Notes on Statistics II

Xavier Vilà
Universitat Autònoma de Barcelona

Year 2004-2005

Universitat Autònoma de Barcelona

Contents

Introduction: What is Statistics ?

Think of a researcher who seeks to explain some fact form the real world. For instance, imagine Newton trying to explain why apples fall. As a closer example, imagine an economist trying to explain why unemployment does exist.
Usually, the task of a researcher consists of three parts:

1. Observe the world in order to determine the problem to study and gather information about it
2. Think about the problem
3. Produce an explanation or *Theory* for the problem.

Statistics become extremely important for the first of these three elements.[1]

It is clear that, in order to study a "real problem", the researcher must observe the "real" world. Nevertheless, it is also clear that no researcher can observe the *whole* reality. Newton can not observer all the falling apples, neither can an economist interview the whole population of a country to determine the unemployment rate. It is hence necessary to somehow "summarize" the reality, but this task has to be done so that such "summary" closely fits the reality. Then, and only then, conclusions drawn from the "summary" can be reliably applied to the whole population.

> Statistics (more precisely, statistical inference) is a collection of techniques by means of which we can draw conclusions with regards to a *reality* from the study of a *summary* of such reality

Hereafter we will study in detail how this is done. Chapter one explains how the *reality* is rigorously *summarized* and what are the main features of the results obtained in this process. Chapter two deals with the first approach on how to generate conclusions about some *real* issues based on what we observe in the *summary*. Chapters three and four introduce more sophisticated techniques to make inferences about the *reality* using some of the more elemental results seen in Chapter two. Finally, Chapter five introduces the linear regression analysis, a technique widely used in the economic

[1] Very often the researcher does not start up by gathering information using statistical techniques. On the contrary, in many cases his initial activity consists of detecting general patterns of behavior for a given fact. From here, researchers are able to build up an abstract theory in order to explain the phenomenon at study. This is, for example, Newton's way, and also the way Economic Theory works. Once this "abstract theory" is logically constructed, statistical techniques are often used to check whether such theory fits the reality, as we will see in Chapter 5.

analysis (and other sciences) to study the relationship between variables. It is worth saying that a clear understanding of the topics in Chapter one are important in order to easily understand what other chapters deal with, and also to get an global idea of the whole process of statistical inference.

Chapter 1

Sampling Theory

This chapter formally intoduces how the *summary* mentioned above is done and which are the main features of the conclusions drawn from it.

At this point, it is important to understand that statistics is based on *probabilistic* techniques. Hence, any statistical conclusion drawn from this kind of *summary* will not be *true for sure* when applied to the whole *reality*, but only with a certain probability. For instance, when an electoral survey is conducted it is clear that its results do not exactly coincide with the results in the final election. Nevertheless, if the survey is "well done", that is, if the *summary* of the *reality* (which in this case is the set of people interviewed) closely represents the whole *reality* (which in this case is the whole population that has the right to vote), then the survey result will be close to the final results with a high probability

In the sections below we will see which are the basic ingredients of any statistical analysis and its probabilistic features

1.1 Population, Sample, Parameter, Statistic, and Statistical Inference

Statistical inference is mainly built upon four main concepts, which will be defined and described below. These concepts are closely related to each other and it is very important to clearly understand each of them and not to mistake one by the other.

Population Is the set of elements that are the object of study[1]. The goal will be to draw some conclusion regarding some specific feature of this population.

Example 1.1.1 *All the apples in the world. The feature at study is whether an apple falls down or not.*

Example 1.1.2 *Labor force in the European Union. The feature at study is whether a worker is unemployed or not.*

Example 1.1.3 *Production of Pentium IV chips in a given day. The feature at study is whether a chip is faulty or not.*

[1] In this sense, the **Population** is what we have called *reality* in the Introductory chapter

Sample Subset of the **Population** used to draw conclusions about the population

Example 1.1.4 *50 apples in Newton's garden.*

Example 1.1.5 *Unemployment statistics at the European Union.*

Example 1.1.6 *25 Pentium IV chips manufactured in a given day.*

Parameter Is the feature of the population that we want to know something about. This feature has to be a numerical one[2] and, obviously, its true value must be unknown[3]

Example 1.1.7 *What is the proportion of falling apples.*

Example 1.1.8 *What is the unemployment rate at the European Union*

Example 1.1.9 *What is the proportion of faulty chips among those produced in a given day.*

Statistic Computation made using the elements in the **sample** and used to get an approximation to the true value of the **parameter**. It is important to notice that this value will be known (since we will compute it) and will be used to draw conclusions on the true value of the **parameter**, which is unknown and is what is of interest to us.

Example 1.1.10 *Proportion of falling apples among the 50 sampled apples in Newton's garden.*

Example 1.1.11 *Unemployment rate among the workers interviewed in the unemployment statistics in the European Union.*

Example 1.1.12 *Proportion of faulty chips among the 25 selected chips produced in a given day.*

From this four main concepts, the process of statistical inference works as follows:

1. Using sampling techniques that will be explained below, a **sample** is selected from the **population** that is going to be studied.
2. From this **sample**, the proper computations are done in order to obtain a **statistic**.
3. From this **statistic**, using some statistical inference technique that we will see in other chapters, some conclusions are drawn regarding the unknown population **parameter** that represents the feature of the population that is to be studied.

This process can be represented as in Figure 1.1

We can now provide a definition for Statistics (or Statistical Inference, to be more precise) which is more formal than the one offered in the introduction.

Definition 1.1.13 *Statistical Inference is a subject whose main objective is to draw conclusions regarding a **population** thru the study of one **sample** by means of probabilistic techniques.*

[2] Although non numerical features can be studied as well, the techniques used in such cases are different from those that we will see here. Nevertheless, Chapter four will introduce some of these analysis.

[3] For otherwise it will not be necessary to do any statistical analysis at all !

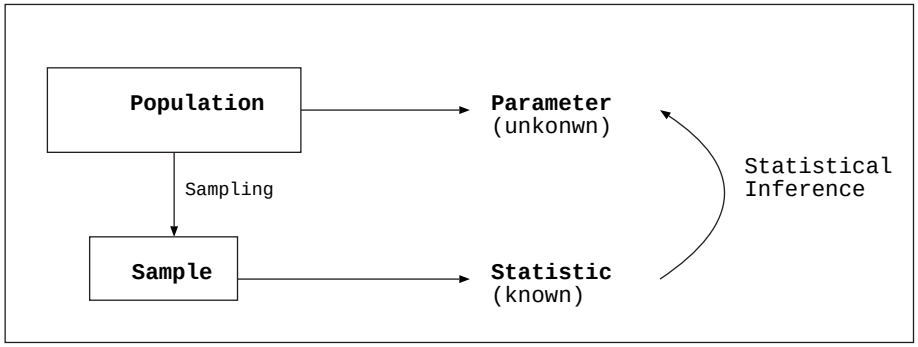

Figure 1.1: The process of Statistical Inference

1.2 Sampling types

We will see what a **sample** is, that is, how a **sample** can be selected out of a **population**. Since we want to study this sample to produce conclusions about the **population**, it can not be selected arbitrarily. In this sense, there exist rigorous techniques specially tailored for this purpose. In what follows, the more basic techniques will be introduced, while more sophisticated analysis are out of the scope of these notes. The following definition approaches the idea of *sampling*

Definition 1.2.1 ***Sampling*** *is a systematic technique to select a sample out of a population in such a way that it is representative of the population*

Here, the keyword is *representative*. Indeed, if we want our sample to be used in order to produce "reliable" conclusions regarding the original population, we would better have a sample that closely resembles (in its structure) the original population. For instance, if we want to conduct an electoral survey and we only interview people living in a "rich" neighborhood, then it is clear that their answers will not be representative of the whole population.

There are different types of sampling techniques, depending on the specific features of the study at hand. The more important are:

1.2.1 Probabilistic Sampling.

Consist of all the sampling techniques that are based on random methods to select the sample from the population. There are different kinds of random samplings:

1.2.1.1 Simple Random Sampling (SRS).

This is the "most random" of all the probabilistic sampling methods, and throughout this notes we will normally assume that samples are obtained using this technique. Its main feature is that *all elements in the population have the same probability of being selected to be incorporated to the sample*. In other words, the sample is constructed *completely* at random. If we think for a moment of all the possible different samples that can be selected from a given population, simple random sampling means that each of these samples has the same probability of been selected as "the sample", i.e., they are equally likely

Example 1.2.2 *Consider a population consisting of only 4 elements*

$$\textit{Population} = \{A, B, C, D\}$$

If, for instance, we want to draw a sample of size 2, there are 6 possible samples

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
$\{A, B\}$	$\{A, C\}$	$\{A, D\}$	$\{B, C\}$	$\{B, D\}$	$\{C, D\}$

Table 1.1: Possibles Samples

Hence, in a Simple Random Sampling, each of this samples has the same probability of being selected, $\frac{1}{6}$ in this case. Analogously, we may also say that each of the four elements in the Population has the same probability of being drawn to enter the selected sample. Indeed, since each of the elements belongs to exactly 3 of the possible sample and each possible sample has probability $\frac{1}{6}$ of being the selected sample, then the probability for any of the elements in the Population of entering the selected sample is $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$. This probability ($\frac{1}{2}$) can also be understood as each element in the Population having probability $\frac{1}{4}$ of being the first element to enter the sample and probability $\frac{1}{4}$ of being the second element in the sample, which yields a total of $\frac{1}{2}$ probability of being one of the two elements in the sample.

1.2.1.2 Systematic Sampling.

The Systematic Sampling consist of a variant of a SRS. It is useful when the population to sample is not "static", but changes often. The following example shows how this method works.

Example 1.2.3 *Consider a factory that manufactures Pentium "chips". The managers want to study how many of these chips turn out to be faulty every day. The factory has a "chain" process so that once the "chip" has been assembled, it automatically enters in the packaging process and then moves into warehouse. Let us suppose that the factory produces 100 "chips" a day, and that a sample of size 5 is going to be selected. It is clear that the managers can not wait until the end of the day, then stop all processes, randomly select 5 chips, and start all processes over again. This would be very costly. What is needed is a way to randomly select "chips" but without having to stop the manufacturing chain process. Here is what can be done in cases like this.*

1. *Select "a priori" which "chips" will be systematically taken out from the chain process to enter the sample. In this example, if 100 chips are produced daily and only 5 need to go to the sample, then we must select one chip out of every 20 produced chips.*

2. *Randomly pick a number between 1 and 20 (here is were "randomness" play a role). Let us suppose that the selected number is 6.*

3. *Following what resulted in the previous items, we must then select chips numbered 6, 26, 46, 66 i 86. That is, starting from chip number 6 (in order of production), we count every 20 chips to construct the sample.*

4. *We can "program" the machines in the chain production so that the selected chips are automatically "deviated" out of the process. Other chips continue their way to packaging without any interference.*

The method just described allows us to obtain a random sample without having to disrupt the production process.

1.2.1.3 Stratified Sampling.

This sampling is another variant of the SRS that makes sense when there is some information regarding the structure of the population. Using this information, it is possible to construct samples which are more representative that those obtained directly with a SRS. The following example shows who this sampling technique works.

Example 1.2.4 *An electoral survey is to be conducted in the city of Barcelona. It is known that voting is very correlated with the district of residence. In other words, a person living in Pedralbes has a higher probability of voting to he right than a person living in the Poble Sec. In order to avoid that a SRS selects too many people from the same district, the sample (of size n) can be splitted in several "subsamples" (one for each district in the city) so that the union of these samples is n Then, each of these subsamples is obtained by means of a SRS in each district.*

The results from this type of sampling are usually more representative, the only problem being we need to know the relative weight of each district with respect to the total of the city. Once this is known, the relative weight of each subsample with respect to the whole sample must be adjusted to mimic the true weights in the city.

1.2.1.4 Step by step sampling.

This is another variant of the SRS that makes sense when, given the structure of the population to study, the realization of a SRS would be very costly. The following example shows how this sampling technique works.

Example 1.2.5 *We want to conduct a survey to know the situation of the public schools in Catalonia. Since this is a very delicate topic, we must travel to each of the schools that have been picked to belong to the sample and interview the Director. In this context, a SRS might very well select a sample composed of schools disseminated all over the territory, which would imply a high level of travel expenditure. To avoid this, we can do the following:*

1. *Perform a SRS within all the "comarques" in Catalonia, so that 10 "comarques" are selected to visit.*

2. *Within each of the 10 selected "comarques", perform another SRS to select 20 towns to visit. Hence, we will have a total of 200 cities to visit.*

3. *Finally, within each of the selected cities perform one SRS more to select one public school to visit.*

In this way, we have selected 200 public schools to visit in Catalonia with travel costs lower than using a SRS. The problem, though, is that the sample obtained will be less representative.

Each of this sample techniques has its own pros and cons.
In what follows, we will always assume (implicitly) that the sample at hand has been obtained by means of a SRS.

1.2.2 Non Probabilistic Sampling

In some cases the sample is obtained without any randomness at all. For instance, if we want to test a new drug against malaria, we can not just randomly select "subjects" and force them to take the drug. In cases like this, a call for volunteers is made. This techniques are usually much less representatives that a random technique. Furthermore, since there are no random components in the sample, we can not use probabilistic tools to study the sample and, therefore, statistical inference techniques can not be correctly applied.

1.3 Sample Statistics Distributions

Once the sample is obtained (we will always assume that using a SRS), the process of working with it and draw conclusions starts.

In this sense, the main task is now to obtain a **statistic**, one of the main elements in statistical inference. We will use it to produce conclusions regarding the unknown population **parameter** that is of interest to us.

The definition that follows will remind us what a **statistic** is (as introduced in the previous section). Then, the concept of **estimate** is defined. Although these two concepts are very similar and closely related, it is very important to notice that they are not the same thing.

Definition 1.3.1 *A **statistic** (or **estimator**)[4] is a formula that uses the values in the sample at hand (observations) in order to produce an approximation to the true value of an unknown population parameter.*

Definition 1.3.2 *An **estimate** (or **estimation**) is the particular value of an estimator that is obtained from a particular sample of data and normally used to indicate the value of an unknown population parameter.*

Hence, a **statistic** is not a number but a formula while an **estimate** is the number that is obtained when the formula (the estimator) is applied to the observations of the specific sample that we have at hand.

At this point, it becomes crucial to understand that, given that the sample is obtained by means of a random technique, the **statistic** will produce different **estimates** with different probabilities (depending on the specific sample that is finally "selected" at

[4]The fact that the the same "object" can have two names must no lead to confussion. Depending on the kind of analysis that we want to perform, the same "formula" is referred to with one name or the other. In Chapter 2 we will use the term **estimator**, while in the chapters that follow we will rather use the name **statistic**. It is always the same idea, but used purposes for different .

random). To put it more formally, a **statistic** is a *Random Variable*, that is, a variable that takes different values with different probabilities In this sense, an **estimate** is a specific realization of this random variable. The following example aims to clarify this idea.

Example 1.3.3 *We want to know the average number of cars per family in a given population. To keep the example simple, we will assume that the population is very small, only 4 families,* $Population = \{A, B, C, D\}$

Let us now assume that family A *owns one car, families* B *and* C *have 2 cars each, and family* D *has 4.* [5]

For the study, we want to obtain a random sample of size 2. We can then compute the average number of cars in the sample and use it to infer some conclusion regarding the true average in the population. Hence, the sample mean (or just mean, for short) will play the role of ***statistic*** *in this example, and we can use it to draw conclusions on the true population* ***parameter*** *that is of interest to us: the average number of cars per family in the whole population, that is, the* ***population mean****.*

Table 1.3 summarizes:

1. *The 6 possible samples than can be the result of a sampling process on this population,*

2. *for each of the possible samples, the probability of being selected (all of them will have the same probability as we are assuming SRS)*

3. *the* ***estimate*** *value that would result from applying the sample average formula to the corresponding sample*

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Elements	$\{A, B\}$	$\{A, C\}$	$\{A, D\}$	$\{B, C\}$	$\{B, D\}$	$\{C, D\}$
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Estimate	1.5	1.5	2.5	2	3	3

Table 1.3: Possible samples, probability for each sample, and **estimate** value in each case.

