Chapter 1

Descriptive Statistics

Descriptive statistics deals with organizing, summarizing, classifying, and representing data obtained in a statistical study, with the aim of getting a global idea of the data, discovering possible relationships between them, identifying which values are similar, which differ greatly from the rest, and so on. In many cases, it should be a preliminary step before making inferences. We can carry out the summary of the information obtained in the sample in two ways: numerically and graphically.

1.1 Frequency Tables

The investigation of a group always provides a more or less extensive set of data, which we aim to analyze to understand the nature of the data in relation to the characteristics we want to study. We will numerically summarize the data using frequency tables. Before that, let us define some concepts that we will need.

Characteristic

The observation of an individual translates into the description of some inherent quality of the individual. This quality is called a *characteristic*. The characteristic can be qualitative (non-numeric), such as eye color, marital status, etc., or quantitative (numeric), such as height, age, weight, etc.

Modality

These are the variations of a characteristic. For example, for eye color, we can have different modalities: green, blue, brown, etc. When we work with quantitative characteristics, the modalities can be grouped into intervals.

Statistical Variable

It is the set of values resulting from the observation of a certain characteristic on a group of individuals. Obviously, it can be qualitative or quantitative. A quantitative variable can be classified as *discrete*, when there are values between which it is not possible to find other values, for example, when we study the number of children (we cannot observe 2.5 children in a family), or

continuous, when it is possible to find values, for example, height or cranial perimeter.

1.1.1 Frequencies

Now we will define the different types of frequencies associated with a set of data.

Absolute frequency

It is the number of observations that present a certain modality, x_i. It is written as n_i. The sum of all absolute frequencies is equal to the total number of observations.

Relative frequency

It is denoted as f_i, and it is the quotient between the absolute frequency and the total number of observations. The sum of all relative frequencies is always equal to one.

Cumulative frequencies

They are calculated for quantitative variables. We have two types:

- Cumulative absolute frequency: $N_i = n_1 + n_2 + \cdots + n_i$
- Cumulative relative frequency: $F_i = f_1 + f_2 + \cdots + f_i$

1.1.2 Data sorting in a table

When we have the data of a discrete numerical variable, we first order the modalities x_i from lowest to highest, and then calculate the frequencies, representing them as follows:

frequencies table				
	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_m	$N_m = n$	f_m	$F_m = 1$
Total	$\sum_{i=1}^{m} n_i = n$		$\sum_{i=1}^{m} f_i = 1$	

If we work with a large amount of data, they can be grouped into intervals (called *class intervals*), which we will choose conveniently to lose the minimum amount of information possible.

The number of intervals is usually between 4 and 15 (it is often the value closest to $\sqrt{n}$), so that there are at least 5 observations in each interval. Of course, the intervals cannot overlap. To tabulate this data, we will change the first column of the previous table to a column indicating the class intervals.

1.2 Statistical Parameters

Statistical parameters are representative values of a set of data, in the sense that they condense much of the information that these data provide us. This allows us to capture the structure of the set of observations and the distribution from which the data in the sample come in a simple and fast way. We classify them into:

1. Measures of central tendency and position.
2. Measures of dispersion.
3. Measures of shape.

1.2.1 Measures of central tendency and position

Mean

Undoubtedly, it is the most well-known parameter of those intended to provide a central value. We define it as follows:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{m} n_i x_i \tag{1.1}$$

For values grouped into intervals, x_i will be the midpoint of each interval (called the *class mark*).

Median

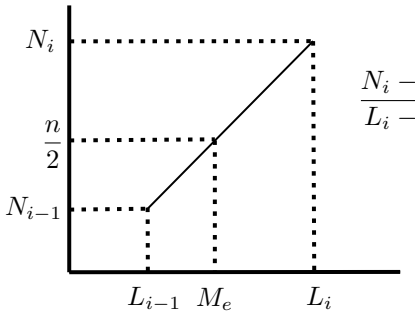
The median is a central parameter that is not linked to the numerical value of the observations but to their relative position within the set of data. To calculate it, we must first consider whether the observations are grouped into intervals or not. If we have n numerical observations $x_1, \ldots, x_n$, ordered from lowest to highest, we must distinguish between two situations:

- If n is odd, the median will be the central value of the observations.
- If n is even, the median will be the sum of the two central values divided by two.

For grouped data, we first select the first interval $[L_{i-1}, L_i)$ in which the cumulative absolute frequency exceeds $n/2$. Then, we calculate the median by simple linear interpolation in which we only have to do a rule of three and

solve for M_e (see graph), with which we arrive at the following formula:

$$M_e = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot (L_i - L_{i-1}) \tag{1.2}$$



$$\frac{N_i - N_{i-1}}{L_i - L_{i-1}} = \frac{\frac{n}{2} - N_{i-1}}{M_e - L_{i-1}}$$

Percentiles y Cuartiles

We define the k-th percentile, P_k, as the value below which $k\%$ of the observations (ordered from smallest to largest) lie. A particular case is that of the *quartiles*, which are three values, Q_1, Q_2 and Q_3, that leave behind 25%, 50% and 75% of the observations, respectively. Clearly, $M_e = Q_2 = P_{50}$.

For the calculation of a percentile of order k, we also need to consider whether we have discrete observations or observations grouped into intervals. If we have n discrete numeric observations $x_1, \ldots, x_n$, ordered from smallest to largest, we proceed as follows:

- We assign to each value x_i a value of $f = \dfrac{i-1}{n-1}$, so that the minimum value has $f = 0$ and the maximum value has $f = 1$.
- The percentile P_k is calculated by linear interpolation, and is the one that has $f = k \cdot n/100$.

Example. Let's calculate Q_1 and Q_3 for the following data: 4,8,9,16,18,22. First, we calculate the value of f for this data:

x_i	4	8	9	16	18	22
f	0	0.2	0.4	0.6	0.8	1

As the value of f for Q_1 is $f = 0.25$ (25% of the observations) and for Q_3 it is $f = 0.75$ (75% of the observations), we know that the first quartile is between 8 and 9, and the third between 16 and 18. Therefore, the proportions will be:

- $\frac{9-Q_1}{0.4-0.25} = \frac{9-8}{0.4-0.2} \implies Q_1 = 9 - \frac{(9-8)\cdot(0.4-0.25)}{(0.4-0.2)} = 8.25$
- $\frac{18-Q_3}{0.8-0.75} = \frac{18-16}{0.8-0.6} \implies Q_3 = 17 - \frac{(18-16)\cdot(0.8-0.75)}{(0.8-0.6)} = 17.5$

When the data is grouped, we operate similarly to when calculating the median, except that instead of n/2, we use $n \cdot k/100$:

$$P_k = L_{i-1} + \frac{\frac{n\cdot k}{100} - N_{i-1}}{n_i} \cdot (L_i - L_{i-1}) \tag{1.3}$$

Mode

The mode is the value (or values) with the highest absolute frequency. If the data is grouped into intervals, the mode will be the result of the following formula:

$$M_o = L_{i-1} + \frac{\delta_1}{\delta_1 + \delta_2} \cdot (L_i - L_{i-1}) \tag{1.4}$$

where $[L_{i-1}, L_i)$ is the interval with the highest absolute frequency (modal interval); $\delta_1 = n_i - n_{i-1}$ and $\delta_2 = n_i - n_{i+1}$.

1.2.2 Dispersion parameters

When analyzing a set of observations, it is often convenient to express numerically whether these values are close to each other or widely dispersed. For example, if I have three type A batteries that have lasted for 4, 5, and 6 hours, and three type B batteries that have lasted for 2, 5, and 8 hours, I can say that the average duration of both types has been 5 hours, but the variability between type A and type B batteries is very different. We, therefore, need to define statistical measures that provide us with information about the dispersion of a set of data. Let's see the most important ones.

Variance

We define the variance of a set of data as:

$$var[x] = s^2 = \frac{1}{n}\sum_{i=1}^{m} n_i(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{m} n_i x_i^2 - \overline{x}^2 \tag{1.5}$$

For values grouped in intervals, x_i will be the midpoint of each interval. The standard deviation, s, is defined as the square root of the variance.

