PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research-University of Relizane

Faculty of Science and Technology

ST Domain

Educational handout

Descriptive statistics

Courses and exercises on descriptive statistics

HOCINE KAMEL

2023-2024

Preamble

To solve engineering problems, data must be collected, described and analysed to produce summary information. The role of descriptive statistics is to give a summary idea of the data by calculating a number of statistics and using graphical representations.

The aim of the course is to introduce students to the basic principles of statistics. Learn the main techniques of univariate and bivariate descriptive statistics.
Its main aim is to introduce the fundamental concepts and elementary methods of statistics, so that students can later learn complementary methods independently.

The aim is to develop the critical thinking required when implementing and interpreting statistical processing. To achieve this, a rigorous mathematical framework will be introduced and used. We will provide as many examples as are necessary for a better understanding of the course.

Table of contents	
Chapter 1: **Descriptive statistics**	1
Introduction	1
1. Statistical vocabularies	2
Exercises	3
Chapter 2: One-variable statistical series	5
2.1. Statistical tables	5
2.2. Graphical representation.	9
2.3. Characteristic parameters.	13
2.3.1 Central tendency characteristics	13
2.3.2. Dispersion characteristics.	18
Exercises	22
Chapter 3: Bivariate statistical series	28
Introduction	28
3.1 Relationship and dependence	28
3.2 . Types of relationship between two quantitative characteristics	28
3.3. The correlation diagram	28
3.3.1. The intensity of the relationship	29
3.3.2. The form of the relationship	30
3.3.3. The meaning of the relationship	30
3.4. Calculating the correlation coefficient	31
3.4.1. The Bravais-Pearson linear correlation coefficient	33
3.4.2. Properties and interpretation of r(XY)	34
3.4.3. Limits of the Pearson coefficient	35
3.5. Linear regression	36

Introduction :

Statistics is the study of collecting data, analysing it, processing it, interpreting the results and presenting them in such a way that the data can be understood by everyone. It is a science, a method and a set of techniques.

Data analysis is used to describe the phenomena studied, make predictions and take decisions about them. In this way, statistics is an essential tool for understanding and managing complex phenomena.

The data studied can be of any kind, which makes statistics useful in all disciplinary fields and explains why it is taught in all university courses, from economics to biology, psychology and of course engineering sciences. Statistics involves :

– Gathering data.

– Present and summarise the data.

– To draw conclusions about the population studied and help decision-making.

– In the presence of weather-dependent data, we try to make forecasts.

Definition 1: ***Statistics*** is the set of scientific tools used to collect, order, analyse and draw conclusions from a certain amount of data.

From this definition, we can deduce that statistical problems are studied in the following stages:

- Gathering data.
- Organise and order this data.

- Represent data graphically.
- Analyse the data.
- Explain the results obtained.
- Drawing conclusions.
-

1. Statistical vocabularies :

- Population (statistical universe) is the set on which the statistical study is based.

- A statistical unit (individual) is an element of the population.

- Sample is a part (subset) of the population.

- Characteristic is a characteristic taken on by the individuals in a population.

- Modalities are the different cases likely to be taken by a character.

Examples

1) Study of the different specialities chosen by students in the 2ème S. T. year.
   - Population: Students in 2ième year S. T.
   - Individual: A student.
   - Feature studied : The specialities chosen.
   - Modalities: electronic, mechanical, etc.
2) Study of the weight of newborn babies.
   - Population: Newly born.
   - Individual : A newborn.
   - Characteristic studied : Weight.
   - Modalities 3,400 kg, 2,900 kg, ... etc.
3) Study of the number of workers in a number of small companies.
   - Population: Small businesses.
   - Individual : A small company
   - Characteristic studied : Number of workers.
   - Modalities: 10, 25, 5, . . . etc.

These examples show that there are two types of character:

- A qualitative characteristic whose characteristics cannot be measured.

- Quantitative characteristic whose modalities are measurable. It is often referred to as **statistical variable.**

According to these terms, a statistical variable can be **discrete** if it takes isolated values (in *IN*), or **continuous** if it takes values in an interval.

Definition 2 A **statistical series** is the correspondence between the individuals in a population and the modalities of the characteristic under study.

Ratings

- Characters are denoted by : *X, Y,* . . ., etc.
- The modalities are denoted by lower case letters: $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_k$.

Exercises :

Exercise 1:

Specify which of the characteristics below are qualitative (nominal and ordinal) and which are quantitative (discrete and continuous):

Blood type, sex, baccalaureate grade, number of accidents on a motorway, diameter of the parts made by a machine, region of residence of a student, brand of mobile phone, weight, height, number of separate coins in a pocket, marital status of a person, colour of houses in a district, gross income, number of houses sold per town.

Exercise 2:

The following list is made up of the first names of a group of students, followed in brackets by the number of films each of them has seen in the last month:

Pierre(3),Paul(2),Jacques(2),Ralph(3),Abdel(1),Sidonie(2),Henrie(0),Paulette(1),Farida(2),Laure(2),Kevin(0),Carole(3),Marie-Claire(0),Jeanine(3),Julie(2),Ernest(3),Cindy(3),Vanessa(2),José(1),Aurélien(1).