In this example we can see how the ***statistic*** *at use (****sample mean****) can take 4 different values, depending on which of the six possible samples is selected by the SRS. From here, it is easy to see that the value 1.5 corresponds to two possible samples (Sample 1 and Sample 2). Hence, since each sample has the same probability of being selected (*$\frac{1}{6}$*), the probability that the* ***statistic*** *takes the value 1.5 is:*

$$P(\textit{statistic} = 1.5) = P(\textit{Sample } 1) + P(\textit{Sample } 2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Next, we summarize what are the possible values the ***statistic*** *can take an what is the probability associated to each of them:*

[5] Obviously, we are supposed not to know that for otherwise there will be no need for any kind of statistical analysis !!

$$\textit{statistic value} = \begin{cases} 1.5 & p = \frac{1}{3} \\ 2 & p = \frac{1}{6} \\ 2.5 & p = \frac{1}{6} \\ 3 & p = \frac{1}{3} \end{cases}$$

*In this example, we have seen how the **statistic** can take different values (4 in this case) with different probabilities. Hence, the **statistic** is a random variable*

Hence, given that **statistics** are random variables, it will be necessary to know their main properties and, specially, the probability distributions of the ones that are more frequently used. In this sense, the main **statistics** (or estimators) that are studied are the **sample mean**, the **sample variance**, and the **sample proportion**.

In all cases, we will assume that a sample of size n has been obtained by means of a SRS. The elements of the sample will be denoted by

$$\{x_1, x_2, \cdots x_n\}$$

Also, we will assume that the sample has been selected form a population that follows a given distribution. To know this distribution is very important as it will influence the sampling result and, hence, the possible values of the **statistic** as we have seen in the previous example. Indeed, in this example we have seen that the population is distributed so that there is 1 element with 1 car, 2 elements with 2 cars, and 1 element with 4 cars. Therefore, if we pick the sample element x_i at random from this population, we will have that:

$$p(x_i = a) = \begin{cases} \frac{1}{4} & \textbf{if } a = 1 \\ \frac{1}{2} & \textbf{if } a = 2 \\ \frac{1}{4} & \textbf{if } a = 4 \\ 0 & \textbf{otherwise} \end{cases}$$

This is, in this case, the *distribution* of the population. Figure 1.2 shows it.

In general[6], we will assume that the SAMPLE has been obtained by means of a SRS from a population distributed according to a NORMAL DISTRIBUTION with some POPULATION MEAN μ and some POPULATION VARIANCE σ^2.

What does it mean ? Easy, it means that for any two numbers a and b, we have that for any element in our sample x_i,

$$p(a \leq x_i \leq b) = p(a - \mu \leq x_i - \mu \leq b - \mu) =$$
$$= p(\frac{a-\mu}{\sigma} \leq \frac{x_i - \mu}{\sigma} \leq \frac{b-\mu}{\sigma}) = p(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$$

where Z represents the Standard Normal *distribution*, usually denoted by $N(0, 1)$, whose associated probabilities are found in tables. Graphically, Figure 1.3 shows it

We turn next to the study of the distributions of the three main **statistics**. These, as we have discussed above, will depend on the distribution of the population from which we obtain the sample. For each case, we will also be interested in knowing what is the *expectation* and the *variance* of these statistics.

[6]There are special cases that we will discuss in due time

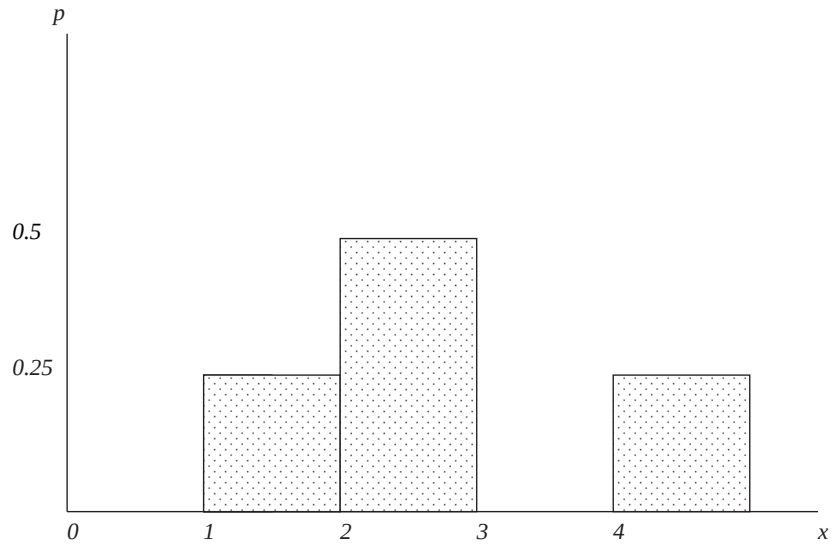

Figure 1.2: Population distribution in example 1.3.3

1.3.1 Sample Mean

Sample mean, denoted by $\bar{X}$, is the **statistic** that is obtained from the sample using the formula:

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

It is normally used to infer conclusions regarding the true value of the *Population mean* μ. Its distribution depends on the characteristics of both the population and the sample

1. If the population is *Normal*, that is, $X_i \sim N(\mu, \sigma^2)\, \forall i$, then we have that

   $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

   because of the *sample mean* being a *linear combination* of *Normal* random variables

2. If the population is not *Normal* but the sample is big enough, then:

   $$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \quad \text{(approx.)}$$

   because of the Central Limit Theorem

3. If the population is not *Normal* and the sample is small, then the distribution of the *sample mean* $\bar{X}$ is unknown in general.

4. If the population variance σ^2 is unknown and the population is *Normal*, then

   $$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

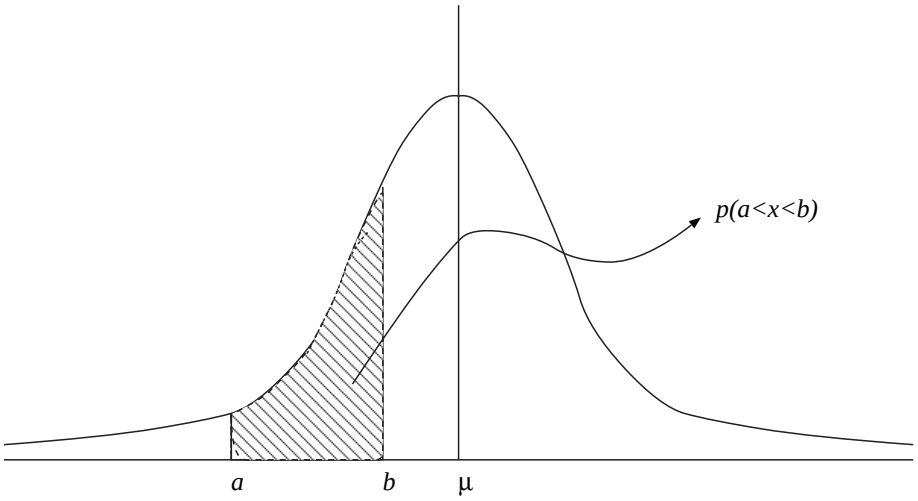

Figure 1.3: Normal Distribution

where S^2 is the *Sample variance* (that we will see next) and t_{n-1} is the $t - student$ *distribution with* $n-1$ *degrees of freedom*, which is very similar to the $N(0,1)$ distribution and whose values can also be found in tables.

We turn next to the study of the *expectation* and *variance* of this statistic. To do so, we will use the mathematical properties of the expectation and variance of a random variable. [7] As usual, we will assume that the sample has been obtained from a population with *population mean* μ and *population variance* σ^2. That is, $E(x_i) = \mu$ and $V(x_i) = \sigma^2$ for any element x_i in the sample. Then,

$$E(\bar{X}) = E(\sum_{i=1}^{n} \frac{x_i}{n}) = \sum_{i=1}^{n} E(\frac{x_i}{n}) = \sum_{i=1}^{n} \frac{1}{n} E(x_i) = \sum_{i=1}^{n} \frac{\mu}{n} = \mu$$

and

$$V(\bar{X}) = V(\sum_{i=1}^{n} \frac{x_i}{n}) = \sum_{i=1}^{n} V(\frac{x_i}{n}) = \sum_{i=1}^{n} \frac{1}{n^2} V(x_i) = \sum_{i=1}^{n} \frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Therefore, for the case of the *sample mean* $\bar{X}$ we have that

$$E(\bar{X}) = \mu, \;\; V(\bar{X}) = \frac{\sigma^2}{n}$$

1.3.2 Sample Variance

Sample variance, denoted by S^2, is the **statistic** that is obtained from the sample using the formula:

$$S^2 = \frac{1}{n-1} \sum (x_i - \bar{X})^2$$

[7] For instance, the expectation of the sum of random variables is the sum of expectations, and so.

It is normally used to infer conclusions regarding the true value of the *Population variance* σ^2. Its distribution depends on the characteristics of the population.

1. If the population is *Normal*, ($X_i \sim N(\mu, \sigma^2)\, \forall i$), then:

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

   where χ^2_{n-1} is the *chi-square distribution with* $n-1$ *degrees of freedom*, whose values are also in tables. (This distribution corresponds to the sum of $n-1$ squared standard *Normals*)

2. If the population is not *Normal*, then the distribution of the *sample variance* is unknown in general, even for large samples.

Since we only know the distribution of the *sample variance* when the population is *Normal*, we will use the fact that in that case its distribution is χ^2_{n-1} to find the expectation and variance easily. In this sense, we know that for any χ^2 variable we have that $E(\chi^2_{n-1}) = n-1$ and $V(\chi^2_{n-1}) = 2(n-1)$. Hence, we will assume the the sample has been obtained from a *Normal* population with *sample mean* μ and *sample variance*l σ^2. That is, $x_i \sim N(\mu\, \sigma^2)$ for any element x_i in the sample. Hence:

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

and therefore

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

$$V(\frac{(n-1)S^2}{\sigma^2}) = 2(n-1) \Rightarrow \frac{(n-1)^2}{(\sigma^2)^2}V(S^2) = 2(n-1) \Rightarrow V(S^2) = \frac{2\sigma^4}{n-1}$$

1.3.3 Sample Proportion

Sample proportion is a special case. It is used when we are interested in knowing which is the true *proportion* of elements in a population that have a given characteristic. For instance, it might be of interest to know what is the proportion of smokers among the second year students in this school (in this case, the *characteristic* that is of interest is "whether a student smokes or not"), or what is the proportion of faulty Pentium IV chips in a day (in this case, the *characteristic* of interest is "whether a chip is faulty or not")
Sample proportion, denoted by $\hat{\pi}$, is the **statistic** that is obtained from the sample using the formula:

$$\hat{\pi} = \sum \frac{x_i}{n}$$

where $x_i = 1$ if the i-th element in the sample has the characteristic that we are studying and $x_i = 0$ if it does not.
Sample proportion is normally used to infer conclusions regarding the true *population sample* π. In this case, the population is never *Normal* since each observation x_i comes from a Bernoulli random variable. Indeed, let us assume that we are looking at a

population of 100 individuals out of which 45 are smokers. That is, the true *population proportion* is 45% or $\pi = 0.45$. Imagine that from this population we want to obtain a sample of size 10. It is clear that for any element x_i of the sample we will have that:

$$p(x_i = 1) = \frac{45}{100} = 0.45$$

$$p(x_i = 0) = \frac{55}{100} = 0.55$$

Hence, we see that each element x_i in the sample follows a *Bernoulli* distribution with parameter π (where π is the true and unknown *population proportion*
It can be shown then that $\hat{\pi} = \sum x_i/n$ is a *Binomial* random variable. Also, given that when samples are large a *Binomial* distribution can be approximated by a *Normal* distribution, we can conclude that, in general:

1. If the sample is large enough, then (approx.):

$$\hat{\pi} \sim N(\pi, \frac{\pi(1-\pi)}{n})$$

   This approximation is better the closer to $0,5$ is π and the larger is the sample

2. If the sample is not large, then the approximation is very bad.

With regards to the expectation and variance of the *sample proportion*, we have:

$$E(\hat{\pi}) = \pi$$

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

1.4 Exercises

1. In each of the sentences below, identify the population, the parameter and the estimate

   (a) A survey conducted with 1000 youngsters between 15 and 17 years old reveals that 432 are regular smokers

   (b) According to a survey conducted by the Ajuntament de Barcelona, one out of three people has received a traffic sanction during 2003.

   (c) A media recording factory produces 50.000 CD ROMs a day, being 25 the average number of faulty units.

2. Let $x_1, x_2, \ldots, x_n$ be a random sample drawn from a population distributed according to a Normal with expectation $\mu$and variance σ^2. In the formulae below, which ones correspond to an estimator ?

   (a) $\sum x_i - \mu$

   (b) $\sigma x_1 + \sigma x_2$

   (c) $x_i, i = 1, 2, \ldots, n$

(d) $x_1^2 + x_2^2 - e^{x_3}$

(e) $\frac{x_i - \mu}{\sigma}, i = 1, 2, \ldots, n$

(f) $\sum (x_i - \bar{X})^2$

3. Based on past cases, we know that the average score in a given quiz is 100, being 125 the standard deviation. Compute the probabilities below for the case when 100 people take the same quiz.

   (a) $P(98.5 < \bar{X} < 101.5)$

   (b) $P(96 < \bar{X} < 104)$

   (c) $P(\bar{X} > 102)$

   (d) $P(\bar{X} < 97.5)$

4. A professional electrician wonders about buying a large quantity of light bulbs to a manufacturer. The later claims that, in average, his bulbs last for 1000 hours, with a standard deviation of 80 hours. The electrician decides that he will buy the bulbs only if in a random sample of 64 bulbs the average live is at least 1000 hours. What is the probability that the electrician does finally buy the bulbs ?

5. A TV sets producer wants to estimate how long does it take (in average) for one of his appliances to malfunction. He wants to do it so that the probability of the difference between the estimate and the true value being more that 10 hours is 0.05. Assuming that the standard deviation is 100, how large should the sample be ?

6. Using the χ^2 table, find the values for χ_1^2 and χ_2^2 such that $P(\chi^2 > \chi_1^2) = 0.95$ and $P(\chi^2 > \chi_2^2) = 0.05$ when the degrees of freedom are 5, 10, 20, 60 and 100.

7. The manager of a manufacturing plant wants to know the variation in the thickness of a plastic element that they produce. It is known by engineering analysis that the distribution of the thickness in that kind of manufacturing processes is Normal with a standard deviation of 0.01 cm. A random sample consisting of 25 such pieces yields a sample standard deviation of 0.015 cm. The manager is surprised, if the population variance is $(0.010)^2$, what is the probability that the sample variance is larger or equal that $(0.015)^2$?

8. Having a random sample of size $n = 16$ drawn from a Normal distribution with unknown mean and variance, find $P(S^2/\sigma^2 \leq 2.041)$.

Chapter 2

Estimation

2.1 General Criteria for Estimation

We have seen in the previous chapter the main statistics used in statistical inference. In Definition 1.3.1 we have stressed that the concept of "statistic" can have different names depending of its use. In this sense, in this chapter we will use the term **estimator** as we will be using different statistics to obtain approximations (i.e. estimations) to the true value of the population parameter that is of interest. Later, in other chapters, we will return to the term **statistic** since we will not use the statistics to do estimations but as a part of a more complex analysis.

2.2 Properties of Estimators

Once the main statistics and their probabilistic features (i.e. probability distribution, expectation and variance) are known, we focus in this chapter on the "good" properties that we would like estimators to have in order for them to provide good approximations to the parameter. In this sense, an estimator might, among others, satisfy the properties of being *unbiassed*, *efficient*, and *consistent* that we will see next. After that, we will learn how this estimators can be used to produce conclusions (very preliminary at this point) regarding the true population parameters. Point estimation and confidence intervals will be the techniques that we will use. Finally, more advanced topics will be introduced. *Maximum likelyhood* estimation will allow us to design good estimators for the case we do not know which one to use. The *Cramer-Rao bound* will help us to know if one specific estimator is efficient.

2.2.1 Bias

Definition 2.2.1 *Let $\hat{\theta}$ be an estimator of the population parameter θ. The* bias *of $\hat{\theta}$ is defined as the difference between the expected value of the estimator and the true value of the population parameter*

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

Definition 2.2.2 *An estimator $\hat{\theta}$ is said to be an unbiassed estimator of the population parameter θ if it has zero bias.*

$$B(\hat{\theta}) = 0 \text{ or } E(\hat{\theta}) = \theta$$

Example 2.2.3

$$E(\bar{X}) = \mu \Rightarrow \bar{X} \text{ is an unbiassed estimator of } \mu$$

Example 2.2.4

$$E(S^2) = \sigma^2 \Rightarrow S^2 \text{ is an unbiassed estimator of } \sigma^2$$

Example 2.2.5

$$E(\hat{\pi}) = \pi \Rightarrow \hat{\pi} \text{ is an unbiassed estimator of } \pi$$

The interpretation of the unbias property is simple. For what we have seen in the previous chapter, we know that an estimator is a random variable, that is, takes different values with different probabilities. Hence, it is clear that it is highly unlikely that the specific value (estimate) that we get once we apply the sample to the estimator exactly coincides with the true parameter value. What the unbias property means is that the above is true "in the sense of expectation". In other words, although when we apply the specific sample we have to the estimator the estimate will not coincide (in general) with the true value of the parameter, if we had 100 different samples to apply to the estimator then the *average* of the 100 different estimates produced would be very close to the true parameter value. This kind of approximation would be more precise the larger is the number of samples to use.