Interquartile Range

It is the length of the central interval that contains 50% of the observations. We will see later that it is the width of the box in a Box-Plot diagram. It is defined as follows:

$$R_I = Q_3 - Q_1 \tag{1.6}$$

Coefficient of Variation

It is a measure that allows us to make comparisons between groups that are measured in different units or magnitudes, taking into account only the proportion between mean and standard deviation. It also helps us to study the variability of a set of data relative to its mean. It is defined as follows:

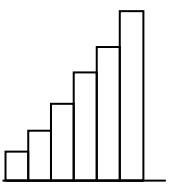
$$C_v = \frac{s}{|\overline{x}|} \tag{1.7}$$

1.2.3 Shape parameters

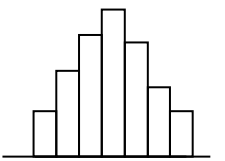
Symmetry

Symmetry parameters help determine whether the data is distributed symmetrically around a central value, which in our case will be the mean. We can have symmetric data or with skewness, which will be positive or negative. We will say that there is a positive skewness when $M_o \leq M_e \leq \overline{x}$; negative, when $\overline{x} \leq M_e \leq M_o$. When the three values coincide, the data is symmetric. To have a numerical value that refers to the symmetry of a set of data, Fisher's skewness coefficient is used:

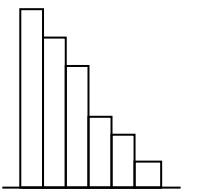
$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{m} n_i (x_i - \overline{x})^3}{s^3} \tag{1.8}$$



negative non-symmetric



Symmetric



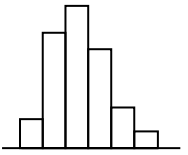
positive non-symmetric

Kurtosis

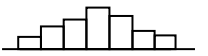
This parameter measures how pointed or flat the frequency polygon is compared to that of the standard normal distribution. We define the coefficient of kurtosis as:

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{m} n_i(x_i - \overline{x})^4}{s^4} - 3 \tag{1.9}$$

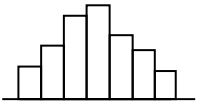
We subtract 3 from the value because the theoretical coefficient of kurtosis for a normal distribution is 3. If $g_2 < 0$, we say that the data distribution is platykurtic; if $g_2 > 0$, leptokurtic, and if $g_2 = 0$, mesokurtic.



leptokurtic distribution



platykurtic distribution



mesokurtic distribution

1.3 Graphical Representations

Graphs are very useful because they highlight and clarify trends that are not easily captured in tables, help make a first estimation of population parameters, and provide information about the distribution from which the data come.

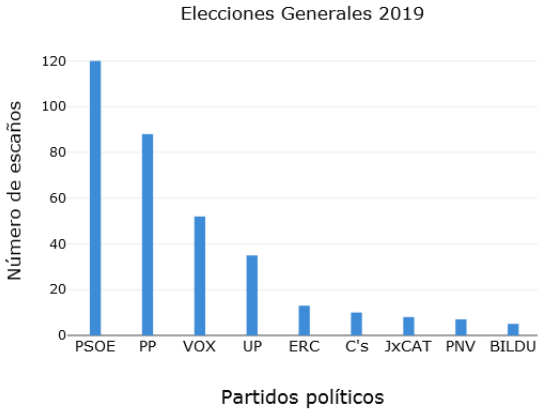
1.3.1 Bar Chart

These graphs are specific to qualitative and non-grouped quantitative variables. On the horizontal axis, the different modalities are placed, and the height of the bars depends on the value of the frequency of each modality.

Example

We will show a bar chart that represents the seats obtained by the main political parties in the General Elections held on November 10, 2019. In this case, the chart we will represent will be one of absolute frequencies. The height of each bar will be the number of seats obtained by each party. To represent these tables, we could also have used a *pie chart*, which is a circle divided into circular sectors in which the arc represented by each sector is proportional to the number of seats obtained.

PSOE	PP	VOX	UP	ERC	C's	JxCAT	PNV	BILDU
120	88	52	35	13	10	8	7	5



1.3.2 Histogram

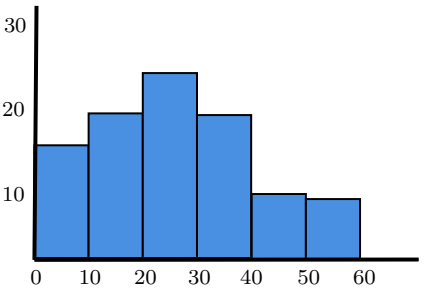
This graph consists of a series of rectangles that are used to represent frequencies of quantitative variables grouped into intervals. The base of the rectangle is equal to the class interval width, while the height is determined by ensuring that the area is proportional to the frequency. Unlike the bar chart, there is no separation between the rectangles in the histogram. Usually, this graph is accompanied by the so-called *frequency polygon*, which is formed by joining the points of greatest height of the columns in the histogram.

Example

A telephone company records the duration of calls made by 100 clients, obtaining the following data:

duration (minutes)	number of clients
[0 , 10)	16
[10 , 20)	20
[20 , 30)	25
[30 , 40)	20
[40 , 50)	10
[50 , 60)	9

We are going to represent the histogram along with the frequency polygon. In this case, the height of each rectangle will coincide with the corresponding absolute frequency:



1.3.3 Stem-and-leaf plot

In this type of graph, the observed values are represented directly instead of their frequencies. With its representation, we can observe how symmetric the data is, how dispersed it is, if there are concentrations of values, etc. The stem-and-leaf plot is used when the number of data is not excessively large, usually less than 50. To represent it, we will follow the following procedure:

- The data is arranged in two columns separated by a vertical line. On the left, the stem is placed, formed by values ordered from lowest to highest in descending order. Each stem defines a class and is written only once; the number of leaves represents the frequency of that class. For example, if we have the values 125, 126, and 127, we would represent them as 12|567. The class [120,130) is represented with the stem 12, and there are 3 values in that interval (absolute frequency 3), which form the leaves.
- For data with 2 digits, the tens will form the stem, and the units will form the leaves, while if we have 3 digits, the stem will be formed by the hundreds and tens, separated by the units, which will form the leaves. And so on, successively.
- To understand a graph of this type, it is convenient to specify how the original values have been treated, the number of significant figures with which it has been rounded, etc.

Example

We have the following data:

3, 4, 4, 5, 5, 5, 8, 8, 9, 9, 10 ,12, 13, 13, 14, 16, 16, 25, 26, 28, 31, 32, 33, 33, 35

To create the stem-and-leaf plot, we separate the tens from the units. The number 4 in the second row - in bold - represents 14 because it is on stem 1, and the number 5 in the fourth row - in bold and italic - represents 35 because

it is on stem 3.

0	3 4 4 5 5 5 8 8 9 9
1	0 1 2 3 3 **4** 6 6
2	5 6 8
3	1 2 3 3 ***5***

Example

The following numbers represent the amount of coal, in kilograms, collected from a mine in the last 10 days:

7563, 8767, 6567, 9898, 7678, 5456, 4543, 6786, 6756, 7876

Let's set the stem as the first significant digit. Therefore, the class interval will be 1000 units. We will have 1 four-thousand, 1 five-thousand, 3 six-thousand, 3 seven-thousand, 1 eight-thousand, and 1 nine-thousand.

4	5
5	4
6	5 **7** 7
7	5 6 8
8	7
9	8

The number 7 in the third row -in bold- represents the value 6756.

Note

If we were shown the diagram without having access to the data, we could approximate the values as:

4500, 5400, 6500, 6700, 6700, 7500, 7600, 7800, 8700, 9800

Example

We are shown the following stem-and-leaf plot:

0	7 0 2 3 6
2	6 9 6 6 9
4	2 3
6	2 7 8

We are also told that the decimal point is located 4 digits to the right of the stem. This indicates that we are dealing with digits of 4 zeros. Therefore, the

only thing we can deduce is that the amplitude of the class interval is 20,000 units because each part of the stem increases by 2 units. Let's look at the first stem:

Its leaves are 70236. It may seem wrong because it is usually written in ascending order. What is deduced is that the first value approximates to 7,000, the second to 10,000, the third to 12,000, the fourth to 13,000, and the fifth to 16,000. We now show the original data, so we can see it more clearly:

66814.195	42144.338	25697.767	35976.874	39060.606	**13362.839**
61674.641	53451.356	**16446.571**	**9867.943**	35976.874	**7195.375**
78121.212	28781.499	**12334.928**			

The five bold numbers are those corresponding to the leaves of the first stem. What has been done is to round it to the nearest digit of 4 zeros. For example, 7195.375 corresponds to 7 because it is closer to 7,000 than to 8,000. The other stems are formed following this same criterion.

