B. Murteira, C. Ribeiro, J. Andrade e Silva, C. Pimenta, F. Pimenta, (2015) Introdução à Estatística (3rd Edition). Escolar Editora. [Portuguese]