We can compare an estimator with a "shooter" whose target is the true value of the parameter. A good "shooter" (unbiassed) always aims at the center of the target, although there is always a small probability that the shot slightly deviates from the center. On the contrary, a bad "shooter" (biassed) never aims at the center of the target.

2.2.2 Efficiency

The efficiency criterion for an estimator, that we will see next , has two different versions depending on whether the estimator is biassed or unbiassed.

2.2.2.1 Unbiassed Estimators

Definition 2.2.6 *Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiassed estimators of θ. Then, the more efficient estimator is that of the lesser variance.*

2.2.2.2 Biassed Estimators

Definition 2.2.7 *Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be any two estimators of θ. Then, the more efficient estimator is that of the lesser Mean Quadratic Error (MQE) where:*

$$MQE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + B(\theta)^2$$

It is easy to see that, in fact, the second "version" contains the first one as a special case. Indeed, it an estimator has zero bias's then its MQE and Variance are equal.

Example 2.2.8 *Let us consider the following alternative estimators of the population mean μ which will be applied to a sample obtained from a population with population mean μ and population variance σ^2*

$$\hat{\mu}_1 = \frac{x_1 + x_2 + x_3}{3}$$

$$\hat{\mu}_2 = \frac{x_1 + x_2}{2}$$

Let us check first the bias's of each of these estimators:

$$B(\hat{\mu}_1) = E(\hat{\mu}_1) - \mu = E(\frac{x_1 + x_2 + x_3}{3}) - \mu = \frac{1}{3}(E(x_1) + E(x_2) + E(x_3)) - \mu = \frac{1}{3}3\mu - \mu = \mu - \mu = 0$$

$$B(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu = E(\frac{x_1 + x_2}{2}) - \mu = \frac{1}{2}(E(x_1) + E(x_2)) - \mu = \frac{1}{2}2\mu - \mu = \mu - \mu = 0$$

Hence, both estimators and unbiassed. Let us now check which one has less variance:

$$V(\hat{\mu}_1) = V(\frac{x_1 + x_2 + x_3}{3}) = \frac{1}{9}(V(x_1) + V(x_2) + V(x_3)) = \frac{1}{9}3\sigma^2 = \frac{\sigma^2}{3}$$

$$V(\hat{\mu}_2) = V(\frac{x_1 + x_2}{2}) = \frac{1}{4}(V(x_1) + V(x_2)) = \frac{1}{4}2\sigma^2 = \frac{\sigma^2}{2}$$

Therefore, $\hat{\mu}_1$ is more efficient as it has less variance ($\frac{\sigma^2}{3} < \frac{\sigma^2}{2}$)

The intuition behind the efficiency of an estimator is also clear. If we compare an unbiassed estimator with a "good shooter" (as we have done before) that always aims at the center of the target, then an estimator is more *efficient* than another one if it "trembles" less. In other words, the more efficient estimator is the one whose values are more concentrated around the mean.

2.2.3 Consistency

Very often it becomes very difficult to find efficient estimators for a specific parameter. In this case we look at the so called *asymptotic properties*, that consist of the properties that the estimators have when the sample is as large as needed. In this sense, we will introduce the *asymptotic bias's* and the *asymptotic efficiency* or *consistency*.

2.2.3.1 Asymptotically unbiassed estimators

Definition 2.2.9 *An estimator $\hat{\theta}$ of the population parameter θ is said to be asymptotically unbiassed if its bias vanishes as the sample size goes to infinity. Formally, $\hat{\theta}$ is an unbiassed estimator of θ if*

$$\lim_{n\to\infty} B(\hat{\theta}) = 0$$

Example 2.2.10 *Let us consider the following estimator of the population variance (σ^2)*

$$\tilde{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$$

It is easy to check that if

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$$

then

$$\tilde{S}^2 = \frac{n-1}{n}S^2$$

and hence

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

Therefore

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

That is, $\tilde{S}^2$is a biassed estimator of σ^2 since $E(\tilde{S}^2) \neq \sigma^2$. Nevertheless, $\tilde{S}^2$ is an asymptotically unbiassed *estimator of σ^2, for its bias vanishes as the sample grows. Indeed,*

$$\lim_{n\to\infty} B(\tilde{S}^2) = \lim_{n\to\infty} -\frac{\sigma^2}{n} = 0$$

2.2.3.2 Consistent Estimators

The property of consistency not only considers the behavior of the bias as the sample grows large, but also looks at the variance. That is, *consistency* refers to the behavior of the MQE of the estimator as the sample size goes to infinity.

Definition 2.2.11 *An estimator $\hat{\theta}$ of the population parameter θ is said to be consistent it its Mean Quadratic Error vanishes as the size of the sample goes to infinity. Formally, $\hat{\theta}$ is a consistent estimator of θ if*

$$\lim_{n\to\infty} EQM(\hat{\theta}) = 0$$

Example 2.2.12 *Let us consider the estimator of σ^2 that we have seen before, $\tilde{S}^2$. We already know that it it a biassed estimator for σ^2 and that its bias is $B(\tilde{S}^2) = -\frac{\sigma^2}{n}$. We will compute now its variance in order to study the behavior of its EQM as the sample size goes to infinity*

$$V(\tilde{S}^2) = V(\frac{n-1}{n}S^2) = (\frac{n-1}{n})^2 V(S^2) = \frac{(n-1)^2}{n^2}\frac{2(\sigma^2)^2}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$

Hence

$$EQM(\tilde{S}^2) = V(\tilde{S}^2) + B(\tilde{S}^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + (-\frac{\sigma^2}{n})^2 = \frac{(2n-1)\sigma^4}{n^2}$$

and then

$$\lim_{n\to\infty} EQM(\tilde{S}^2) = \lim_{n\to\infty} \frac{(2n-1)\sigma^4}{n^2} = 0$$

Therefore, $\tilde{S}^2$ is a consistent estimator of σ^2

2.3 Point Estimation

A *point estimation* is the simplest method to produce estimations for a population parameter, that is, an approximation to its true value. To obtain a point estimation or *estimate* we just need to apply our *estimator* to the specific sample at hand.

Example 2.3.1 *Imagine that we want to obtain an approximation to the true value of the population mean μ of a given population. For what we have seen before, we know that the sample mean $\bar{X}$ is a good estimator of μ for it is unbiassed*[1]*. Hence, this will be the estimator we use. Imagine that the sample we have is*

$$\textit{Sample} = \{1, 2, 3, 4\}$$

Then

$$\bar{X} = \frac{1+2+3+4}{4} = 2.5$$

*Hence, in this case the **point estimation** (or estimate) we get for μ is 2.5*

This method of estimation has the good property that is quick and simple. The main drawback, though, is that gives very little information and, moreover, with very little precision. In the previous example, we know that the sample mean is an unbiassed estimator of the population mean. Hence, the true value of μ will be "around" 2.5, but we do not have any further information (above 2.5 ? below 2,5 ? close to 2.5 ? far from 2.5 ? ...). In other words, we do not know anything about the accuracy of this estimate. This can be solved, in some sense, using a different estimation method.

2.4 Confidence Intervals

We will use now the knowledge we have about the probability distribution of the sample statistics to supplement the point estimation with additional information. In this way, we will produce an *interval* that will contain, with some probability, the true value of the unknown population parameter.

That is, we will be able now to "measure" the accuracy of our estimation. In this sense, the outcome of an *estimation by confidence intervals* will be something similar to (for the case of the mean):

$$\mu \in [2.25\,,\ 2.75] \text{ with probability } 95\%$$

The intervals obtained using this method are called **confidence intervals**, and the probability that the population parameter lies within this interval is the **confidence level**, usually denoted by $1-\alpha$.

[1] We will see later that, moreover, it is the most efficient estimator of the population mean

2.4.1 Confidence Interval for the mean

We will see next how to build the confidence interval for the case when we need to produce an estimation for the population mean μ

2.4.1.1 Case I: *Normal* Population or large sample (σ^2 known)

We know that in this case,

$$\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

hence

$$p(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha$$

where $z_{1-\frac{\alpha}{2}}$ is the value that corresponds to a $N(0,1)$ whose left tail contains an area of $1-\frac{\alpha}{2}$. That is,

$$P(Z \leq z_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$$

where Z represents a $N(0,1)$ and this value can be found in tables.
Doing some algebra inside the inequalities we get,

$$p(-\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \leq -\mu \leq -\bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}) = 1-\alpha$$

multiplying by -1 we reverse the "direction" of the inequalities, and hence

$$p(\bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} \geq \mu \geq \bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}) = 1-\alpha$$

at the end we get the interval we were looking for,

$$\mu \in [\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}] \text{ with probability } 1-\alpha$$

2.4.1.2 Case II: *Normal* Population or large sample (σ^2 unknown)

In the previous case we need to know the true value of the population variance σ^2 in order to compute the interval. This is highly unusual. To overcome this problem we can replace σ^2 by its unbiassed estimator S^2. The only difference is that now we can not use the $N(0,1)$, but the $t-student$ with $n-1$ degrees of freedom.

$$\mu \in [\bar{X} - t_{1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}}, \bar{X} + t_{1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}}] \text{ with probability } 1-\alpha$$

where $t_{1-\frac{\alpha}{2}}$ is the value that corresponds to a $t-student$ whose left tail contains an area of $1-\frac{\alpha}{2}$ and that can be found in tables as well.

(when n is large, then $t_{1-\frac{\alpha}{2}}$ is approximately equal to a $z_{1-\frac{\alpha}{2}}$)

2.4.1.3 Confidence Interval for the variance

In a similar manner, we can also construct a *confidence interval* for the case of the population variance. We must remember, though, that in this case the population must follow a *Normal* distribution for otherwise the distribution of the sample variance S^2 would be unknown. We know the that

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

and hence

$$p(\chi_{\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\frac{\alpha}{2}}) = 1-\alpha$$

where $\chi_{\frac{\alpha}{2}}$ is the value of a χ^2_{n-1} whose left tail contains an area of $\frac{\alpha}{2}$ and that can be found in tables. Similarly, $\chi_{1-\frac{\alpha}{2}}$ is the value of a χ^2_{n-1} whose left tail contains an area of $1-\frac{\alpha}{2}$.

As before, we can work the inequalities out to obtain

$$p(\frac{1}{\chi_{\frac{\alpha}{2}}} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi_{1-\frac{\alpha}{2}}}) = 1-\alpha$$

$$p(\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}}) = 1-\alpha$$

that is,

$$\sigma^2 \in [\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}}] \text{ with probability } 1-\alpha$$

2.4.1.4 Confidence Interval for the proportion

The case of the *proportion* is special for, as said before, the approximation to the *Normal* requires a large sample and that the true population proportion π is close to $\frac{1}{2}$. Therefore, to have a good approximation to the Normal, the confidence interval for the proportion will be different depending on the sample proportion ($\hat{\pi}$) being close to 0.5 or not.

If $\hat{\pi} \approx \frac{1}{2}$

$$\pi \in [\hat{\pi} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}]$$

If $\hat{\pi} \neq \frac{1}{2}$

$$\pi \in [\hat{\pi} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.5(1-0.5)}{n}}, \hat{\pi} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.5(1-0.5)}{n}}]$$

2.5 Maximum Likelihood estimation

So far we have seen that when we need to produce estimations for population parameters that are "standard", ($\mu\, \sigma^2, \pi$), there are *good estimators* at hand: ($\bar{X}, S^2, \hat{\pi}$). We have studied the main features and properties of these estimators.

The problem arises when we need to estimate a different population parameter (for instance the median or the kurtosis) for which do not have a "candidate" for estimator.

The *Maximum Likelihood method* provides a technique to build good estimators of a given population parameter.

The intuition of the method is as follows: After performing a totally random sampling (SRS) we obtain a specific sample, and there must be a reason for it (since we could have obtained a different one). Well, probably we have obtained this specific sample because the parameter value we want to estimate is such that the sample we have obtained is the one with the highest probability of been selected. In this sense, the *maximum likelihood method* finds the value of the parameter that maximizes the probability of obtaining the sample at hand. The process takes three steps, starting with the sample we have, $\{x_1, x_2, \cdots x_n\}$ and the probability density function of the population that contains the parameter (θ) we want to estimate, $f(x;\theta)$. We will first introduce the general method, and later we offer an example to clarify it. Imagine that we want to estimate the parameter θ of a population with a distribution given by $f(x;\theta)$ using the sample that we have obtained $\{x_1, x_2, \cdots x_n\}$. These are the 3 steps:

Step 1 BUILD THE LIKELIHOOD FUNCTION

The Likelihood function is the "formula" that computes the probability of having obtained the sample we have conditional on the population parameter we want to estimate. In other words, is a function (denoted by L) that depends on both the *sample* obtained and the *parameter* we want to estimate:

$$L(x_1, x_2, \cdots x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n; \theta)$$

Since the sample has been obtained from a population with a probability distribution given by $f(x;\theta)$ and that the elements in the sample are independent from each other, the joint probability $P(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n; \theta)$ can be computed as

$$P(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n; \theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta)$$

hence,

$$L(x_1, x_2, \cdots x_n; \theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Step 2 Apply logarithms

The functional form of the likelihood function is often involved (the product of functions), and working directly with it is rather difficult. Hence, using logarithms we can simplify the function so that it becomes easier to deal with. Therefore, in this step we simply apply "ln" and then use the properties of logarithms in order to simplify the form of the likelihood function

$$\ln L(x_1, x_2, \cdots x_n) = \ln \prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

Step 3 MAXIMIZE

The last step is to maximize the likelihood function, that is, to find the value of θ that maximizes the function L (the probability of having obtained the sample we have). Thus, we must compute the derivative of the likelihood function L with respect to

the parameter θ and make it equal to zero to find the value of θ that maximizes it. Usually this is complicated, that is why we have applied logarithms in Step 2. Indeed, since the function "logarithm" is strictly increasing, the value of θ that maximizes $\ln L$ maximizes L as well. Hence, in practice, what we do is:

$$\frac{\partial \ln L(x_1, \cdots x_n; \theta)}{\partial \theta} = 0$$

and from here we find the value of θ solves the above equation. The solution will be the *maximum likelihood estimator* of θ, usually denoted by $\hat{\theta}_{MV}$

Example 2.5.1 LET $\{x_1, x_2, \cdots x_n\}$ BE A SAMPLE (INDEPENDENT) OBTAINED FROM A NORMAL POPULATION WITH POPULATION MEAN μ AND POPULATION VARIANCE σ^2. FIND THE MAXIMUM LIKELIHOOD ESTIMATOR OF μ.

First, let us remember what is the probability density function corresponding to a $N(0, \sigma^2)$*:*

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

Step 1 LIKELIHOOD FUNCTION

$$L(x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

This would be hard to work with !. That's why we need to use logarithms.

Step 2 LOGARITHMS

$$\ln L(x_1, \cdots x_n) = \ln\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$$

It still looks hard, but after using some of the properties of logarithms[2] *the simplification will be important*

$$\ln\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n + \ln e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2} =$$

$$= \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \ln e$$

Hence

$$\ln L(x_1, \cdots x_n) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

Pas 3 MAXIMIZE

[2]The logarithm of the product is the sum of logarithms, etc.

We have to compute the derivative of $L(x_1, \cdots, x_n)$ with respect to μ and equate it to zero.

$$\frac{\partial L(x_1, \cdots x_n)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right) =$$

$$= \frac{\partial}{\partial \mu}\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n - \frac{\partial}{\partial \mu}\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 =$$

$$= 0 - \frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}\left(\frac{x_i-\mu}{\sigma}\right)^2 = -\frac{1}{2}\sum_{i=1}^{n}2\left(\frac{x_i-\mu}{\sigma}\right)(-\frac{1}{\sigma}) = \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma^2}\right)$$

Hence,

$$\frac{\partial L(x_1, \cdots x_n)}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma^2}\right) = 0 \Rightarrow \frac{1}{\sigma^2}(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mu) = 0$$

and finally,

$$\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}\mu \Rightarrow \sum_{i=1}^{n}x_i = n\mu \Rightarrow \mu = \frac{\sum_{i=1}^{n}x_i}{n}$$

That is, the maximum likelihood estimator of the population mean μ is the sample mean $\bar{X}$

$$\hat{\mu}_{MV} = \frac{\sum_{i=1}^{n}x_i}{n} = \bar{X}$$

2.6 The Cramer-Rao lower bound

In section 1.2.2 we have seen that one of the "good" properties of an estimator is that of *efficiency*, that is, having a variance as low as possible (and accuracy as high as possible). Nevertheless, we have seen that this is a "relative" property in the sense that we are not able to tell whether one estimator is the "most" efficient[3] but only to compare a few of them and then say which one has the lower variance.