1.3.4 Box-Plot Diagram

It is a graph that allows us to represent a series of numerical data through its quartiles. It is especially useful for detecting outliers. The diagram is formed by a box and two vertical lines that extend from it. DWe will draw a segment inside the box that corresponds to the median. Let's consider some important aspects:

- We say that an observation is an outlier if it exceeds the lower limit $L_I = Q_1 - 1.5(Q_3 - Q_1)$ or the upper limit $L_S = Q_3 + 1.5(Q_3 - Q_1)$, that is, 1.5 times the interquartile range.
- The upper whisker extends to the largest value that is not an outlier, while the lower whisker extends to the smallest non-outlier value.
- If the minimum value coincides with the first quartile, the lower side of the box will coincide with the end of the lower whisker, while if the maximum value coincides with the third quartile, the upper side of the box will coincide with the end of the upper whisker.

1.4 Excercises

Excercise 1.

Complete the following frequencies table:

	n_i	N_i	f_i	F_i
[0 , 5)			0,25	
[5 , 10)				0,35
[10 , 15)	20			
[15 , 20)		80		
[20 , 25)		100		

Solución. The table should be completed starting from the first row. We are given two important pieces of information: $f_1 = 0.25 = F_1$ and $N_5 = n = 100$, which implies that $n_1 = f_1 \cdot n = 25$. Since $F_2 = 0.35$, we have $f_2 = 0.10$, so $n_2 = f_2 \cdot n = 10$. Using the same reasoning, we complete the third and fourth rows. To complete the fifth row, we use the fact that the sum of the relative frequencies must be equal to 1, hence $F_5 = 1$. The completed table will look like this:

	n_i	N_i	f_i	F_i)
[0 , 5)	25	25	**0.25**	0.25
[5 , 10)	10	35	0.10	**0.35**
[10 , 15)	**20**	55	0.2	0.55
[15 , 20)	25	**80**	0.25	0.8
[20 , 25)	20	**100**	0.2	1

Exercise 2.

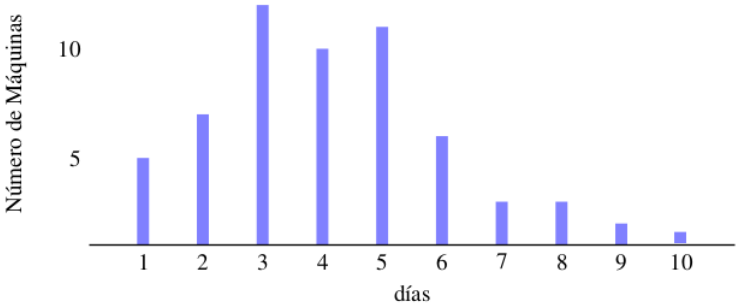
In a factory, the number of days that machines operate without needing repair has been recorded, obtaining the following table:

días	1	2	3	4	5	6	7	8	9	10
Máquinas	5	7	12	10	11	6	3	3	2	1

1. Draw the bar chart for the absolute frequencies and its associated frequency polygon.
2. Calculate the mean, median, and mode.
3. Calculate the variance, coefficient of variation, skewness, and kurtosis.

Solution.

1. We represent the bar chart:



2. In this exercise, x_i represent the days and n_i the absolute frequency, that is, the number of machines associated with each day.

   - $\overline{x} = \dfrac{1}{60}\displaystyle\sum_{i=1}^{10} n_i x_i = 4.32$
   - Since $N_3 = 24$ y $N_4 = 34$, the two central values will be $x_i = 4$, the median then will be $M_e = 4$.
   - $M_o = 3$, since it is the value with the highest frequency.

3. Let's now calculate the measures of dispersion and shape:
   - $s^2 = \dfrac{1}{60}\displaystyle\sum_{i=1}^{10} n_i(x_i - (4.32))^2 = 4.58$
   - $C_v = \dfrac{\sqrt{4.58}}{|\overline{4,32}|} = 0.49$
   - $g_1 = \dfrac{\frac{1}{60}\displaystyle\sum_{i=1}^{10} n_i(x_i - 4.32)^3}{2.14^3} = 0.5637$ (positive skewness)
   - $g_2 = \dfrac{\frac{1}{60}\displaystyle\sum_{i=1}^{10} n_i(x_i - 4.32)^4}{2.14^4} - 3 = -0.139$ (platykurtic distribution)

Exercise 3.

A VTC company records every day the number of trips made and the distances traveled by its workers in each service. The following table shows the data obtained in the city of Madrid on a randomly chosen working day:

Distance (km)	Number of trips
[0 , 2)	15
[2 , 4)	20
[4 , 6)	25
[6 , 10)	10
[10 , 20)	6
[20 , 50)	4
[50 , 100)	2

1. Calculate the mean, median, and mode.
2. Calculate the standard deviation and the coefficient of variation.
3. Calculate the skewness and kurtosis coefficients.

Solution. When dealing with variables grouped in intervals, x_i will be the midpoint of each interval.

1. Measures of central tendency:
   - $\overline{x} = \dfrac{1}{82}\sum_{i=1}^{7} n_i x_i = 8.05$
   - To calculate the median, we need to find the first interval whose cumulative frequency exceeds $\dfrac{n}{2} = 41$. In this case, $[L_{i-1}, L_i) = [4, 6)$. Applying formula (1.2):
     $M_e = L_{i-1} + \dfrac{n/2 - N_{i-1}}{n_i} \cdot (L_i - L_{i-1}) = 4 + \dfrac{41 - 35}{25} \cdot (6 - 4) = 4.48$
   -
   - To calculate the mode, we look for the interval where the highest absolute frequency is located. In this case,
     $[L_{i-1}, L_i) = [4, 6)$. Aplicamos ahora la fórmula (1.4):
     $M_o = L_{i-1} + \dfrac{\delta_1}{\delta_1 + \delta_2} \cdot (L_i - L_{i-1}) = 4 + \dfrac{25 - 20}{(25 - 20) + (25 - 10)} \cdot (6 - 4) = 4.5$
2. Measures of dispersion:
   - $var[x] = s^2 = \dfrac{1}{82}\sum_{i=1}^{7} n_i (x_i - (8.05))^2 = 166.43 \quad \Longrightarrow \quad s = 12.9$
   - $C_v = \dfrac{12.9}{8.05} = 1.6$
3. Measures of shape:
   - $g_1 = \dfrac{\dfrac{1}{82}\sum_{i=1}^{7} n_i (x_i - 8.05)^3}{12.9^3} = 3.81$

- $g_2 = \dfrac{\dfrac{1}{82}\displaystyle\sum_{i=1}^{7} n_i(x_i - 8.05)^4}{12.9^4} - 3 = 15.64$

Ejercicio 4.

In an industrial quality control, we take a sample of trucks with an average weight of 5,000 kilograms and a standard deviation of 400 kilograms, and another sample of electronic components with an average weight of 15 grams and a standard deviation of 5 grams. Calculate the coefficient of variation for both samples and interpret the results.

Solución. Both variables measure the same magnitude, weight, but with different units. If we convert grams to kilograms, the standard deviation of the sample of electronic components will be 0.005 kg, which is very small compared to the sample of trucks. The coefficient of variation will allow us to compare these groups since it is a dimensionless measure that only takes into account the proportion between standard deviation and mean. Let's calculate it in both cases:

- Trucks: $C_v = \dfrac{\sqrt{400}}{5000} = 0.08$
- components: $C_v = \dfrac{\sqrt{5}}{15} = 0.33$

We can see that the sample of electronic components has a higher variability compared to the sample of trucks.

Excercise 5.

A population of students has an average height of 160 cm with a standard deviation of 16 cm. These same students have an average weight of 70 kg with a standard deviation of 14 kg. Which of the 2 variables has greater relative variability?

Solution. To study the variability of two variables of different magnitudes, we use the coefficient of variation:

- height: $C_v = \dfrac{\sqrt{16}}{160} = 0.1$
- weight: $C_v = \dfrac{\sqrt{14}}{70} = 0.2$

Contrary to what may seem at first glance, the height of the students has less variability.

Excercise 6.