The CRAMER-RAO LOWER BOUND that we will see next, allows us to know which is the *minimum variance* that any unbiassed estimator of a given parameter can have. Hence, if we find an unbiassed estimator and find that its variance reaches this bound, then we can be sure that it is, at least, as efficient as any other unbiassed estimator.

We define next what this lower bound is,

Definition 2.6.1 *Given a population parameter θ of a population with probability density function given by $f(x;\theta)$, the* CRAMER-RAO LOWER BOUND *establishes which is the lowest variance of any unbiassed estimator $\hat{\theta}$ of this parameter. It is computed as*

$$C - R = \frac{1}{nE\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right]}$$

[3]That would imply being able to compare any estimator with "all" other possible estimators !

Hence, for any unbiassed estimator $\hat{\theta}$ of the population parameter θ we have that $V(\hat{\theta}) \leq C - R$. Therefore, if we find an unbiassed estimator whose variance equals $C - R$, then we can say that it is the "most" efficient, no other unbiassed estimator can have a lower variance.

To obtain this bound $C - R$,, we must perform every computation in the formula that defines the Cramer-Rao lower bound:

1. Do $\ln f(x;\theta)$ and apply the properties of logarithms to simplify as much as possible
2. Compute the derivative $\frac{\partial \ln f(x;\theta)}{\partial \theta}$
3. Square the previous result $\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2$
4. Compute the expectation of the previous result $E\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right]$ (usually, this is the most difficult step)
5. Multiply by n, $nE\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right]$
6. Finally, invert the result above

$$\frac{1}{nE\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right]}$$

Example 2.6.2 FIND THE CRAMER-RAO LOWER BOUND FOR ANY UNBIASSED ESTIMATOR OF THE POPULATION MEAN μ OF A NORMAL POPULATION WITH POPULATION VARIANCE σ^2

Remember the density function of a $N(\mu, \sigma^2)$:

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

From here, we follow the 6 steps described above to obtain the $C - R$ lower bound in this case.

1. *Logarithms*

$$\ln f(x;\mu\,\sigma^2) = \ln\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

2. *Derivative*

$$\frac{\partial \ln f(x;\mu\,\sigma^2)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) = \frac{x-\mu}{\sigma^2}$$

3. *Square it*

$$\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2 = \left(\frac{x-\mu}{\sigma^2}\right)^2$$

4. *Expectation (the trickiest step !)*

$$E\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right] = E\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4}E(x-\mu)^2 =$$

$$\frac{1}{\sigma^4}E(x-E(x))^2 = \frac{1}{\sigma^4}V(x) = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}$$

5. *Multiply by* n

$$nE\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right] = \frac{n}{\sigma^2}$$

6. *Finally, invert*

$$\frac{1}{nE\left[\left(\frac{\partial \ln f(x;\theta)}{\partial \theta}\right)^2\right]} = \frac{\sigma^2}{n}$$

Therefore, the Cramer-Rao lower bound, $C-R$*, in this case is*

$$C - R = \frac{\sigma^2}{n}$$

Hence, any unbiassed estimator of the population mean μ *will, necessarily, have a variance greater or equal to* $\frac{\sigma^2}{n}$*. Remenber now that* $V(\bar{X}) = \frac{\sigma^2}{n}$*, and hence the* SAMPLE MEAN $\bar{X}$ IS THE MOST EFFICIENT UNBIASSED ESTIMATOR OF μ*.*

2.7 Exercicis

1. The *Mean Square Error* of an estimator $\hat{\theta}$ is defined as $MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2$. Prove that $EQM(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$

2. Assuming that $X_i \sim N(\mu, \sigma^2)$, which of the statistics below are unbiased estimators of μ ?

   (a) $\hat{\mu}_1 = \frac{X_1+X_2+X_3+X_4}{4}$

   (b) $\hat{\mu}_2 = \frac{2(X_1+X_2)}{6} + \frac{X_3+X_4}{6}$

   (c) $\hat{\mu}_3 = \frac{X_1-X_2+X_3-X_4}{4}$

   Among all the unbiased estimators, which one is the most efficient ? Which one is the most efficient among all three estimators ?

3. Imagine that we have a random sample of size n drawn from a population $N(\mu, \sigma^2)$ and we want an estimate for μ. Among all the estimators for μ that are of the form:

$$\hat{\mu} = \lambda x_1 + \theta x_2$$

   find the values for λ and θ so that the estimator in unbiased and has the minimum variance.

4. A random sample of hourly wages for nine mechanics yields the following data:

$$10.5, 11, 9.5, 12, 10, 11.5, 13, 9, 8.5$$

Assuming that the sample is obtained from a Normal population, find the confidence intervals for the average hourly wage (both, when $\alpha = 0.05$ and $\alpha = 0.0.1$) when:

(a) It is known that $\sigma^2 = 1.5$

(b) σ^2 is unknown

5. The thickness of the metal pieces that a machine produces is expected to present some fluctuation. A random sample of 12 pieces i selected and the thickness of each of them is recorded, which yields

$$12.6, 11.9, 12.3, 12.8, 11.8, 11.7, 12.4, 12.1, 12.3, 12.0, 12.5, 12.9$$

Assuming that thickness is a Normal random variable, obtain a 95% confidence interval for the variance of thickness.

6. A manufacturer claims that the percentage of faulty items in any lot of the articles he produces is 1%. A random sample of 200 articles is selected and 8 are found to be faulty. Find 95% and 99% confidence intervals for the true proportion of faulty items. Based on these results, what can you say about the manufacturer's claim ?

7. A physician is interested in the proportion of men that smoke and develop lung cancer. The physician wants to select a sample of smokers and observe whether they develop cancer or not. What has to be the sample size so that with a 95 % probability the difference between the sample proportion and the true proportion is less than 0.02 ?

8. Let $x_1, x_2, \ldots, x_n$ a random sample drawn from a Poisson distribution with true parameter λ. Compute the maximum likelihood estimator of λ.

9. Let $x_1, x_2, \ldots, x_n$ a random sample drawn from a Exponential distribution with true parameterλ. Compute the maximum likelihood estimator of λ.

10. By means of the Cramer-Rao lower bound, find the variance of the most efficient unbiased estimator of $\lambda$when the sample is drawn from a population distributed according to an exponential:

$$f(x, \lambda) = \frac{1}{\lambda} e^{\frac{-x}{\lambda}}, \quad x > 0$$

Prove that the sample mean is an efficient estimator of λ.

Chapter 3

Hypothesis Testing

3.1 Hypothesis Testing

So far we have learned how to approximate the value of a population parameter by means of some estimation technique. In many cases, though, what is of interest is not knowing what is the value of the parameter but rather some question regarding the parameter. For instance: Is the average wage in Cerdanyola this year greater than in the previous year ? Does this Pentium Chip works at 3Ghz ? Is the proportion of smokers in the UAB smaller than in the UPF ?

In all these cases, we are interested in *testing* whether a belief, idea, or conjecture regarding the population parameter seems true or not. To do so, to *test our hypothesys*, we must base our analysis upon the data we have, *the sample*, because this is the only information we have regarding the population. We can now be more precise in the definition of "hypothesis testing"

Definition 3.1.1 *Hypothesis testing is a statistical technique by means of which we can verify whether the data in the sample backs up, or not, a specific hypothesis stablished on some population parameter.*

In general, the "structure" of the hypothesis testing technique can be decomposed into 6 steps. To understand how the procedure works, let's imagine that we want to test whether the population parameter θ equals θ_0 or not. In this case, the 6 steps we have mentioned above are:

1. To specify the *Null Hypothesis*. This is the hypothesis that we believe is true and that we want to test if the data supports it.

$$H_0 : \theta = \theta_0$$

2. To specify an *Alternative Hypothesis*, which represents what is true when the Null Hypothesis is false. This Alternative Hypothesis may have four different specifications, depending on the information we have on the poulation parameter we are studying.

$$H_1 : \theta \neq \theta_0$$

ó $\quad H_1 : \theta < \theta_0$

$$\text{ó} \quad H_1 : \theta > \theta_0$$
$$\text{ó} \quad H_1 : \theta = \theta_1$$

The first kind of Alternative Hypothesis is the more general. It corresponds to the case when there is no information at all regarding the population parameter. In this sense, if the parameter is not what we believe (θ_0,the Null Hypothesis), then we simply specify than it is different from what we think.

The second kind corresponds to the case when there is some information regarding the parameter we are studying. In this specific case such information points out that if the parameter is not what we think, then it must be smaller (for some reason it is known that it can not be larger).

The third kind of Alternative Hypothesis is the opposite to the previous case. We use this specification when the information says that if the parameter is not what we think then it must have a larger value.

Finally, the last specification, which is rather rare, corresponds to the case in which there is a lot of information regarding the parameter. In this case, we know that either the parameter takes the value we believe (θ_0) or it takes another specific $\theta_1$value (Finalment, el quart tipus de hipòtesi alternativa es dona rarament i correspon al cas en que es té molta informació sobre el paràmetre que s'estudia de forma que se sap que si no pren el valor alashores l'única possibilitat és que sigui igual a un altre valor θ_1.

Later we will see that the first kind of Alternative Hypothesis produces a *two-tailed test*, whereas the second a the third correspond to a *left-tail test* and a *right-tail test* respectively. Finally, the last kind of hypothesis produces a *left-tail test* or a *right-tail test* depending on whether $\theta_1 < \theta_0$ or $\theta_1 > \theta_0$ respectively.

3. To specify a *test statistic* (*TE*) and to compute the *observed value of the test statistic*, (*OVTE*) using the data in the sample.

   In practice, what distinguishes one hypothesis test from another is the *test statistic* used. Hence, we will see these *test statistics* in detail when we introduce each specific test.

4. To determine what is the *probability distribution* of the test statistic in the previous step under the assumption the the Null Hypothesis is true. This, as in the previous step, depends on what kind of test we are conducting. Hence, we will see the details later.

5. To define a *Rejection Area*, (*RA*) of size α (level of significance). This is the place where the test actually takes place. For this, we need to use the tables that correspond to the distribution determined in step 4 to find a region with the property that if the *null hypothesis* is true then the probability that the *test statistic* lies within this *RA* is α.

   $$p(TE \in RA) = \alpha$$

   In general, this Rejection area consists of only one tail (left- or right-tailed tests) of size α or can be splitted into two symetric tails of size $\frac{\alpha}{2}$ each.

6. Finally, the last step consists of, simply, verify whether the *Observed Value of the Test Statistic*, (*OVTE*) lies, or not, iside the *Rejection Area*. Therefore,

(a) If the *OVTS* is inside the *RA* $\Rightarrow$ THE NULL HYPOTHESIS IS REJECTED

(b) If the *OVTS* is NOT inside the *RA* $\Rightarrow$ THE NULL HYPOTHESIS IS NOT REJECTED

Notice that the final conclusion is always of the form "REJECT" or "DON'T REJECT" the *Null Hypothesis*. The term "ACCEPT" is never used. The reason is as follows: If the output of the test is that the *Null Hypothesis* is REJECTED, then we interpret this as not having "enough empirical evidence" to support the hypothesis. In the same sense, if the test results in the Null Hypothesis being rejected, then the interpretation is that we do not have "enough empirical evidence" against the hypohtesis.

3.2 Hypothesis Testing Types

We will see next what the three basic types of hypothesis testing are:

1. Hypothesis test on the *population mean* μ
2. Hypothesis test on the *population variance* σ^2
3. Hypothesis test on the *population proportion* π

We will learn that all three cases share a common structure, the "six steps" we have learned above. The difference is, mainly, the *test statistic* that will be used in each case. Also, in each case the test can be of one or two tails, depending on the form of the corresponding *alternative hypothesis*.

3.2.1 Hypothesis Test for the Population Mean (μ)

1. NULL HYPOTHESIS
   Is the value of the population mean we want to test $(\mu_0 = \text{value to test})$

$$H_0 : \mu = \mu_0$$

2. ALTERNATIVE HYPOTHESIS
   Corresponds to what would be true if the null hypothesis is false. Depends on what information we have regarding the population mean. There are 4 cases

	INFORMATION REGARDING μ	TEST TYPE
$H_1 : \mu \neq \mu_0$	General case. No information regarding μ. Hence, if it is not equal to μ_0 we can only say that it is different	Two Tails test
$H_1 : \mu > \mu_0$	We have some information regardingμ. This information states that if it is not equal to μ_0 the it is larger	right-tail test
$H_1 : \mu < \mu_0$	We have some information regarding μ. This information states that if it is not equal to μ_0 then it is smaller	left-tail test
$H_1 : \mu = \mu_1$	We have a lot of information regarding μ. We know that if it is not equal to μ_0 then it must be equal to the value μ_1	right-tail test if $\mu_1 > \mu_0$ or left-tail test if $\mu_1 < \mu_0$

3. TEST STATISTIC
   The Test Statistic (*TE*) to use in this case depends on whether we know the population variance σ^2 or not.

σ^2 KNOWN	σ^2 UNKNOWN
TE=$\frac{\bar{X}-\mu_o}{\sqrt{\frac{\sigma^2}{n}}}$	TE=$\frac{\bar{X}-\mu_o}{\sqrt{\frac{S^2}{n}}}$

   In any of these cases, the *observed value of the test statistic (OVTE)* is obtained by pluging the values into the corresponding formula, where

$\bar{X}$	Sample Mean
μ_0	Null Hypothesis value
σ^2	Population Variance (if known)
S^2	Sample Variance (if σ^2 is unknown)
n	Sample size

4. DISTRIBUTION OF THE TEST STATISTIC when the Null Hypothesis is true
   As we have learned in previous chapters, we have that if the null hypothesis is true, that is, if $\mu = \mu_0$ then

σ^2 KNOWN	σ^2 UNKNOWN
$\frac{\bar{X}-\mu_o}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$	$\frac{\bar{X}-\mu_o}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$

5. REJECTION AREA of size α
   The way to determine the Rejection Area will different depending on whether the test is of one or two tails.

   (a) TWO TAILS TEST. Corresponds to the case when the Alternative hypothesis is like $H_1 : \mu \neq \mu_0$

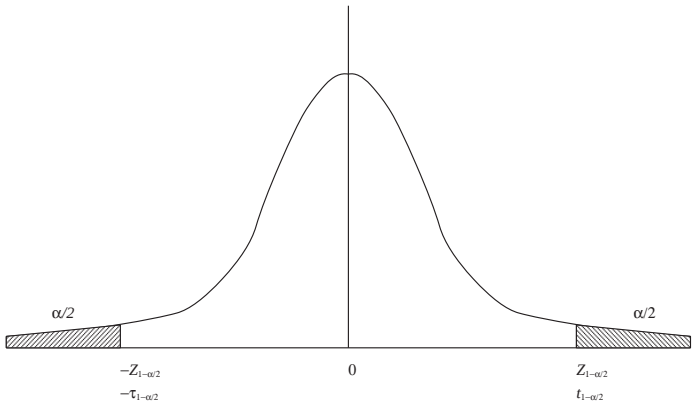

Figure 3.1: Rejection Area in a two tails test

The limit values of the rejection area, $Z_{1-\frac{\alpha}{2}}$ and $t_{1-\frac{\alpha}{2}}$ can be found using the tables of the $N(0,1)$ or $t-student$ with $n-1$ degrees of freedom respectively depending on whether we know σ^2 or not. (See Figure 3.1)

(b) RIGHT-TAIL TEST. Corresponds to the case when we have an alternative hypothesis of the type $H_1: \mu > \mu_0$ (or the type $H_1: \mu = \mu_1 \,\mathrm{i}\, \mu_1 > \mu_0$)

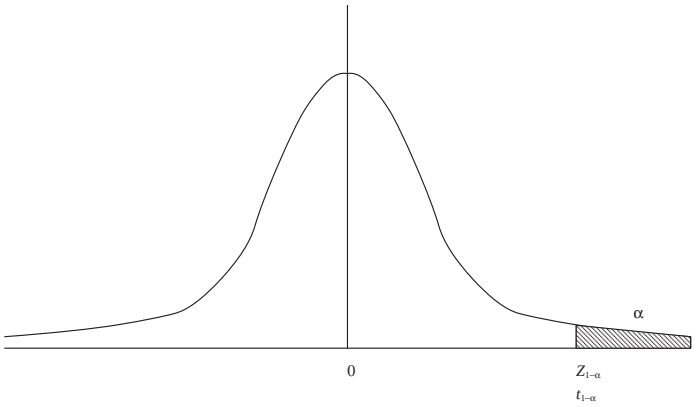

Figure 3.2: Rejection Area in one right-tail test

The limit values of the rejection area, $Z_{1-\alpha}$ i $t_{1-\alpha}$ can be found in the tables of the $N(0,1)$ or $t-student$ with $n-1$ degrees of freedom respectively depending on whether we know σ^2 or not. (See Figure 3.2)

(c) LEFT-TAIL TEST. Corresponds to the case when we have an alternative hypothesis of the type $H_1: \mu < \mu_0$ (or $H_1: \mu = \mu_1 \,\mathrm{i}\, \mu_1 < \mu_0$)

The limit values of the rejection area, $Z_{1-\alpha}$ i $t_{1-\alpha}$ can be found in the tables of the $N(0,1)$ or $t-student$ with $n-1$ degrees of freedom respectively, depending on whether we know σ^2 or not. (See Figure 3.3)

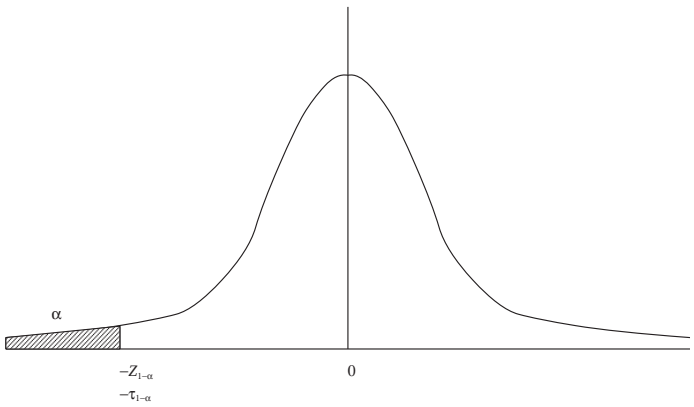

Figure 3.3: Rejection Area in one left-tail test

6. TEST CONCLUSION
   Finally, we have to check if the OBSEVED VALUE OF THE TEST STATISTIC (OVTS) falls, or not, inside the REJECTION AREA. If it does, we then say that the test rejects the NULL HYPOTHESIS. If it does not belong to the rejection area, then we say that the test DOES NOT REJECT THE NULL HYPOTHESIS.

3.2.2 Hypothesis Test for the Population Variance (σ^2)

1. NULL HYPOTHESIS
   Is the value of the population variance that we want to test. ($\sigma_0^2 =$ value to test)

$$H_0 : \sigma^2 = \sigma_0^2$$

2. ALTERNATIVE HYPOTHESIS
   Corresponds to what would be true if the null hypothesis is false. Depends on what information we have regarding the population mean. There are 4 cases

	INFORMATION REGARDING σ^2	TEST TYPE
$H_1 : \sigma^2 \neq \sigma_0^2$	General case. There is no information about σ^2. Therefore, if it does not equal σ_0^2 the only think we can say is that it will be different	Two Tails Test
$H_1 : \sigma^2 > \sigma_0^2$	We have some information about σ^2 indicating that if it is not equal to σ_0^2 then it must be greater	Right-Tail Test
$H_1 : \sigma^2 < \sigma_0^2$	We have some information about σ^2 indicating that if it is not equal to σ_0^2 then it must be smaller	Left-Tail Test
$H_1 : \sigma^2 = \sigma_1^2$	We have some information about σ^2 indicating that if it is not equal to σ_0^2 then it must be equal to another value σ_1^2	Right-Tail Test if $\sigma_1^2 > \sigma_0^2$ and Left-Tail Test if $\sigma_1^2 < \sigma_0^2$

3. TEST STATISTIC
   The Test Statistic (TE) to use in this case is:

$$\text{T.E.} = \frac{(n-1)S^2}{\sigma_0^2}$$

   The Observed Value of the Test Statistic (OVTS) is obtained by subtituting the corresponding values in the formula, where

σ_0^2	Null Hypothesis Value
S^2	Sample Variance
n	Sample Size

4. DISTRIBUTION OF THE TEST STATISTIC when the Null Hypothesis is true
   From what we know from the previous chapter, if the Null Hypothesis is true, that is, if $\sigma^2 = \sigma_0^2$ then

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

5. REJECTION AREA of size α
   The way to determine the Rejection Area depends on the test being of one or two tails, that is, depending on what is the Alternative Hypothesis.

   (a) TWO-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \sigma^2 \neq \sigma_0^2$

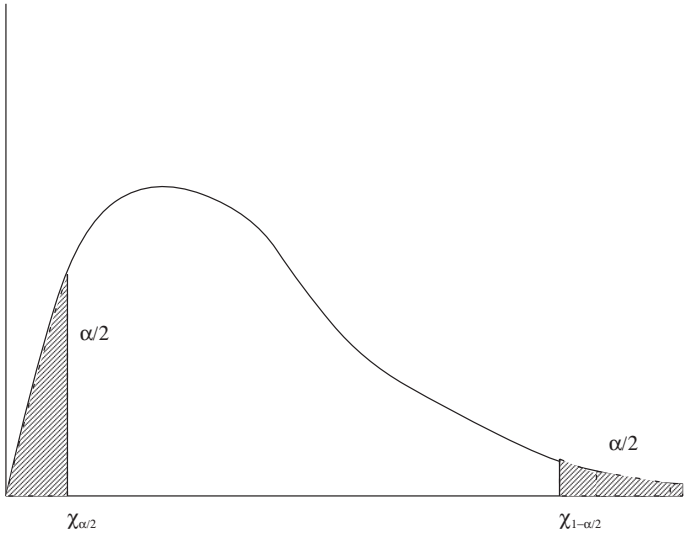

Figure 3.4: Rejection Area in a Two-Tails Test

   Le limit values in the Rejection Area, $\chi^2_{1-\frac{\alpha}{2}}$ and $\chi^2_{\frac{\alpha}{2}}$, can be found using the tables of a χ^2 with $n-1$ degrees of freedom. (See Figure 3.4)

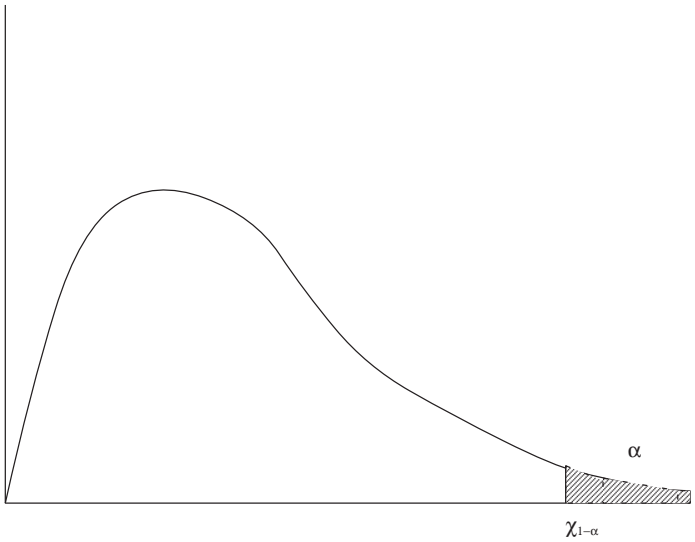

Figure 3.5: Rejection Area in a Right-Tail Test

(b) RIGHT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \sigma^2 > \sigma_0^2$ (also of the type $H_1 : \sigma^2 = \sigma_1^2 \,\text{i}\, \sigma_1^2 > \sigma_0^2$)

The limit value in the Rejection Area, $\chi^2_{1-\alpha}$, can be found in the tables of a χ^2 with $n-1$ degrees of freedom. (See Figure 3.5)

(c) LEFT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \sigma^2 < \sigma_0^2$ (also of the type $H_1 : \sigma^2 = \sigma_1^2 \,\text{i}\, \sigma_1^2 < \sigma_0^2$)

The limit value in the Rejection Area, χ^2_{α}, can be found in the tables of a χ^2 with $n-1$ degrees of freedom. (See Figure 3.6)

6. TEST CONCLUSION
Finally, we have to check if the OBSEVED VALUE OF THE TEST STATISTIC (OVTS) falls, or not, inside the REJECTION AREA. If it does, we then say that the test rejects the NULL HYPOTHESIS. If it does not belong to the rejection area, then we say that the test DOES NOT REJECT THE NULL HYPOTHESIS.

3.2.3 Hypothesis Test for the Population Proportion (π)

1. NULL HYPOTHESIS
Is the value of the Population Proportion that we want to test. ($\pi_0 =$ valor a contrastar)

$$H_0 : \pi = \pi_0$$

2. ALTERNATIVE HYPOTHESIS
Corresponds to what would be true if the null hypothesis is false. Depends on what information we have regarding the population proportion. There are 4 cases

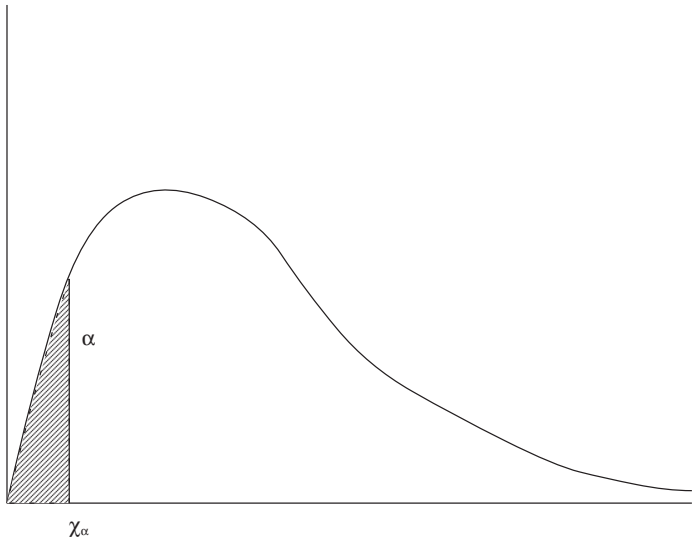

Figure 3.6: Rejection Area in a Left-Tail Test

	INFORMATION REGARDING π	TEST TYPE
$H_1 : \pi \neq \pi_0$	General case. There is no information regarding π. Hence, if it is not equal to π_0 the only thing we can say is that it is different	Two Tails Test
$H_1 : \pi > \pi_0$	We have some information regarding π indicating that if it is not equal to π_0, then it must be greater	Right-Tail Test
$H_1 : \pi < \pi_0$	We have some information regarding π indicating that if it is not equal to π_0, then it must be smaller	Left-Tail Test
$H_1 : \pi = \pi_1$	We have some information regarding π indicating that if it is not equal to π_0, then it must be equal to another value π_1	This is a Right-Tail Test if $\pi_1 > \pi_0$ and a Left-Tail Test if $\pi_1 < \pi_0$

3. TEST STATISTIC
   The Test Statistic (TE) to use in this case is.

$$\text{TE} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

The Ovserved Value of the Test Statistic (OVTS) is obtained when the elements in the formula are replaced by their corresponding values from the sample, where

$\hat{\pi}$	Sample Proportion
π_0	Null Hypothesis Value
n	Sample Size

4. DISTRIBUTION OF THE TEST STATISTIC when the Null Hypothesis is true
   As we have seen in previous chapters, if it is true that $\pi = \pi_0$ then[1]

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$$

5. REJECTION AREA of sizeα
   As in the other tests, the determination of the Rejection Area depends on the test type

   (a) TWO TAILS TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \pi \neq \pi_0$

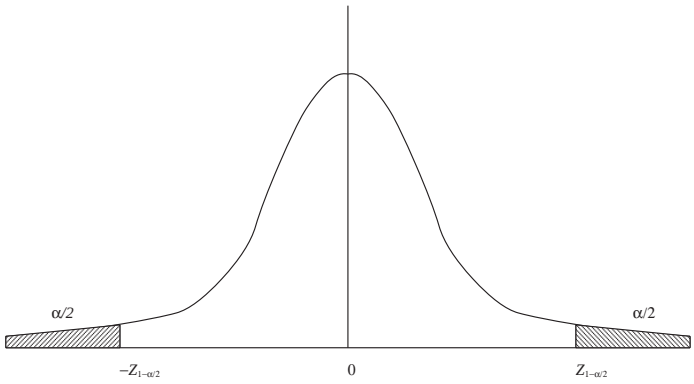

Figure 3.7: Rejection Area in a Two Tails Test

   The limit values of the Rejection Area, $Z_{1-\frac{\alpha}{2}}$ and $-Z_{1-\frac{\alpha}{2}}$, can be found in the table of the $N(0,1)$. (See Figure 3.7)

   (b) RIGHT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \pi > \pi_0$ (or $H_1 : \pi = \pi_1 \,\mathrm{i}\, \pi_1 > \pi_0$)

   The limit value of the Rejection Area, $Z_{1-\alpha}$, can be found in the tables of the $N(0,1)$. (See Figure 3.8)

   (c) LEFT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \pi < \pi_0$ (or $H_1 : \pi = \pi_1 \,\mathrm{i}\, \pi_1 < \pi_0$)

   The limit value in the Rejection Area, $Z_{1-\alpha}$, can be found in the tables of the $N(0,1)$. (See Figure 3.9)

6. TEST CONCLUSION
   Finally, we have to check if the OBSEVED VALUE OF THE TEST STATISTIC (OVTS) falls, or not, inside the REJECTION AREA. If it does, we then say that the test rejects the NULL HYPOTHESIS. If it does not belong to the rejection area, then we say that the test DOES NOT REJECT THE NULL HYPOTHESIS.

[1] Remember that this is an approximation, which is better the larger is the sample n and the closer to $0,5$ is $\hat{\pi}$

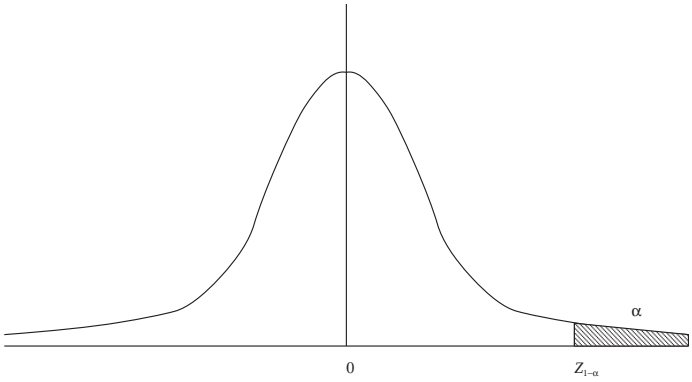

Figure 3.8: Rejection Area in a Right-Tail Test

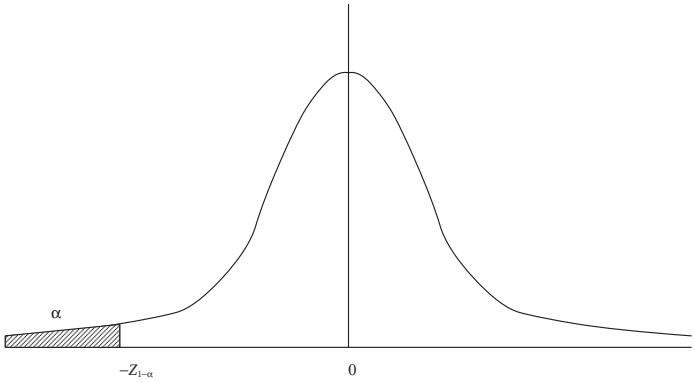

Figure 3.9: Rejection Area in a Left-Tail Test

3.3 Two Samples Tests

In many cases what is of interest is not some question regarding some population parameter (as in the previous section), but rather to compare one parameter in one population with the corresponding parameter in other population. For instance, we might want totest whether the average income in Cerdanyola this year is equal or greater that in the previous year, or if the average income in Cerdanyola is equal to that in Sant Cugat. That is, now we are interested in COMPARING POPULATION PARAMETERS BETWEEN TWO POPULATIONS, eihter two different populations (as when comparing the average income in Cerdanyola and Sant Cugat) or the same population at two different dates or after some action (as when comparing the average income in Cerdanyola this year whith that of the previous year).