The height of a group of 50 high school students has been measured. The results are collected in the following table:

138	167	151	170	175	138	148	153	178	142
137	157	145	146	148	155	167	142	154	133
133	152	157	149	169	159	148	150	153	145
140	161	156	149	152	140	146	151	143	140
152	138	160	153	165	157	158	162	155	144

The following tasks are requested:

1. Construct a frequency table of grouped data in class intervals of length 10.
2. Calculate the interquartile range.
3. Calculate the mean, variance, and coefficient of variation.
4. Plot the Box-Plot diagram.

Solution.

1. The table will be as follows:

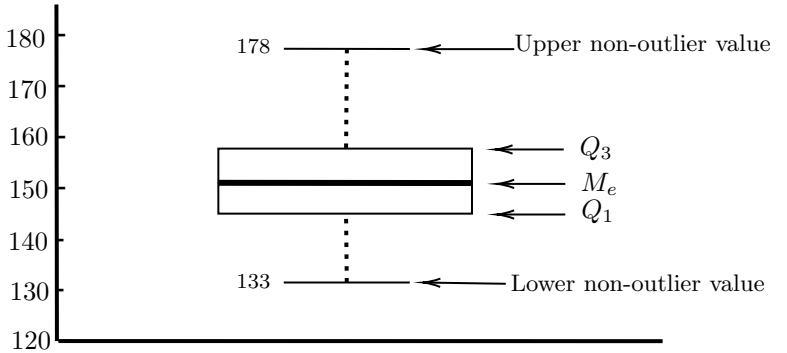
	n_i	N_i	f_i	F_i
[130 , 140)	6	6	0.12	0.12
[140 , 150)	16	22	0.32	0.44
[150 , 160)	18	40	0.36	0.8
[160 , 170)	7	47	0.14	0.94
[170 , 180)	3	50	0.06	1

2. We have to calculate Q_1 y Q_3.
   - To calculate Q_1 we need to find the first interval whose cumulative frequency exceeds $\frac{n \cdot 25}{100} = 12.5$. In this case, $[L_{i-1}, L_i) = [140, 150)$.
     $$Q_1 = P_{25} = L_{i-1} + \frac{\frac{n \cdot 25}{100} - N_{i-1}}{n_i} \cdot (L_i - L_{i-1}) = 140 + \frac{\frac{50 \cdot 25}{100} - 6}{16} \cdot (150 - 140) = 144.06$$
   - As $\frac{n \cdot 75}{100} = 37.5$, then:
     $$Q_3 = P_{75} = L_{i-1} + \frac{\frac{n \cdot 75}{100} - N_{i-1}}{n_i} \cdot (L_i - L_{i-1}) = 140 + \frac{\frac{50 \cdot 25}{100} - 22}{18} \cdot (150 - 140) = 158.61$$

Therefore, the interquartile range will be $IQR = Q_3 - Q_1 = 158.61 - 144.06 = 14.55$

3. Mean, variance, and coefficient of variation:
   - $\overline{x} = \dfrac{1}{50}\sum_{i=1}^{5} n_i x_i = 152$
   - $s^2 = \dfrac{1}{50}\sum_{i=1}^{5} n_i (x_i - (152))^2 = 53.6$
   - $C_v = \dfrac{7.32}{152} = 0.048$
4. In order to represent the Box-plot diagram, we will need this five values:
   - $M_e = 150 + \dfrac{25-22}{18} \cdot (160 - 150) = 151.67$
   - $Q_1 = 144.06$
   - $Q_3 = 158.61$
   - $L_I = Q_1 - 1.5R_I = 122.23$
   - $L_S = Q_3 + 1.5R_I = 180.43$

   There is no outlier, so the box will be as follows:



Exercise 7.

In a sociological study on the number of children in families in Seville, the following data has been obtained by asking 50 families:

2	4	2	3	1	2	4	2	3	0
2	2	2	3	2	6	2	3	2	2
3	2	3	3	4	1	3	3	4	5
2	0	3	2	1	2	3	2	2	3
1	6	2	3	2	4	3	3	2	2

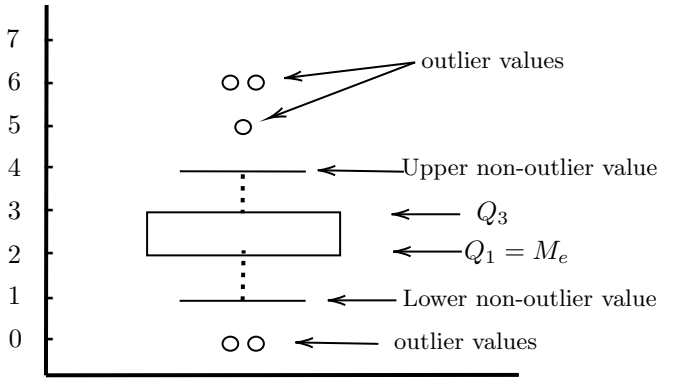
1. Calculate the quartiles.
2. Draw the Box-Plot diagram.

Solution. We will calculate the quartiles taking into account that they are ungrouped data. It will help us to have the table of cumulative absolute frequencies:

x_i	0	1	2	3	4	5	6
N_i	2	6	27	42	47	48	50

1. Quartiles:
   - As n is even, the median will be the average of the two central values: $M_e = 2$
   - We will first calculate the position of the two quartiles within the ordered data group from lowest to highest. As $\frac{n}{4} = 12.5$, we will interpolate between the values $x_{13} = 2$ and $x_{14} = 2$, then $Q_1 = 2$. Reasoning in the same way, we will arrive at $Q_3 = 3$.
2. To draw the box, we need these 5 values:
   - $M_e = 2$
   - $Q_1 = 2$
   - $Q_3 = 3$
   - $L_I = 2 - 1.5R_I = 0.5$
   - $L_S = 3 + 1.5R_I = 4.5$

   There are 2 outlier values: 5 and 6. The maximum and minimum non-outlier values are 4 and 1, respectively, so the box would look like this:



Exercise 8.

450 candidates have applied for a local police force in a specific town in Spain. To prepare for the physical exam, the candidates have been grouped by height. The data is shown in the following table:

Height (meters)	Number of candidates
[1.65 , 1.70)	30
[1.70 , 1.75)	70
[1.75 , 1.80)	120
[1.80 , 1.85)	150
[1.85 , 1.90)	80

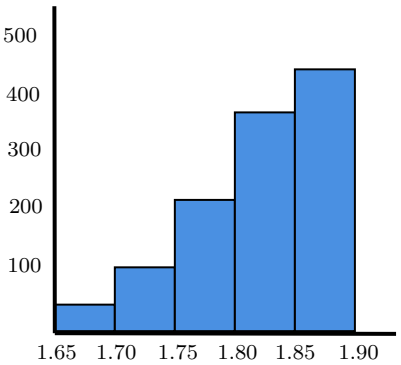
1. Discuss the meaning of the third cumulative absolute frequency.
2. Plot the histogram of cumulative absolute frequencies.
3. Calculate the first and third quartiles.
4. If a candidate measures 1.79 m, will he be above or below the 80th percentile?

Solution.

1. We calculate the table of cumulative absolute frequencies:

Height (meters)	N_i
[1.65 , 1.70)	30
[1.70 , 1.75)	100
[1.75 , 1.80)	220
[1.80 , 1.85)	370
[1.85 , 1.90)	450

   The third cumulative absolute frequency, $N_3 = 220$, is the number of candidates with a height less than 1.80 meters.
2. We plot the histogram:



3. Let's calculate the quartiles:

- To calculate Q_1, we need to find the first interval whose cumulative frequency exceeds $\frac{n}{4} = 112.5$. In this case, $[L_{i-1}, L_i) = [1.75, 1.80)$, hence: $Q_1 = P_{25} = 1.75 + \frac{112.5 - 100}{120} \cdot (1.80 - 1.75) = 1.755$
- As $\frac{3n}{4} = 337.5$, then: $Q_3 = 1.80 + \frac{337.5 - 220}{150} \cdot (1.85 - 1.80) = 1.839$

4. Let's calculate the 80th percentile:
   The first thing is to determine in which interval it is located. Since $\frac{80 \cdot 450}{100} = 360$, it will be located in [1.80,1.85), so the percentile will be: $P_{80} = 1.80 + \frac{360 - 220}{150} \cdot 0.05 = 1.846$. Therefore, this aspirant will be below the 80th percentile.

Exercise 9.

Using the data from the previous exercise on heights, we want to know:

1. Mean and variance.
2. Coefficient of variation.

Solution.