In any of these cases, what we do is a TWO SAMPLES TEST. Now we have two populations (Population 1 and Population 2) each one with its corresponding population parameters (μ_1, σ_1^2 and π_1 for the first population and μ_2, σ_2^2 i π_2 for the second population). We then draw two independent samples from each of these populations (Sample 1 and Sample 2) which might have different sizes (n_1 and n_2). From these samples we compute the corresponding Sample Statistics that will be used to perform the tests ($\bar{X}_1$, S_1^2 and $\hat{\pi}_1$ for the first sample and $\bar{X}_2$, S_2^2 i $\hat{\pi}_2$ for the second sample)

Summarized, the information we can gather is:

Population 1	Population 2
μ_1, σ_1^2 and π_1	μ_2, σ_2^2 and π_2
Sample 1	Sample 2
x_{11}	x_{12}
x_{21}	x_{22}
$\vdots$	$\vdots$
$x_{n_1 1}$	$x_{n_2 2}$
$\bar{X}_1$, S_1^2 and $\hat{\pi}_1$	$\bar{X}_2$, S_2^2 and $\hat{\pi}_2$

From here, we can do tests regarding:

1. The difference between the *means* of the two populations: $\mu_1 - \mu_2$

2. The difference between the *variances* of the two populations $\sigma_1^2 - \sigma_2^2$

3. The difference between the *proportions* of the two populations $\pi_1 - \pi_2$

3.3.1 Test for the Difference of Means

We want to test if the difference between the means of two populations equals some specific value δ_0 or not ($\delta_0 = 0$ if we want to test if the means are equal to each other). For instance, we could test if the average income in Cerdanyola and Sant Cugat are equal to each other ($\mu_1 - \mu_2 = 0$). Another example would be to test whether the average sleeping time after taking a new pill equals (or, alternativelly, is larger) than without taking any pill.

For this thest the six corresponding steps are:

1. NULL HYPOTHESIS
   Is the value for the difference that we want to test ($\delta_0 =$ difference value to test)

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

2. ALTERNATIVE HYPOTHESIS
   Corresponds to what would be true if the null hypothesis is false. Depends on what information we have regarding the population proportion. There are 4 cases

	INFORMATION REGARDING μ_1 AND μ_2	TEST TYPE
$H_1 : \mu_1 - \mu_2 \neq \delta_0$	General case. We have no information for the population means. Hence, if the difference is not equal to δ_0 the only thing we can say is that it is different	Two Tails Test
$H_1 : \mu_1 - \mu_2 > \delta_0$	We have some information about the means indicating that if the difference is not equal to δ_0 then it must be greater	Right-Tail Test
$H_1 : \mu_1 - \mu_2 < \delta_0$	We have some information about the means indicating that if the difference is not equal to δ_0 then it must be smaller	Left-Tail Test
$H_1 : \mu_1 - \mu_2 = \delta_1$	We have some information about the means indicating that if the difference is not equal to δ_0 then it must be equal to a known alternative value δ_1	Right-Tail Test if $\delta_1 > \delta_0$ and Left-Tail Test if $\delta_1 < \delta_0$

3. TEST STATISTIC
   The Test Statistic (TS) to use in this case depends on whether the population variances σ_1^2 i σ_2^2 are both known or not.

$$\sigma_1^2 \text{ and } \sigma_2^2 \text{ KNOWN} \qquad \sigma_1^2 \text{ or } \sigma_2^2 \text{ UNKNOWN}$$

$$\text{TS}=\frac{(\bar{X}_1-\bar{X}_2)-\delta_o}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \qquad \text{TS}=\frac{(\bar{X}_1-\bar{X}_2)-\delta_o}{\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}}}$$

   In any of these cases, the Ovserved Value of the Test Statistic (OVTS) is obtained by replacing the correspondig values in the formula, where

$\bar{X}_1$ i $\bar{X}_2$	Sample Means
δ_0	Null Hypothesis Value
σ_1^2 i σ_2^2	Population variances (if known)
S^2	Common Sample Variance (if σ_1^2 or σ_2^2 are not known)
n_1 i n_2	Sample sizes

   In the formulae above the common value for the Sample Variance, S^2(that we use if we do NOT know either σ_1^2 ór σ_2^2) (or any of them) is computed as

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

   where S_1^2 i S_2^2 are the Sample Variances of the first and the second sample respectively. The rason for using this *common estimation of the sample variance* is that for the test to make sense the two populations must be somehow "homogeneous". Tecnichally, this is equivalent to requiring that the two populations have a similar population variance.

4. DISTRIBUTION OF THE TEST STATISTIC WHEN THE NULL HYPOTHESIS IS TRUE
   If it is true that $\mu_1 - \mu_2 = \delta_0$ then

σ_1^2 and σ_2^2 KNOWN	σ_1^2 or σ_2^2 UNKNOWN
$\frac{(\bar{X}_1-\bar{X}_2)-\delta_o}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$\frac{(\bar{X}_1-\bar{X}_2)-\delta_o}{\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}}} \sim t_{n_1+n_2-2}$

5. REJECTION AREA of size α
   The Rejection Area depends on whether we have a Two Tails Test, a Right-Tail Test, or a Left-Tail Test. This, in turn, depends on what is the specification of the Alternative Hypothesis.

   (a) TWO TAILS TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \mu_1 - \mu_2 \neq \delta_0$

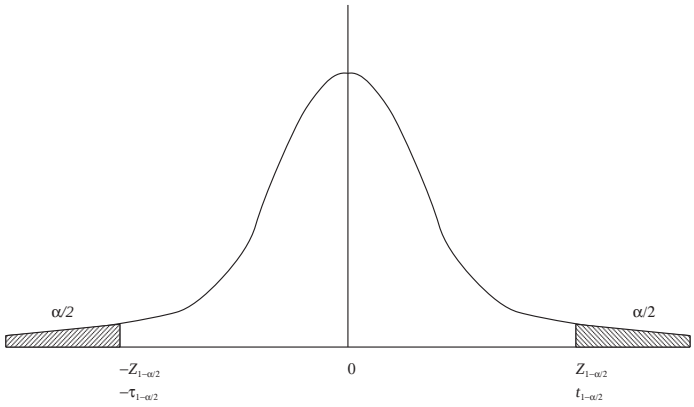

Figure 3.10: Rejection Area in a Two Tails Test

   The limit values of the Rejection Area, $Z_{1-\frac{\alpha}{2}}$ and $t_{1-\frac{\alpha}{2}}$ can be found in the tables of a $N(0,1)$ or a $t-student$ with n_1+n_2-2 degrees of freedom respectively, depending on whether we know the two population variances or not as explained above (See Figure 3.10)

   (b) RIGHT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \mu_1 - \mu_2 > \delta_0$ (or $H_1 : \mu_1 - \mu_2 = \delta_1$ and $\delta_1 > \delta_0$)

   The limit value of the Rejection Area, $Z_{1-\alpha}$ or $t_{1-\alpha}$ can be found in the tables of the $N(0,1)$ or the $t-student$ with n_1+n_2-2 degrees of freedom respectively depending on whether we know the two population variances or not as explained before. (See Figure 3.11)

   (c) LEFT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \mu_1 - \mu_2 < \delta_0$ (or $H_1 : \mu_1 - \mu_2 = \delta_1$ and $\delta_1 < \delta_0$)

   The limit value of the Rejection Area, $Z_{1-\alpha}$ or $t_{1-\alpha}$ can be found in the tables of the $N(0,1)$ or $t-student$ with n_1-n_2-2 degrees of freedom

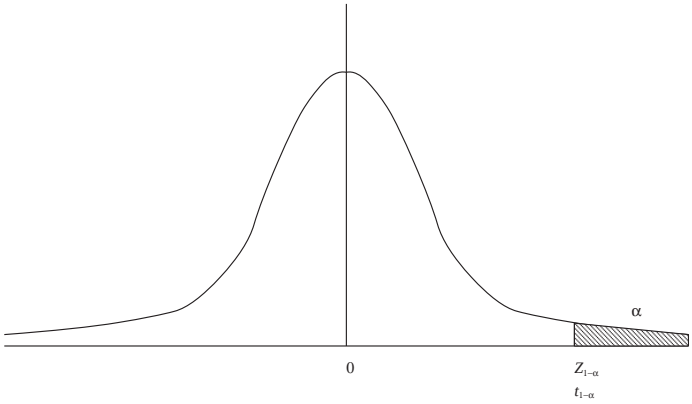

Figure 3.11: Rejection Area in a Right-Tail Test

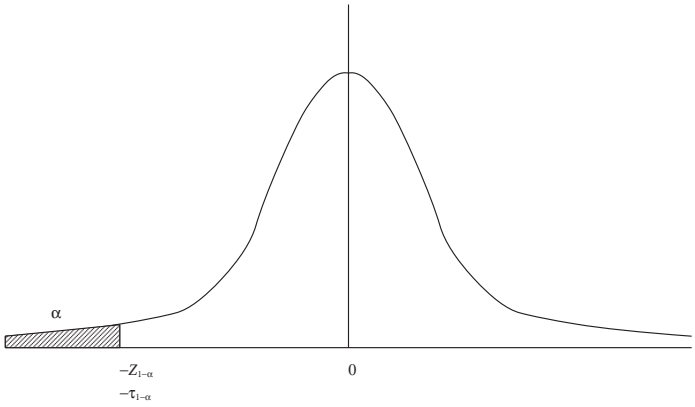

Figure 3.12: Rejection Area in a Left-Tail Test

respectively depending on whether we know the two population variances or not. (See Figure 3.12)

6. TEST CONCLUSION
   Finally, we have to check if the OBSEVED VALUE OF THE TEST STATISTIC (OVTS) falls, or not, inside the REJECTION AREA. If it does, we then say that the test rejects the NULL HYPOTHESIS. If it does not belong to the rejection area, then we say that the test DOES NOT REJECT THE NULL HYPOTHESIS.

3.3.2 Test for the Difference of Variances

In this case we only test if the two populations have the same variance or not. This is special test for three reasons:

1. We can only test if the two variances are equal or not, that is, the Null Hypothesis is always the same
   $$H_0 : \sigma_1^2 = \sigma_2^2$$

2. The test must be conducted following a strict order, that must be stablished in a "extra" step before starting the usual 6 steps

3. This test is important because it allows to check if two different populations seem to have the same variance. This, as we have seen before, is important for other tests. Indeed, the test for the difference of means only makes sense if the two populations are "homogeneous", that is, have a similar variance.

Hence, this "special" test will begin with an extra step (Step 0) where we stablish the order of the elements of the test.

0. EXTRA STEP

We change the "denomination" of our two samples so that **ALWAYS** the sample with the highest Sample Variance is the Sample 1, being the Sample 2 the one with the lowest variance. This way, once we have followed this rule, we will always have:

$$S_1^2 > S_2^2$$

1. NULL HYOPTHESIS
   Is always the same and, as said before, it consists of testing whether the two variences are the same or not. Beacuse of the special structure of this test, the correct way to specify this hypothesis is:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

2. ALTERNATIVE HYPOTHESIS
   As usual, it represents what is true when the Null Hypothesis is false. I this specific case, there are only two possible specifications for this hypothesis (once more, this is so because of the special structure of this test)

	INFORMATION REGARDING σ_1^2 AND σ_2^2	TEST TYPE
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	General Case. We have no information on σ_1^2 nor about σ_2^2. Hence, if they are not equal, the only thing we can say is that they are different	Two Tails Test
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$	We have some information about σ_1^2 and σ_2^2 indicating that if they are not equal then one of them is greater. Because of the “denomination“ in Step 0, the greater will always be σ_1^2	Right-Tail Test

3. TEST STATISTIC

   In this case, the Test Statistic (TC) to use is:

$$textrmTE = \frac{S_1^2}{S_2^2}$$

The observed value of the Test Statistic (OVTS) is easily obtained replacing the corresponding sample variances in the formula, where

S_1^2 Sample Variance of Sample 1
S_2^2 Sample Variance of Sample 2

Notice that, because of Step 0 whe have that $S_1^2 > S_2^2$, and hence we will always find that VOEC> 1

4. DISTRIBUTION OF THE TEST STATISTIC when the Null Hypothesis is true
In this case, the Test Statistic follows a distribution that is known as a F of Snedecor. This distribution, the same as the $t - student$ or the χ^2 is also characterized by its "degrees of freedom". Unlike those cases, though, the $F - snedecor$ has a "pair" of degrees of freedom, those corresponding to the numerator and those corresponding to the denominator. Hence, the notation:

$$\frac{S_1^2}{S_2^2} \sim F_{(n_1-1,n_2-1)}$$

indicates that the Test Statistic $\frac{S_1^2}{S_2^2}$ follows a $F - snedecor$ distribution with $n_1 - 1$ degrees of freedom in the numerator (that is, the size of the sample that corresponds to S_1^2 in the numerator minus 1) and $n_2 - 1$ degrees of freedom in the denominator (that is, the size of the sample that corresponds to S_2^2 in the denominator minus 1).
Remember that it is very important to keep the order stablished in Step 0, that is, sample 1 corresponds to to the sample that has the highest sample variance. In this sense, the "degrees of freedom in the numerator" is the size of such sample minus 1: $n_1 - 1$. This is important whem looking at the tables of the $F - snedecor$ in order to determine the Rejection Area.

5. REJECTION AREA of size α
The Rejection Area depends on whether the test has one or two tails, as given by the Alternative Hypothesis. In this special test, the tail that "matters" will always be the Right-Tail, even if the test is a "Two Tails Test".

   (a) TWO TAILS TEST. Corresponds to thec case when we have an Alternative Hypothesis of the type $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

   For the limit values of the Rejection Area, $F_{1-\frac{\alpha}{2}}$ and $F_{\frac{\alpha}{2}}$, we only need to find $F_{1-\frac{\alpha}{2}}$ in the tables of the F with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator. The other value, $F_{\frac{\alpha}{2}}$, is not needed in any case since the OVTS is always > 1, Hence, if it falls into the Rejection Area, it will be on the Right-Tail. BECAUSE OF WHAT IS DONE IN STEP 0 (THE "DENOMINATION"OF THE SAMPLES), THE OBSERVED VALUE OF THE TEST STATISTIC WILL NEVER BE IN THE LEFT-TAIL. (See Figure 3.13)

   (b) RIGHT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$

   The limit value of the Rejection Area, $F_{1-\alpha}$, can be foundin the tables of a F with $n_1 - 1$ degrees of freedom in the numerator and $n_2 - 1$ degrees of freedom in the denominator. (See Figure 3.14)

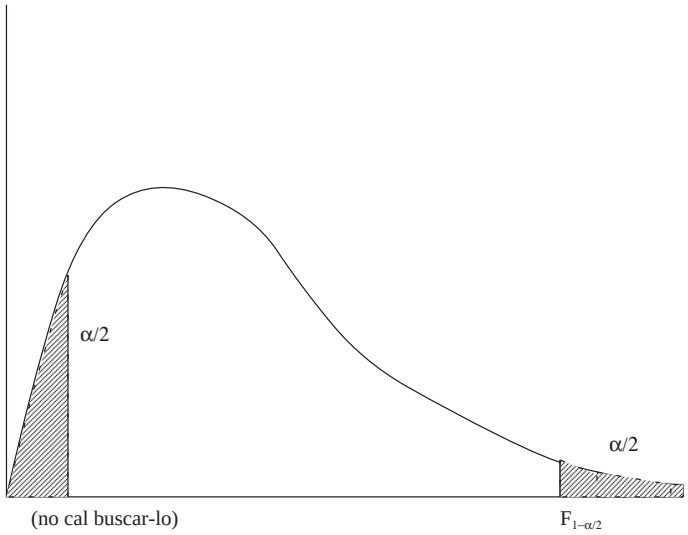

Figure 3.13: Rejection Area in a Two Tails Test

6. TEST CONCLUSION
   Finally, we have to check if the OBSEVED VALUE OF THE TEST STATISTIC (OVTS) falls, or not, inside the REJECTION AREA. If it does, we then say that the test rejects the NULL HYPOTHESIS. If it does not belong to the rejection area, then we say that the test DOES NOT REJECT THE NULL HYPOTHESIS.