1. Mean and variance:
   - $\overline{x} = \frac{1}{450} \sum_{i=1}^{5} n_i x_i = 1.795$
   - $var[x] = \frac{1}{450} \sum_{i=1}^{5} n_i (x_i - (1.795))^2 = 0.016$
2. $C_v = \frac{\sqrt{0.016}}{1.795} = 0.071$

Exercise 10.

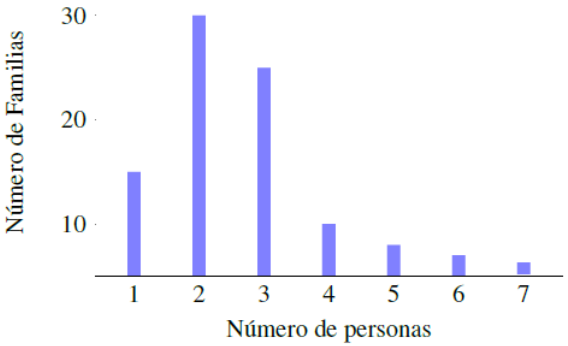
In a community consisting of 100 neighbors, they have been classified according to the number of people in each household:

Persons per household	Number of households	Cumulative
1	15	15
2	30	45
3	25	70
4	10	80
5	8	88
6	7	95
7	5	100

1. Represent the bar chart for the absolute frequencies.
2. Calculate the mean, median, and mode.
3. Calculate the skewness and kurtosis.
4. Obtain the value of the 80th percentile.

Solution.

1. Bar chart:



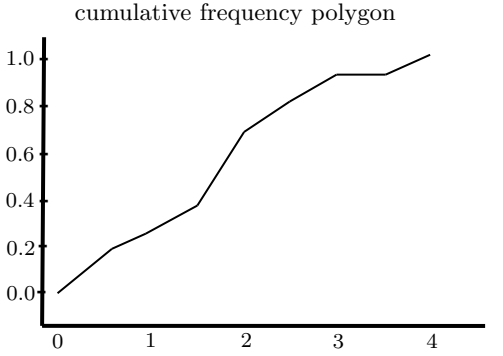
2. Mean, standard deviation, median, and mode:
   - $\overline{x} = \dfrac{1}{100}\displaystyle\sum_{i=1}^{7} n_i x_i = 3.07$
   - $s = \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{7}(x_i - 3.07)^2\right)^{1/2} = 1.6568$
   - Since 100 is odd, the median will be the average of the two central values, that is, $M_e = 3$.
   - To calculate the mode, select the x_i with the highest absolute frequency. Therefore, $M_o = 2$.
3. Skewness and kurtosis:
   - $g_1 = \dfrac{\frac{1}{100}\displaystyle\sum_{i=1}^{7} n_i (x_i - 3.07)^3}{1.6568^3} = 0.8251$
   - $g_2 = \dfrac{\frac{1}{100}\displaystyle\sum_{i=1}^{7} n_i (x_i - 3.07)^4}{1.6568^4} - 3 = -0.1574$
4. To calculate the 80th percentile, we have to look at the cumulative frequency column: it is clear that it will be between $x_{80} = 4$ and $x_{81} = 5$. Now, we only need to calculate f for both values and interpolate linearly:

x_i	$x_{80} = 4$	$x_{81} = 5$
f	0.798	0.818

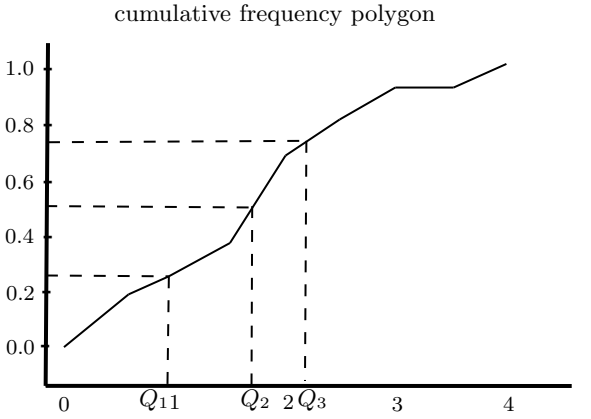
$$P_{80} = 5 - \frac{(5-4)\cdot(0.818-0.80)}{(0.818-0.798)} = 4.1$$

Excercise 11.

In the following cumulative frequency polygon, graphically determine the three quartiles:

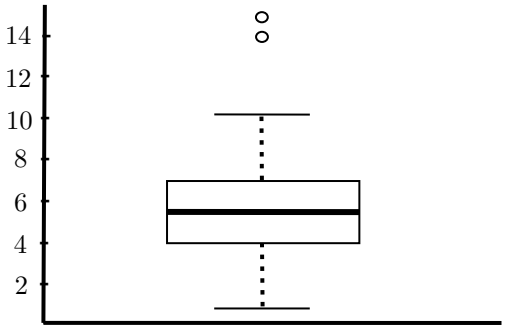


Solution. Q_1, the median, and Q_3 are the values that leave behind 25

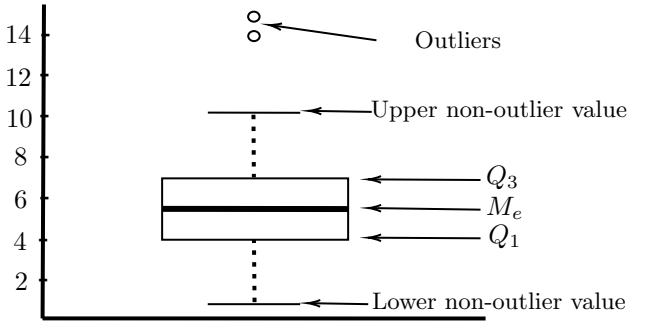


Excercise 12.

Identify the most representative values of the following box plot:



Solution. In the following diagram we indicate the values that make up the Box-Plot diagram:



Excercise 13.

The size of the femur has been measured in a group of children aged between 12 and 18 months, obtaining the following data:

Femur size (cm)	Number of children
[38 - 40)	5
[40 - 42)	7
[42 - 44)	8
[44 - 46)	6
[46 - 48)	4

1. Construct the frequency table.
2. Calculate the mean, median and variance.
3. Calculate the coefficient of variation and interpret the result.

Solution.

1. We represent the frequency table:

	x_i	n_i	N_i	f_i	F_i
[38 - 40)	39	5	5	0.167	0.167
[40 - 42)	41	7	12	0.233	0.4
[42 - 44)	41	8	20	0.267	0.667
[44 - 46)	45	6	26	0.2	0.867
[46 - 48)	47	4	30	0.133	1

2. Mean, median, and variance:
   - $\overline{x} = \dfrac{1}{30}\displaystyle\sum_{i=1}^{5} n_i x_i = 42.8$
   - Para calcular M_e We will have to find the first interval whose cumulative frequency exceeds $\dfrac{n}{2} = 15$.
     In this case, $(L_{i-1}, L_i) = [42, 44)$. Applying the formula (1.2):
     $M_e = L_{i-1} + \dfrac{n/2 - N_{i-1}}{n_i} \cdot (L_i - L_{i-1}) = 42 + \dfrac{30/2 - 12}{8} \cdot (44 - 42) = 42.75$
   - $s^2 = \dfrac{1}{30}\displaystyle\sum_{i=1}^{5} (x_i - \overline{x})^2 = 2.548$
3. $C_v = \dfrac{\sqrt{2.548}}{42.8} = 0.059$. Then there is a 5.9% variability with respect to the mean.

Exercise 14.

The age distribution of the unemployed population in a certain city is as follows:

Age	Number of unemployed
[16 - 26)	621
[26 - 36)	162
[36 - 46)	113
[46 - 56)	93
[56 - 66)	22

1. Find the mean age and variance of the unemployed population.
2. Construct the histogram of cumulative absolute frequencies.

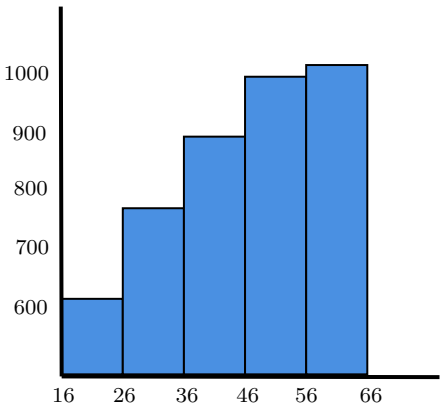
3. Calculate the coefficient of skewness.

Solution.

1. Mean and variance:
   - $\overline{x} = \dfrac{1}{1011}\displaystyle\sum_{i=1}^{5} n_i x_i = \dfrac{28781}{1011} = 28.47$
   - $s^2 = \dfrac{1}{1011}\displaystyle\sum_{i=1}^{5}(x_i - 28.47)^2 = 122.54$
2. First, we construct the table of cumulative absolute frequencies:

Age	Cumulative frequency
[16 - 26)	621
[26 - 36)	783
[36 - 46)	896
[46 - 56)	989
[56 - 66)	1011

   Now we represent the histogram:



3. $g_1 = \dfrac{\dfrac{1}{1011}\displaystyle\sum_{i=1}^{5} n_i (x_i - 28.47)^3}{11.07^3} = 1.3$. The skewness is clearly positive, as already perceived in the histogram.

Excercise 15.

We have the following information about the weekly leisure expenses of a group of university students:

Expenses (euros)	Number of students
[0 - 5)	4
[5 - 10)	11
[10 - 15)	16
[15 - 20)	22
[20 - 25)	8
[25 - 30)	6

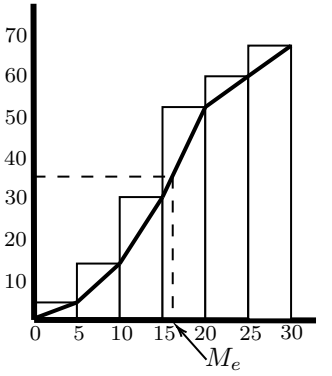
1. Build the frequency table
2. Determine the median graphically.
3. Determine the mode graphically.
4. Calculate the mean and variance.

Solution.

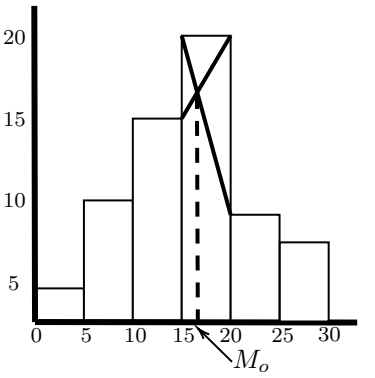
1. Frequency table:

	x_i	n_i	N_i	f_i	F_i
[0 - 5)	2.5	4	4	0.06	0.067
[5 - 10)	7.7	11	15	0.164	0.224
[10 - 15)	12.5	16	31	0.239	0.463
[15 - 20)	17.5	22	53	0.328	0.791
[20 - 25)	22.5	8	61	0.119	0.91
[25 - 30)	27.5	6	67	0.09	1

2. We use the cumulative frequency polygon. As $\frac{n}{2} = 33.5$, then:



3. To calculate the mode graphically, we rely on the histogram of absolute frequencies and apply a proportion in the interval of highest frequency:



4. Let's calculate the mean and variance:
   - $\overline{x} = \frac{1}{67}\sum_{i=1}^{6} n_i x_i = 15.26$
   - $var[x] = \frac{1}{67}\sum_{i=1}^{6}(x_i - 15.26)^2 = 43.4$

Exercise 16.

A study is carried out on the hotel capacity in a certain city, obtaining the following data:

Hotel capacity	Number of hotels
[0 - 20)	25
[20 - 40)	50
[40 - 60)	55
[60 - 80)	30
[80 - 100)	10

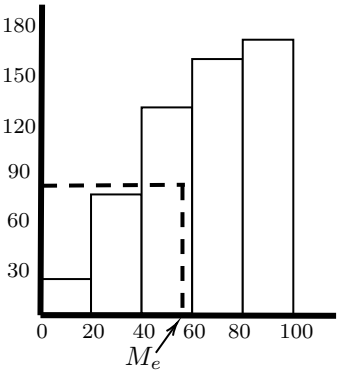
1. Construct the frequency table.
2. Determine the mode and median graphically.
3. Calculate the coefficient of variation and interpret the results.

Solution.

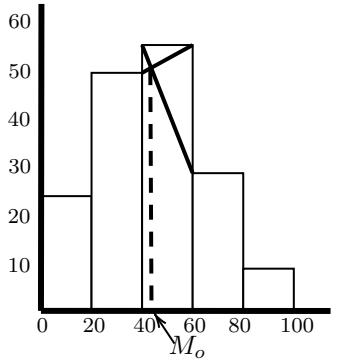
1. We construct the frequency table:

	x_i	n_i	N_i	f_i	F_i
[0 - 20)	10	25	25	0.147	0.147
[20 - 40)	30	50	75	0.294	0.441
[40 - 60)	50	55	130	0.324	0.765
[60 - 80)	70	30	160	0.176	0.941
[80 - 100)	90	10	170	0.059	1

2. We use the cumulative frequency polygon. Since $\frac{n}{2} = 85$:



3. We apply the proportion in the modal interval:



4. We calculate the mean and standard deviation beforehand:
   - $\overline{x} = \frac{1}{170}\sum_{i=1}^{5} n_i x_i = 44.117$
   - $var[x] = \frac{1}{170}\sum_{i=1}^{5}(x_i - 44.117)^2 = 483.032 \implies \text{s}=\sqrt{483.032} = 21.978$

- $C_v = \dfrac{s}{|\overline{x}|} = \dfrac{21.978}{44.117} = 0.498$. This value shows that there is a variability with respect to the mean of 49.8%, which can be considered as very important.

Exercise 17.

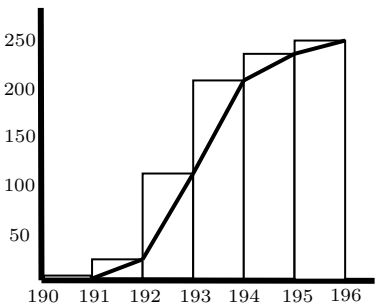
The following table shows the distribution of the lengths of a sample of 250 pieces obtained in the manufacturing process of a certain product:

Diameter (mm)	Number of pieces
[190 - 191)	2
[191 - 192)	18
[192 - 193)	92
[193 - 194)	99
[194 - 195)	26
[195 - 196)	13

1. Draw the cumulative frequency polygon.
2. Calculate the first and third quartiles.

Solution.

1. First we have to represent the histogram and then trace the polygon:



2. Quartile:
   - To calculate Q_1, we will need to find the first interval whose cumulative frequency exceeds $\dfrac{n}{4} = 62.5$.
     In this case, $[L_{i-1}, L_i) = [192, 193)$, then:
     $Q_1 = 192 + \dfrac{62.5 - 20}{92} \cdot (193 - 192) = 192.46$

- Since $\frac{3n}{4} = 187.5$, Q_3 will be in $[L_{i-1}, L_i) = [193, 194)$. Therefore: $Q_3 = 193 + \frac{187.5 - 112}{99} \cdot 1 = 193.76$

Ejercicio 18.

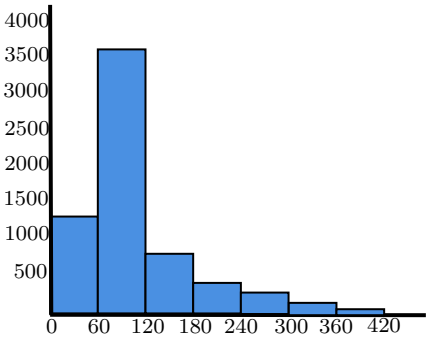
A parking lot charges 1.5 euro cents per minute of parking. Data on the occupancy of the parking lot are collected for a week and are shown in the following table:

Time (minutes)	Number of vehicles
[0 - 60)	1240
[60 - 120)	3575
[120 - 180)	746
[180 - 240)	327
[240 - 300)	218
[300 - 360)	130
[360 - 420)	40

1. Determine the most frequent parking time.
2. Plot the histogram of absolute frequencies and calculate the coefficient of skewness.
3. Calculate the weekly average parking time and the average revenue earned by the company during that week.
4. The company is considering changing the existing rate as follows: a fee of 50 euro cents will be charged at entry and 1.1 euros per minute of parking. Which alternative would be more advantageous for the company?.

Solution.