3.3.3 Test for the Difference of Proportions

We now test what is the difference between the proportion of elements that have a given characterisic in two populations. For instance, we can test if the proporion of voters of the PP in Cerdanyola equals the proportion of voters of the PP in Sant Cugat ($\pi_1 - \pi_2 = 0$). Another example would be to test if the proportion of people that recovers from a given illness is bigger if they take a specific medicine than if they don't (in order to test the goodness of such medicine)

Te six steps for this test are as follows:

1. NULL HYPOTHESIS
   It is the value for the difference between the population proportions that we want to test. ($\delta_0 =$ difference to test)

$$H_0 : \pi_1 - \pi_2 = \delta_0$$

2. ALTERNATIVE HYPOTHESIS
   Represents what is true when the Null Hypothesis is false. As usual, its specification depends on the information we have about the populations

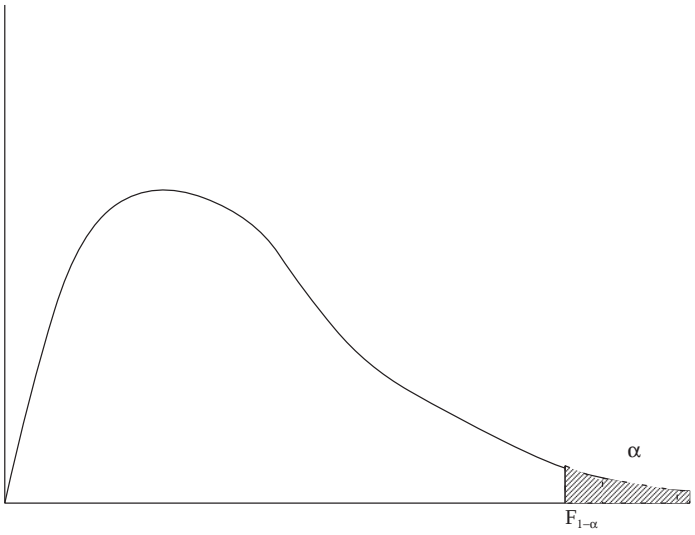

Figure 3.14: Rejection Area in a Right-Tail Test

	INFORMATION REGARDING π_1 I π_2	TEST TYPE
$H_1 : \pi_1 - \pi_2 \neq \delta_0$	General Case. We have no information about the population proportions. Hence, we can only say that if the difference is not δ_0 then it is different	Two Tails Test
$H_1 : \pi_1 - \pi_2 > \delta_0$	We have some information indicating that if the difference is not δ_0 then it must be bigger	Right-Tail Test
$H_1 : \pi_1 - \pi_2 < \delta_0$	We have some information indicating that if the difference is not δ_0 then it must be smaller	Left-Tail Test
$H_1 : \pi_1 - \pi_2 = \delta_1$	We have very specific information about the proportions indicating that if the difference is not equal to δ_0 then it must be equal to a specific alternative value δ_1	Right-Tail Test if $\delta_1 > \delta_0$ and Left-Tail Test if $\delta_1 < \delta_0$

3. TEST STATISTIC
   The Test Statistic (TE) to use in this case is always the same.

$$\text{T.S.} = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \delta_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}}$$

   The Observed Value of the Test Statitic (OVTS) is obtained by replacing in the formula the corresponding values, where

$\hat{\pi}_1$ and $\hat{\pi}_2$	Sample Proportions
δ_0	Null Hypothesis Value
$\hat{\pi}$	Common Sample Proportion
n_1 i n_2	Sample Sizes

The value of the Common Sample Proportion, $\hat{\pi}$, is obtained from

$$\hat{\pi} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2}$$

which is equivalent to computing the proportion of elements in the two samples (jointly) that have the characteristic that is of interest.

4. DISTRIBUTION OF THE TEST STATISTIC when the Null Hypothesis is true
As seen in other cases, when is true that $\pi_1 - \pi_2 = \delta_0$ then

$$\frac{(\hat{\pi}_1 - \hat{\pi}_2) - \delta_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}} \sim N(0,1)$$

5. REJECTION AREA of size α
The Rejection Area will be different depending on the type of test.

   (a) TWO TAILS TEST. Corresponds to the case when we have an Alternative Hypothesis of the type $H_1 : \pi_1 - \pi_2 \neq \delta_0$

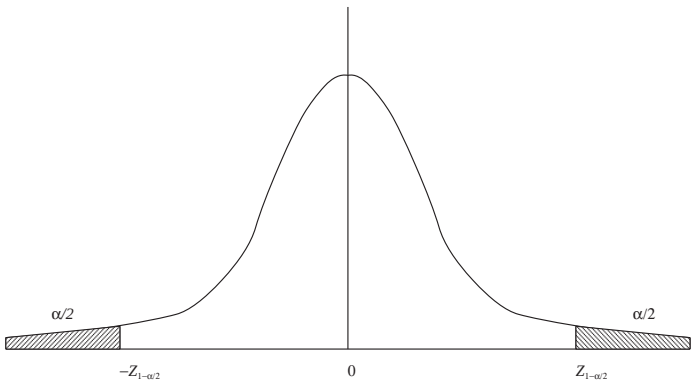

Figure 3.15: Rejection Area in a Two Tails Test

   The limit values of the Rejection Area, $Z_{1-\frac{\alpha}{2}}$ and $-Z_{1-\frac{\alpha}{2}}$, can be found in the table of a $N(0,1)$. (See Figure 3.15)

   (b) RIGHT-TAIL TEST. Corresponds to the case when we have an Alternative Hypothesis of the type $H_1 : \pi_1 - \pi_2 > \delta_0$ (or $H_1 : \pi_1 - \pi_2 = \delta_1$ and $\delta_1 > \delta_0$)

   The limit value of the Rejection Area, $Z_{1-\alpha}$, can be found in the table of a $N(0,1)$. (See Figure 3.16)

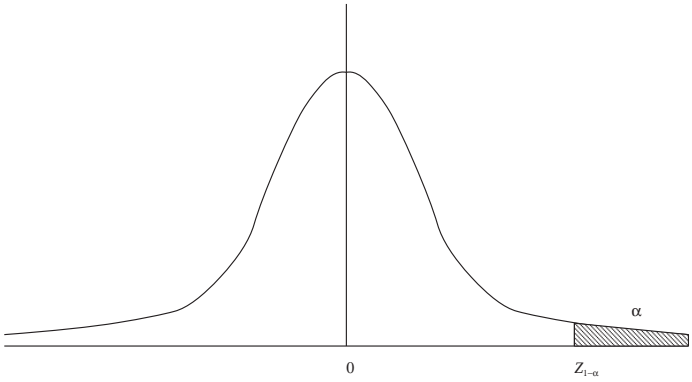

Figure 3.16: Rejection Area in a Right-Tail Test

(c) LEFT-TAIL TEST. Corresponds to the case when the Alternative Hypothesis is of the type $H_1 : \pi_1 - \pi_2 < \delta_0$ (or $H_1 : \pi_1 - \pi_2 = \delta_1$ and $\delta_1 < \delta_0$)

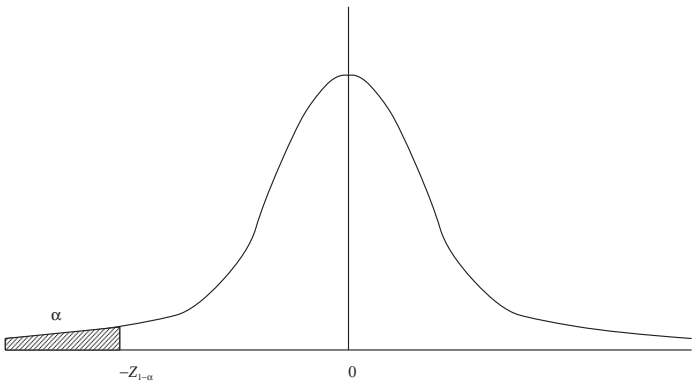

Figure 3.17: Rejection Area in a Left-Tail Test

The limit value of the Rejection Area, $Z_{1-\alpha}$, can be found in the table of a $N(0,1)$. (See Figure 3.17)

6. TEST CONCLUSION
   Finally, we have to check if the OBSEVED VALUE OF THE TEST STATISTIC (OVTS) falls, or not, inside the REJECTION AREA. If it does, we then say that the test rejects the NULL HYPOTHESIS. If it does not belong to the rejection area, then we say that the test DOES NOT REJECT THE NULL HYPOTHESIS.

3.4 Analysis of Variance

The ANalysis Of VAriance (ANOVA) between groups is a statistical technique that allows to simultaneously compare more than two populations. For instance we can compare the productivity of different types of wheat, the performance of several makes of cars, etc. For each of these cases we focus on one specific numerical feature: the weight

of wheat, the gas consumption of cars. What we test with the ANOVA is whether there exists a relationship among the averages of the different populations: have all the varieties of wheat the same weight ? do all the car makes have the same consumption ?

3.4.1 Basic Framework

What we test now is if the means of all the populations are the same. Let k be the number of populations. We will assume that each of the populations $(i = 1, \ldots, k)$ is distributed according to a Normal distribution with the same variance σ^2:

$$x_1 \sim N(\mu_1, \sigma^2) \qquad x_2 \sim N(\mu_2, \sigma^2) \quad \cdots \quad x_k \sim N(\mu_k, \sigma^2)$$

From each of the populations i a sample of size n_i is obtained.
Notation
x_{ij} Observarion j^{th} from Sample i, $(i = 1, \ldots, k; j = 1, \ldots, n_i)$.
N Número total d'observacions

$$N = \sum_{i=1}^{k} n_i$$

$\bar{X}_i$ Mitjana mostral de la mostra de la població i.

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$\bar{\bar{X}}$ Mitjana total o mitjana de totes les observacions

$$\bar{\bar{X}} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{k}(\bar{X}_1 + \cdots + \bar{X}_k)$$

3.4.2 Estadístics

\medskip

- **Variació entre mostres: VEM = SCE** (variació explicada)

$$VEM = \sum_{i=1}^{k} n_i (x_i - \bar{\bar{X}})^2$$

$$\frac{VEM}{\sigma^2} = \chi^2_{k-1}$$

- **Variació dins les mostres: VDM = SCR** (variació no explicada o resudual)

$$VDM = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)^2$$

$$\frac{VDM}{\sigma^2} = \chi^2_{N-k}$$

- **Variació total: VT = STC=VEM+VDM**

$$VT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{X}})^2$$

$$\frac{VT}{\sigma^2} = \chi^2_{N-1}$$

3.4.3 Contrast

El test a realitzar és de la forma

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1: \quad \text{No totes les } \mu_i \text{ soniguals}$$

Cal tenir present que: \medskip

1. Sempre, per qualsevol i, $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2$ és un estimador *inesbiaixat* de σ^2
2. Sota la hipòtesi nul·la, $S_E^2 = \frac{SCE}{k-1}$ és un estimador *inesbiaixat* de σ^2
3. En considerar totes les variàncies com iguals, $S_R^2 = \frac{SCR}{N-k}$ és un estimador *inesbiaixat* de σ^2

Per tant:

$$\frac{\frac{SCE}{\sigma^2(k-1)}}{\frac{SCR}{\sigma^2(N-k)}} \sim \frac{\frac{\chi^2_{k-1}}{k-1}}{\frac{\chi^2_{N-k}}{N-k}}$$

És a dir, l'estadístic de contrast és

$$F^* = \frac{S_E^2}{S_R^2} \sim F(k-1, N-k)$$

3.5 Non-Parametric Tests

In the previous sections we have seen the main tests of the so called "parametric tests", that is, we test hypothesis regarding one specific "parameter" of the population (or comparing the parameters of two populations).

In this section we will see one specific test of the kind named "non-parametric tests", that is, we do not perform tests regarding a specific parameter but we test more general hypothesis. More specifically, we will see how to test if the date in our sample seems to come from a given theoretical distribution.

3.5.1 The Kolmogorov-Smirnov Test for the Goodness of Fit

This test checks whether a given set of data (the sample) seems to "fit" (and how "good" is this "fit") a specific probability distribution. For instance, whe can test whether the distribution of the income per capita in a sample collected in Cerdanyola seem to fit what would be a Normal distributon with the same mean and variance (this is sometimes referred to as the "normality test"). The idea is to test if the "frequencies" observed in the sample coincide with the "frequencies" (probabilities) that we can compute using a Normal distribution with the same mean and variance.

Hence, the procedure focuses on looking at the differences between the "observed frequency" (in the sample) and the "theoretical frequency" (according to a Normal distribution) to determine if these differences are small enough as to conclude that, indeed, the data in the sample seems to follow a distribution close to that of a Normal.

The procedure is as follows when we want to test if the data "fits" a Normal distribution with mean$=\mu$ and variance$=\sigma^2$, that is, $N(\mu, \sigma^2)$

1. The Null Hypothesis for this test is always the same:

   $$H_0 : F_O = F_T$$

   Where F_O is the "observed cummulative frequency" in the sample and F_T is the "theoretical cummulative probability (frequency)" according to a Normal distribution[2]. How these "frequencies" are computed would be explained later.

2. The Alternative Hypothesis also is always the same:

   $$H_A : F_0 \neq F_T$$

   That is, if the "frequencies" are not equal, then they are just different.

3. Test Statistic

   For this test, the computation of the Test Statistic is rather involved and takes a lot of work.

   First, for the "observed frequencies" F_O, we must compute for each element in the sample what is the proportion (or frequency) of elements that are "smaller or equal" to that value

   $$F_O(x_i) = \frac{\text{Number of elements in the sample smaller or equal than } x_i}{\text{Total number of elemnts in the sample}}$$

   Now we must compute (using the $N(0,1)$ tables) what are the corresponding "theoretical frequencies" according to the "Normal" $N(\mu, \sigma^2)$ we are testing for:

   $$F_T(x_i) = P(X \leq x_i) = P(\frac{X-\mu}{\sqrt{\sigma^2}} \leq \frac{x_i - \mu}{\sqrt{\sigma^2}}) = P(Z \leq \frac{x_i - \mu}{\sqrt{\sigma^2}})$$

   where $Z \sim N(0,1)$

   Finally, we compute the differences between each of the "observed frequencies" $F_O(x_i)$ and the corresponding "theoretical frequencies" $F_T(x_i)$ and then select

[2]The test could also be done to check if the data behaves according to another distribution, like an exponentical, a Poisson, a Binomial, etc. Here we focus only on the "normality test", that is, to check if the data in the sample behaves according to a Normal distribution.

the "maximum" (in absolute value) of these differences. This value will be the Observed value of the Test Statistic. That is, the Test Statistic for this test, that we denote by $K - S$ is given by:

$$K - S = \max |F_O(x_i) - F_T(x_i)|$$

and the corresponding Observed Value of the Test Statistic follows from the computation of the differences and the selection of the "maximum" difference as explained above.

4. Distribution of the Test Statistic when the Null Hypothesis is true

   For this test, the distribution to use is a special one named the Kolmogorov-Smirnov distribution, whose values are also in tables

5. Rejection Area of size α

   This test is always a Right-Tail Test (only a tail on the right). In the Kolmogorov-Smirnov tables we find the limit value for this rejection area depending on both the size of the rejectin area α and the size of the sample n

6. Test conlcusion

   Given the special features of this test, we only need to check if the Observed Value of the Test Statistic K-S is bigger or not that the value found in the Kolmogorov-Smirnov tables. If it is bigger, then we reject the Null Hypothesis that says that the sample follows the distribution of a Normal. If it is smaller then we do not reject that hypothesis

3.6 Exercises

1. Among the following sentences, which are true and which are false:

   (a) The larger the significance level, the more likely is to reject $H_0$when it is true.

   (b) The larger the confidence level, the more likely is to reject $H_0$when it is true.

   (c) The larger the significance level, the higher the power of the test.

   (d) The higher the power of the test, the more likely is to reject H_0 when it is false.

2. In the Penedés area, the average grape crop in a normal year is of 100 Tons/Ha. This year that the weather has been specially good 12 selected lots produced 106 Tons/Ha. in average. If the crop per Ha. is a random variable with variance 64, is there any reason to think that this year's crop is better than normal ? ($\alpha = 0.01$). Find the p_value in this case.

3. A manager orders a large quantity of steel girders with an average length of 5 meters. It is known that the length of such girders is a random variable *normally* distributed with 0.02 *standard deviation*. Once the order is received, the manager randomly selects 16 girders and measures their lengths. If the average length in the sample is less than expected, the manager will return the order.

(a) If the probability of rejecting a "good" order is 0.04, what has to be the value of the average length in the sample that makes the manager return the order ?

4. A specific task in a factory takes 5 minutes in average to be completed. The factory manager believes that one of the workers spends more time in this operation. The manager selects a sample of 11 timings for this worker and collects the following data (in minutes): 4.8, 5.6, 5.3, 5.2, 4.9, 4.7, 5.7, 4.9, 5.7, 4.9, 4.6. Assuming that operation time is a Normal random variable,

   (a) Does the data supports the manager's belief ($\alpha = 0.02$).

   (b) How much is the p_value in this case ?

5. A washing machines producer claims that only a 5 % of the whole production need service whithin the first year of normal operation. A consumers organization asks 20 families with the same number of members that have bought this washers to report about any malfunctioning in the first year. At the end, only 3 families reported some kind of problem.
   Test whether the manufacturer's hypothesis that the proportion of "bad" units is 0.05 can be rejected against the consumers organization belief that such proportion is more than 0.05 with $\alpha = 0.1$

6. The manager of the election campaign of candidate A believes that his candidate is in the same position as his opponent, candidate B. Nevertheles, hi is afraid that some recent scandals might have harmed his candidate. Hence, he decides to interview 1500 citizens and 720 show a clear preference for candidate A. Does it exist any reason to think that the scandal has affected the image of candidate A ? ($\alpha = 0.05$)

7. The person in charge of a workshop thinks that the number of items that a particular worker produces oscillates more than normal. He decides to monitor the worker activity during 10 randomly selected days. The number of items produced each of these days was 15, 12, 8, 13, 12, 15, 16, 9, 8, and 14. It is known that the standard deviation of other workers in the workshop is of 2 units, and that the number of produced items per day is distributed according to a Normal. Does this data support the manager's suspicion? ($\alpha = 0.05$). What is the p_value in this case ?