1. To obtain the most frequent parking time, we need to calculate the mode of this data distribution. First, we need to find the interval where the highest absolute frequency is located, in this case, $[L_{i-1}, L_i) = [60, 120)$. We now apply formula (1.4): $M_o = L_{i-1} + \frac{\delta_1}{\delta_1 + \delta_2} \cdot (L_i - L_{i-1}) = 60 + \frac{3575 - 1240}{(3575 - 1240) + (3575 - 746)} \cdot (120 - 60) = 87.13$
2. We plot the histogram:



$g_1 = \dfrac{\frac{1}{6276}\sum_{i=1}^{7} n_i(x_i - 104.67)^3}{67.57^3} = 1.71$. The skewness is clearly positive, as can be seen in the histogram.

3. $\overline{x} = \dfrac{1}{6276}\sum_{i=1}^{7} n_i x_i = 104.67$ minutes.
   The average income will be: 6276 x 104.67 x 0.015 = 9853.63 euros.
4. With the new pricing, the company would earn the following: 6276 x 0.5 + 6276 x 104.67 x 0.011 = 10363.99 euros.
   The pricing modification would benefit the company.

Exercise 19.

A steel parts manufacturing company has taken a sample of 40 pieces and measured their length in millimeters, obtaining the following results:

160	169	173	167	166	172	169	166	170	173
187	181	179	180	177	168	163	169	162	173
161	170	164	167	172	167	164	190	167	173
163	179	170	168	180	179	166	167	168	174

Represente el diagrama Box-Plot.

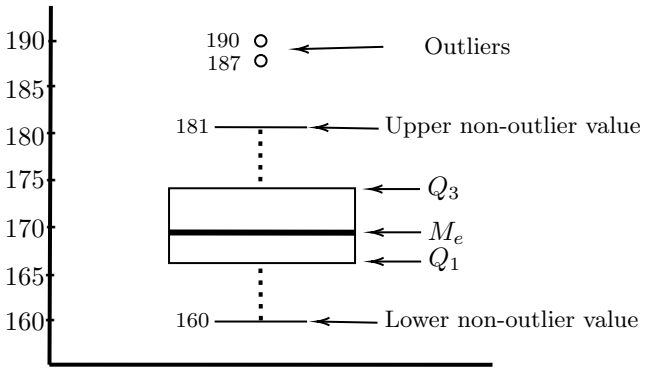
Solución. En primer lugar, vamos a calcular el primer y tercer cuartil. Luego debemos ordenar los datos de menor a mayor y calcular la tabla de frecuencias absolutas acumuladas. Es claro que el primer cuartil estará entre 166 y 167, mientras que el tercero estará entre 173 y 174.

x_i	166	167	173	174
f	0.231	0.256	0.744	0.769

Remember that the value of f was calculated as follows: $f = \dfrac{i-1}{n-1}$, where i is the position of x_i once the data are sorted from smallest to largest. For example, the value of f that we have calculated for 166 refers to the value that occupies the tenth position ($i = 10$, $n = 40$). Now, to find Q_1 and Q_3, we perform linear interpolation:

- $Q_1 = 167 - \dfrac{(167-166)\cdot(0.256-0.25)}{(0.256-0.231)} = 166.76$
- $Q_3 = 174 - \dfrac{(174-173)\cdot(0.769-0.75)}{(0.769-0.744)} = 173.24$
- To calculate the median, we will take the arithmetic mean of the two central values, since $n = 40$ is even. Therefore, $Me = 169$.
- $L_I = Q_1 - 1.5R_I = 166.76 - 1.5(173.24 - 166.76) = 157.04$
- $L_S = Q_3 + 1.5R_I = 173.24 + 1.5(173.24 - 166.76) = 182.96$

Hay 2 valores outliers: 187 y 190. Los valores máximo y mínimo no outliers son 181 y 160, respectivamente, so the box-plot will be like that:



Excercise 20.

Using the data from the previous exercise, calculate the stem-and-leaf plot.

160	169	173	167	166	172	169	166	170	173
187	181	179	**180**	177	168	163	169	162	173
161	170	*164*	167	172	167	*164*	190	167	173
163	179	170	168	**180**	179	166	167	168	174

Solution. Since we have three-digit numbers, we will form the stem with the hundreds and tens digits. The leaves will be formed by the units digit. The class interval amplitude will be 10 units. The plot will look like this:

16	0 1 2 3 3 4 4 6 6 6 7 7 7 7 7 7 8 8 8 8 9 9 9
17	0 0 0 2 2 3 3 3 4 7 9 9 9
18	**0** **0** 1 7
19	0

Chapter 2

Parameter Estimation

In this and the following chapters, we will assume that all individuals in a simple random sample follow a known distribution. However, such distribution will always depend on unknown parameters that we will have to estimate and about which we lack prior information. Only with the help of sample data can we perform the estimation. This is the classical approach to Statistics, and it is in contrast to the Bayesian estimation method.

2.1 Point Estimation

As the name suggests, in point estimation, a point value is assigned that approximates the true value of the parameter being estimated.

Definition. Given a sample of identically distributed random variables $X_1, \ldots, X_n$, we call any function $T(X_1, \ldots, X_n)$ a "Statistic". As it is a function of random variables, it will also be a random variable, and hence will have a probability distribution called the "sampling distribution".

Definition. We call any statistic that does not depend on an unknown parameter θ and helps estimate its value an "Estimator" of θ, denoted as $\hat{\theta}$.

Example

Suppose that a screw manufacturing company wants to check the size of the screws. To do this, they take a sample of size $n = 5$. The results are: $4.97; 5.01; 5.01; 5.2; 4.98$. To make a point estimate of the population mean μ, they use two different estimators:

$$\hat{\mu_1} = \overline{X} \quad \text{y} \quad \hat{\mu_2} = \frac{X_1 - 3X_2 - X_3 - X_4 + 4X_5}{3}$$

Once we have taken the sample data, the random variables X_i take specific values, that is, $X_i = x_i$; then, we can already make the estimation:

$$\hat{\mu_1} = \overline{x} = 5,034 \quad \text{y} \quad \hat{\mu_2} = \frac{x_1 - 3x_2 - x_3 - x_4 + 4x_5}{3} = 3,22$$

How do we know which of the two is more reliable? For an estimator to be

reliable, it must meet a series of properties. Let's look at the three most important ones.

Properties

- **Unbiasedness.** An estimator $\hat{\theta}$ is *unbiased* if $E[\hat{\theta}] = \theta$. This property tells us that the expected value of the estimates is precisely θ.
- **Consistency.** An estimator $\hat{\theta}$ is *consistent* if as the sample size increases, its expected value approaches θ and its variance becomes smaller and smaller.
- **Efficiency.** Given the estimators $\hat{\theta_1}$ and $\hat{\theta_2}$, we say that $\hat{\theta_1}$ is more efficient than $\hat{\theta_2}$ if $Var[\hat{\theta_1}] < Var[\hat{\theta_2}]$.

Mean Squared Error (MSE)

Sometimes the problem arises of choosing between an unbiased estimator, $\hat{\theta_1}$, and a biased one, $\hat{\theta_2}$, but with lower variance. In these cases, the estimator with the lower average error in predicting the parameter is usually chosen:

$$ECM = E[(\hat{\theta} - \theta)^2] = (E[\hat{\theta}] - \theta])^2 + Var[\hat{\theta}]$$

Si $\hat{\theta}$ is unbiased,

$$ECM = Var[\hat{\theta}]$$

2.1.1 Method of Moments

This estimation method is based on equating the theoretical moments of the distribution to the corresponding sample moments, obtaining equations whose solutions will be the estimators.

Example

The useful mean life of a printer can be modeled by a random variable with density function:

$$f(x) = \begin{cases} \theta\, e^{-\theta x} & x > 0 \\ 0 & c.c. \end{cases}$$

We have taken a sample of ten printers with the following results (in years): 7, 4, 3, 5, 6, 4, 3, 4, 5, 7. Let's calculate an estimator of θ using the method of moments.