8. A manufacturer wants to compare the average stress of the linens he produces against that of his competitors. One hundred threads of each brand were selected and their corresponding stress recorded. The results were:

$$\bar{X}_1 = 110.8 \qquad \bar{X}_2 = 108.2$$

$$s_1 = 10.2 \qquad s_2 = 12.4$$

   Assuming that the sampling took place on two normal, independent populations, is there any reason to think that the difference between the average stress of the two brands is significant ? ($\alpha = 0.02$)

9. A survey was conducted to test the degree of influence of alcohol on the ability to concentrate to perform a specific task. Ten people were selected at random to participate in an experiment. First, each person developed the task without any alcohol intake, and then did it again with a 0.1 % of alcohol in blood. The task completion timings recorded before and after the alcohol intake were:

Participant	Before	After
1	28	39
2	22	45
3	55	67
4	45	61
5	32	46
6	35	58
7	40	51
8	25	34
9	37	48
10	20	30

Can we conclude, with a significance level of 5%, that the average timing "before" is lower that the average timing "after" in more than ten minutes ? (assume that he population is normally distributed)

10. An investor wants to compare the risks associated to two different stock markets, A and B. Market risk is measured using the variance of the daily changes in stock prices. The investor believes that the risk in market A is lower than the risk in market A. Two random samples are selected, consisting of 21 observations on the changes i prices in market A and 16 observations on the changes of prices in market B. The results are:

Market A	Market B
$\bar{X}_A = 0.3$	$\bar{X}_B = 0.4$
$s_A = 0.25$	$s_B = 0.45$

Assuming that both samples come from two Normal and independent populations, does the data suport the investor's belief ? ($\alpha = 0.5$)

11. An electrician buys large amounts of electrical components mainly from two suppliers, A and B. Because of a better pricing policy, the electrician will buy only from supplier B is the proportion of faulty items is the same for both suppliers. The electrician selects two random samples, one of size 125 from supplier A and other of size 100 from supplier, discovering that there are exactly 7 faulty components in each sample. Is there any reason not to buy only from supplier B ? ($\alpha = 0.02$).

12. Two people play "heads or tails" with a coin. After 100 tosses A, who chose "heads", won 62 times. Immediately, B claims that the coin is biassed and the probability of getting heads is above 50 %. Is she right ? ($\alpha = 0.05$).

13. In a Hospital 7 patients were selected, observing that they slept 7, 5, 8, 8.5, 6, 7 i 8 hours respectively. All of them were given a new sleeping pill, and then 5 of them were selected, observing 9, 8.5, 9.5, 10 i 8 sleep hours respectively. Is the new pill effective ? (Assume normality and $\alpha = 0.05$)

14. Random errors in two measuring tools follow Normal distributions $N(0, \sigma_1^2)$ i $N(0, \sigma_2^2)$. In a sample of size 7 the following measuring error are observed

    First tool: 0.3, 0.7, -1.1, 2.0, 1.7, -0.8, -0.5

    Second tool: 1.6, -0.9, -2.8, 3.1, 4.2, -1.0, 2.1

    Can we tell that the first tool is more precise than the second tool ?

15. A testing lab is asked to compare the durability of four different brands of golf balls. The lab randomly selects 7 balls from each brand and puts them into a machine that hits them with constant strength. The measurement of interest is the number times the machine hits the ball before its external cover is broken. The following table reports the data gathered during the test:

A	B	C	D
205	242	237	212
229	253	259	244
238	226	265	229
214	219	229	272
242	251	218	255
225	212	262	233
209	224	242	224
204	247	234	245

    Is there any reason to think that the average durability is different across brands ? ($\alpha = 0.05$).

16. In order to test if there exist differences in the average crop of three varieties of corn, a lot is divided in three equal areas and one different variety of corn is planted in each one. In each area a sample of size 5 is collected corresponding to 5 measurements of tons per acre. The following table is an incomplete ANOVA table for this problem

Variation	Sum	Deg. of Freedom	Average Sum	F
VEM	64			
VDM				
VT	100			

    Complete the ANOVA table and determine if the null hypothesis of all the averages being equal can be rejected with $\alpha = 0.01$

Chapter 4

Goodness of Fit and Correlation Analysis

In the previous chapter we have seen the main tests of the so called "parametric tests" (we test hypothesis regarding one specific "parameter" of the population).

In this chapter we will first see one specific test of the kind named "non-parametric tests", that is, we do not perform tests regarding a specific parameter but we test more general hypothesis. More specifically, we will see how to test if the date in our sample seems to come from a given theoretical distribution.

Then, we will introduce the concept of "relationship" between data in two samples. This will be important for the next chapter. More specifically, we will introduce the analysis of the correlation between samples.

4.1 The Kolmogorov-Smirnov Test for the Goodness of Fit

This test checks whether a given set of data (the sample) seems to "fit" (and how "good" is this "fit") a specific probability distribution. For instance, we can test whether the distribution of the income per capita in a sample collected in Cerdanyola seem to fit what would be a Normal distribution with the same mean and variance (this is sometimes referred to as the "normality test"). The idea is to test if the "frequencies" observed in the sample coincide with the "frequencies" (probabilities) that we can compute using a Normal distribution with the same mean and variance.

Hence, the procedure focuses on looking at the differences between the "observed frequency" (in the sample) and the "theoretical frequency" (according to a Normal distribution) to determine if these differences are small enough as to conclude that, indeed, the data in the sample seems to follow a distribution close to that of a Normal.

The procedure is as follows when we want to test if the data "fits" a Normal distribution with mean$= \mu$ and variance$= \sigma^2$, that is, $N(\mu, \sigma^2)$

1. The Null Hypothesis for this test is always the same:

$$H_0 : F_O = F_T$$

Where F_O is the "observed cumulative frequency" in the sample and F_T is the "theoretical cumulative probability (frequency)" according to a Normal distribution[1]. How these "frequencies" are computed would be explained later.

2. The Alternative Hypothesis also is always the same:

$$H_A : F_0 \neq F_T$$

That is, if the "frequencies" are not equal, then they are just different.

3. Test Statistic

For this test, the computation of the Test Statistic is rather involved and takes a lot of work.

First, for the "observed frequencies" F_O, we must compute for each element in the sample what is the proportion (or frequency) of elements that are "smaller or equal" to that value

$$F_O(x_i) = \frac{\text{Number of elements in the sample smaller or equal than } x_i}{\text{Total number of elements in the sample}}$$

Now we must compute (using the $N(0,1)$ tables) what are the corresponding "theoretical frequencies" according to the "Normal" $N(\mu, \sigma^2)$ we are testing for:

$$F_T(x_i) = P(X \leq x_i) = P(\frac{X-\mu}{\sqrt{\sigma^2}} \leq \frac{x_i - \mu}{\sqrt{\sigma^2}}) = P(Z \leq \frac{x_i - \mu}{\sqrt{\sigma^2}})$$

where $Z \sim N(0,1)$

Finally, we compute the differences between each of the "observed frequencies" $F_O(x_i)$ and the corresponding "theoretical frequencies" $F_T(x_i)$ and then select the "maximum" (in absolute value) of these differences. This value will be the Observed value of the Test Statistic. That is, the Test Statistic for this test, that we denote by $K-S$ is given by:

$$K - S = \max |F_O(x_i) - F_T(x_i)|$$

and the corresponding Observed Value of the Test Statistic follows from the computation of the differences and the selection of the "maximum" difference as explained above.

4. Distribution of the Test Statistic when the Null Hypothesis is true

For this test, the distribution to use is a special one named the Kolmogorov-Smirnov distribution, whose values are also in tables

5. Rejection Area of size α

This test is always a Right-Tail Test (only a tail on the right). In the Kolmogorov-Smirnov tables we find the limit value for this rejection area depending on both the size of the rejecting area α and the size of the sample n

[1] The test could also be done to check if the data behaves according to another distribution, like an exponential, a Poisson, a Binomial, etc. Here we focus only on the "normality test", that is, to check if the data in the sample behaves according to a Normal distribution.

6. Test conclusion

   Given the special features of this test, we only need to check if the Observed Value of the Test Statistic K-S is bigger or not that the value found in the Kolmogorov-Smirnov tables. If it is bigger, then we reject the Null Hypothesis that says that the sample follows the distribution of a Normal. If it is smaller then we do not reject that hypothesis

4.2 Relationship between samples

Consider two independent samples randomly obtained from two different populations. For instance, we could think of one sample with data about the unemployment rate in Cerdanyola and another sample with data about income per capita also in Cerdanyola. Then, we might wonder if there is a "relationship" between these data, that is, if it seems to be true that when the unemployment rate is low then the income per capita is high and vice versa. These kind of questions is more ambitious for the economic analysis than those addressed in previous chapters. Indeed, for the design of economic policies it is very important to know what kind of relation exists among the different economic variables.

In this sense, there are two types of relationships that we can observe between two given variables:

1. Casual

   We say that two variables have a "casual" relationship when changes in one of the variables induce changes in the other one. For instance, it seems clear that the lower is the interest rate the higher is the demand for mortgage loans.

2. Spurious

   We say that two variables have a "spurious" relationship when they seem to be related but this relation is not causal but explined from some other factor, like a third variable that is independently related to each of these two or some other unknown factor.

Once we now if two variables are related or not, it is very important to understand what is the kind of relationship they have. Indeed, even if two variables are related to each other we can not "use" this relationship trying to influence one of the variables by means of changes in the other.

4.3 Correlation Analysis: The Correlation Coefficient

The analysis start with a set of paired data sampled from two variables X and Y

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

We can represent these two pairs of data in a $X - Y$ graph to obtain, generically, one of these four kind of graphs, named *Data dispersion diagram*

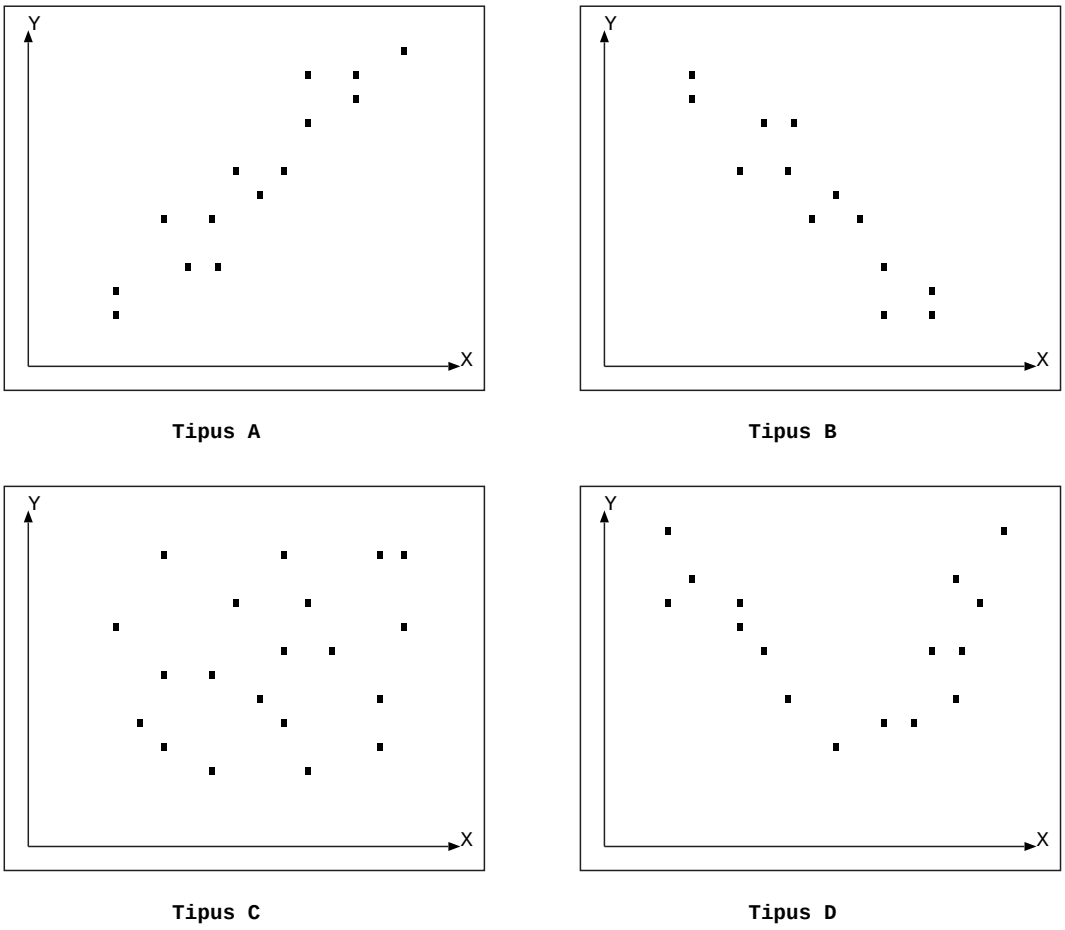

Figure 4.1: Data dispersion types

Each of these four types of data dispersion corresponds to a specific kind of relationship between the variables X i Y

Dispersion Type	Type of relationship between variables	Comment
A	Monotone increasing	The two variables change in the same direction. The higher is the value of one of them, the higher is the value of the other
B	Monotone decreasing	The two variables change in opposite directions. The higher is the value of one of them, the lowest is the value of the other
C	No relationship	There is no apparent relationship between the variables. For some data high values of X correspond to high values of Y, but for some other data they correspond to low values
D	Non-monotone relationship	The two variables seem to be related to each other, but this relationship is sometimes increasing, sometimes decreasing

Figure 4.2: Type of relationship between variables

With the Correlation Analysis we seek to determine:

1. Which is the type of relationship between the variables
2. Which is the "degree" of relationship between the variables

This analysis is done by means of the *Correlation Coefficient* r given by the formula

$$r = \frac{\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i}{n S_X S_Y}$$

where:

$$\tilde{x}_i = x_i - \bar{X}$$
$$\tilde{y}_i = y_i - \bar{Y}$$
$$S_X = \sqrt{\frac{\sum_{i=1}^{n} \tilde{x}_i^2}{n}}$$
$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} \tilde{y}_i^2}{n}}$$

It can be proved that $-1 \leq r \leq 1$. The interpretation of this coefficient is as follows

The correlation coefficient r is an estimator (that is computed using the sample of observations of the variables X and Y) of the population correlation ρ that measures the true correlation between the two variables. In this sense, as we have done in previous chapters with other sample estimators like $\bar{X}$, we can use r to do "inference" regarding

Value of r	Interpretation
$-1 \leq r < 0$	There exists a monotone decreasing relationship (Type B). The closer to -1 is r the stronger the relationship
$0 < r \leq 1$	There exists a monotone increasing relationship (Type A). The closer to 1 is r, the stronger the relationship
$r \approx 0$	When r is close to 0, we do not have any kind of monotone relationship. The problem, though, is that it is not possible to determine if we are in a relationship like in Type D or we do not have any kind of relation like in Type C

Figure 4.3: Interpretation of r

ρ (confidence intervals, hypothesis testing). To do so we must know the distribution of such estimator. It can be proved that:

$$\frac{1}{2}\ln(\frac{1+r}{1-r}) \sim N(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$$

4.4 Exercises

1. Given the three sets of data below regarding variables X and Y, plot the dispersion diagram and compute the correlation coefficient r. Comment on the type of relation in each case.

   SET A

X	Y
1	5
2	9

   SET B

X	Y
1	1
2	9
3	2
4	7
5	6

   SET C

X	Y
1	1
2	16
3	81
4	256
5	625

2. The correlation coefficient computed in a sample of size 39 is $r = 0.35$. Find the 95% confidence interval for the true correlation ρ. Does the confidence interval found imply that the null hypothesis $\rho = 0$ can not be rejected ?

3. The correlation coefficient computed in a sample of size 28 is $r = 0,8$. Test the null hypothesis $\rho = 0,8$.

4. The following data correspond to the class attendance and the final grades in an Statistics II test.

	Pass	Fail	Total
Attended the class regularly	40	20	60
Did not regularly the class	15	25	40
Total	55	45	100

   Does this data set indicate that the class attendance is related to the final grade ?

5. A compmay wants to maximize the number of people that answers to their surveys. The tested three different methods of presenting the survey with a random sample of size 2000 and the results were:

Format of the Survey	Did answer	Did not answer	Total
Typewriter	250	200	450
Cyclist	300	450	750
Computer Laser Printout	300	500	800
Total	850	1150	2000

   Does the format of the survey influence the people's attitude to take the survey ?

6. The number of births per month in a hospital during a given year were:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Set	Oct	Nov	Dec
95	105	95	105	90	95	105	110	105	100	95	100

   If $\alpha = 0.01$, is there any reason to think that the number of births is not distributed uniformly during the year ?