Since we only have to estimate one parameter, we will calculate the theoretical

mean (first moment) and equate it to the sample mean.

$$m_1 = E[X] = \int_0^1 x\,\theta\,e^{-\theta x}dx = \left[-e^{-\theta x}\left(x + \frac{1}{\theta}\right)\right]_0^{\infty} = \frac{1}{\theta} = \overline{x}$$

Solving for θ,

$$\theta = \frac{1}{m_1}$$

Therefore, the estimator of θ using the method of moments will be:

$$\hat{\theta} = \frac{1}{\overline{x}} = \frac{1}{4.8} = 0.208$$

2.1.2 Maximum likelihood method

The idea of this method is based on what is known as *likelihood*. Suppose I select a simple random sample $(x_1, x_2, \ldots, x_n)$ that comes from a random variable X that depends on a parameter θ. We want to define a function that calculates the probability of obtaining that sample. This function, which depends on θ, is known as the *likelihood function* and is denoted as:

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

Being $f(x;\theta)$ the probability density function of the variable X. Our objective is to maximize this function with respect to θ, i.e., find the value of θ that makes the probability of what has already happened (the sample we have) maximum. In general, it is easier to maximize the natural logarithm of the likelihood function; this function is called the *support function*. Therefore, if $L(x_1, \ldots, x_n; \theta) = \ln[l(x_1, \ldots, x_n; \theta)]$, then:

$$\hat{\theta}_{MV} \;/\; L(\hat{\theta}_{MV}) = \max_{\theta} L(x_1, \ldots, x_n; \theta)$$

Example.

We will calculate the maximum likelihood estimator (MLE) using the density function and the sample data from the previous example. First, we will construct the likelihood function:

$$l(x_1, \ldots, x_n; \theta) = \begin{cases} \theta^n\, e^{-\theta \sum_{i=1}^{n} x_i} & min\,\{x_i\} > 0 \\ 0 & c.c. \end{cases}$$

Its natural logarithm will be:

$$L(x_1, \ldots, x_n; \theta) = n \ln(\theta) - \theta \sum_{i=1}^{n} x_i$$

By taking the derivative with respect to θ and setting it to zero, we get:

$$\frac{\partial L(x_1, \ldots, x_n; \theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

Solving for θ, we obtain the maximum likelihood estimator:

$$\hat{\theta}_{MV} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}} = 0.208$$

2.2 Interval Estimation

While point estimation provides a single number that approximates the value of an unknown parameter θ, interval estimation provides an interval $\hat{\theta}_I$, $\hat{\theta}_S$, such that

$$p(\hat{\theta}_I < \theta < \hat{\theta}_S) = 1 - \alpha$$

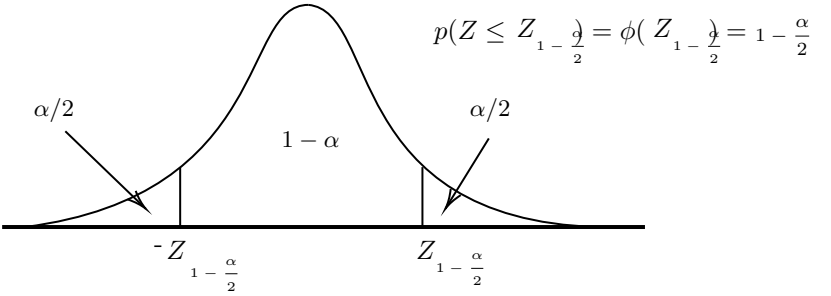
where $\hat{\theta}_I$ and $\hat{\theta}_S$ depend on the distribution of an estimator of θ.

The value α (which is usually between 0.1 and 0.05) is called the *significance level*; $(1 - \alpha) \cdot 100\%$ is called the *confidence level*; and the interval $(\hat{\theta}_I, \hat{\theta}_S)$ is called the *confidence interval* for θ.

2.2.1 Estimation for the mean when the variance is known

Let's suppose that we have a simple random sample $X_1, \ldots, X_n$ that comes from a normal distribution with mean μ and standard deviation σ, the latter being known. To estimate μ with a confidence level of $(1 - \alpha) \cdot 100\%$, we will use the sample mean estimator. We know that

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



As we observe in the graph and taking advantage of the fact that the normal distribution is symmetrical with respect to its mean, we take symmetrical values so that the interval is as small as possible. Therefore,

$$p\left(-Z_{1-\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq Z_{1-\alpha/2}\right) = 1-\alpha \implies$$
$$\implies p\left(\overline{X} - \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}\right) = 1-\alpha$$

Thus, the confidence interval will be as follows:

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2} \ , \ \overline{X} + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}\right)$$

Remarks.

1. Although the sample does not come from a normal distribution, the Central Limit Theorem ensures us that, asymptotically:

$$\overline{X} \sim \mathcal{N}\left(\mu, \sigma/\sqrt{n}\right)$$

   with the approximation being very good when $n > 30$.
2. If, for example, we have a 95

2.2.2 Estimation for the mean when the variance is unknown

Suppose we have a simple random sample $X_1, \ldots, X_n$ that comes from a normal distribution with mean μ and standard deviation σ. If the parameter σ is unknown, instead of using the previous interval we will have to use the t-student distribution. We will rely on the estimators

$$\hat{\mu} = \overline{X} \quad \text{y} \quad \hat{\sigma}^2 = S^2$$

We know that

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Taking advantage of the fact that the t-distribution is symmetrical with respect to its mean, we take symmetrical values so that the interval is as small as possible. Therefore,

$$p\left(-t_{n-1,1-\alpha/2} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) = 1-\alpha$$

Thus, the confidence interval will be as follows:

$$\left(\overline{X} - \frac{S}{\sqrt{n}}\, t_{n-1,1-\alpha/2} \;\;,\;\; \overline{X} + \frac{S}{\sqrt{n}}\, t_{n-1,1-\alpha/2}\right)$$

Donde

$$p(t \leq t_{n-1,1-\alpha/2}) = 1 - \frac{\alpha}{2}$$

2.2.3 Estimation for the difference of means

Known variances

Given independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_k$ from two normal distributions, $\mathcal{N} > (\mu_1, \sigma_1) >$ and $\mathcal{N} > (\mu_2, \sigma_2) >$, respectively, we will calculate the confidence interval for $\mu_1 - \mu_2$ using the following estimator:

$$\overline{X} - \overline{Y}$$

Before constructing the interval, we must know the distribution of this estimator. Its expected value is:

$$E[\overline{X} - \overline{Y}] = E[\overline{X}] - E[\overline{Y}] = \mu_1 - \mu_2$$

and its variance:

$$Var[\overline{X} - \overline{Y}] = Var[\overline{X}] + Var[\overline{Y}] = \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}$$

Therefore,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1}{n_1} + \dfrac{\sigma_2}{n_2}}} \sim \mathcal{N}(0,1)$$

The confidence interval will be:

$$\left(\overline{X}-\overline{Y}-\sqrt{\frac{\alpha_1}{n_1}+\frac{\alpha_2}{n_2}}\cdot Z_{1-\alpha/2}\,,\ \overline{X}-\overline{Y}+\sqrt{\frac{\alpha_1}{n_1}+\frac{\alpha_2}{n_2}}\cdot Z_{1-\alpha/2}\right)$$

Equal unknown variances

The starting assumptions are the same as in the previous section except that, in this case, the variances are unknown but equal, that is, $\sigma_1^2 = \sigma_2^2$. The confidence interval will be as follows:

$$\left(\overline{X}-\overline{Y}-\sqrt{\overline{S}^2\frac{1}{n_1}+\frac{1}{n_2}}\,t_{n_1+n_2-2,1-\alpha/2}\,,\overline{X}-\overline{Y}+\sqrt{\overline{S}^2\frac{1}{n_1}+\frac{1}{n_2}}\,t_{n_1+n_2-2,1-\alpha/2}\right)$$

where $\overline{S}^2$ is a weighted average of the sample variances:

$$\overline{S}^2=\frac{(n_1-1)S_x^2+(n_2-1)S_y^2}{n_1+n_2-2}$$

Unknown and Unequal Variances

For this case, the following estimator is used:

$$T=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_x^2}{n_1}+\frac{S_y^2}{n_2}}}\ \sim\ t_g$$

where g is the integer value closest to the expression

$$\frac{\left(\frac{S_x^2}{n_1}+\frac{S_y^2}{n_2}\right)^2}{\frac{(S_x^2/n_1)^2}{n_1-1}+\frac{(S_y^2/n_2)^2}{n_2-1}}$$

The confidence interval will be as follows:

$$\left(\overline{X}-\overline{Y}-\sqrt{\frac{S_x^2}{n_1}+\frac{S_y^2}{n_2}}\,t_{g,1-\alpha/2}\,,\ \overline{X}-\overline{Y}+\sqrt{\frac{S_x^2}{n_1}+\frac{S_y^2}{n_2}}\,t_{g,1-\alpha/2}\right)$$