Determine: The population studied-The variable studied-The nature of the v a r i a b l e - T h e modalities (values) of the variable.

Exercise3 :

Specify which of these statements are true and which are false.

1. A variable is a characteristic that is being studied.

2. The task of descriptive statistics is to collect data.

3. The task of descriptive statistics is to present data in the form o f tables, graphs and statistical indicators.

4. In statistics, variables are classified according to different types.

5. The values of the variables are also called modalities.

6. For a qualitative variable, each statistical individual can have only one modality.

7. In practice, when a discrete quantitative variable takes on a large number of distinct values, it is treated as continuous.

Chapter 2: Univariate statistical series

To study a statistical characteristic or variable, the first step is to gather all the information required. This information must be arranged in **statistical tables** or **graphical** displays, which provide a clearer summary and make it easier to interpret the data.

2.1. Statistical tables.

(a) The case of a discrete variable.

Consider a population of N individuals, described in terms of a discrete statistical variable X with values ($x_1, x_2, \ldots, x_k$). We are interested in knowing, for each value x_i, the number of individuals taking this value. This number is denoted by n_i, $i = 1,\ldots,k$. This gives us the following ***statistical table***:

The values x_i	***or***
x_1 x_2 . . . x_k	n_1 n_2 . . . n_k
Total	N

Definitions :

- The number of individuals n_i in the population for whom the variable X takes the value x_i is called the **number of individuals** or the **absolute frequency of** the value x_i.

- The **relative frequency** f_i of the value x_i of size n_i is given by the formula $f_i = \frac{n_i}{N}$, where N is the **total size of** the population.
- The **percentage** p_i of the value x_i of the number n_i is given by the formula $p_i = f_i \times 100 = \frac{n_i}{N} \times 100$

Remarks : $\sum_{i=1}^{k} n_i = N \; et \; 0 \leq n_i \leq N$ where k is the number of different values.

- $\sum_{i=1}^{k} f_i = 1$ and $\sum_{i=1}^{k} p_i = 100$
- The correspondence between the values of x_i and their numbers is called a **number distribution.**
-

Example 1: A tissue manufacturer tests a new machine and counts the number of defects in 75 10-metre samples. He obtained the following results:

Number of defects x_i	**Number of samples n_i**
0	38
1	15
2	11
3	6
4	3
5	2
Total	75

Cumulative counts.

It may be interesting to read the table and answer questions of this kind:

- What is the number of individuals for whom the variable X takes at least x_j?

- What is the number of individuals for whom the variable X takes at most x_j?

Question 1ére is answered by adding up the numbers from the first value n_1 to n_j $(1 \leq j \leq k)$. The numbers thus obtained are called **cumulative increasing numbers** or **cumulative increasing absolute frequencies,** denoted by $n_{ic}\uparrow$**.**

Question 2ème is answered by adding up the numbers from n_j $(1 \leq j \leq k)$ to the last value n_k. The numbers thus obtained are called **cumulative decreasing numbers** or **cumulative decreasing absolute frequencies**, denoted by $n_{ic}\downarrow$**.**

For example, the 4ème line of the -1- table also reads :

- 6 samples contained 3 defects.
- 70 samples contain a maximum of 3 defects.
- 11 samples contained at least 3 defects.

The table above can be completed as follows:

Number of defects x_i	**Number of samples n_i**	**nic↑**	**nic↓**
0	38	38	75
1	15	53	37
2	11	64	22
3	6	70	11
4	3	73	5
5	2	75	2
Total	75		

Note:

In the same way, we can define

- **Cumulative relative frequencies** (increasing and decreasing).

- **Cumulative percentages** (ascending and descending).

b) The case of a continuous variable.

In the case of a continuous variable, theoretically the values collected are infinite and very close to each other. So, to simplify the study, we construct **classes** (intervals) by dividing **the range of** the statistical series into several intervals.

Definitions

- The **range** of a statistical series is the difference between the largest and smallest values in the series.
- The **classes** are intervals of the form $[a_i, a_{i+1}[$, such that $\overline{\bigcup_{i=1}^{k-1}[a_i, a_{i+1}[} = [a_0, a_k]$; or a_0 and a_k are respectively the smallest and largest value in the series.
- In the class $[a_i, a_{i+1}[$, the values a_i and a_{i+1} are the **bounds** or **limits** of this class.
- The number $x_i = \frac{a_i + a_{i+1}}{2}$ is called the **centre of** the class $[a_i, a_{i+1}[$.
- The number $a_i = a_{i+1} - a_i$ is called the **range** or **amplitude** of the class $[a_i, a_{i+1}[$.
- The headcount *or* the class $[a_i, a_{i+1}[$.corresponds to the number of values belonging to this class.

Note:

The number of classes *k* must not be too small, as this would result in a loss of information, nor too large, as this would make grouping into classes useless. The number of classes that can be constructed is given by the formula .$k = \sqrt{N}$

Example 2 A study of the weight of 80 newborn babies in a maternity hospital gave the following results (Table 2):

The classes	Class centres x_i	Staff numbers *and*	nic↑	nic↓
[2.2 , 2.5[	2.35	2	2	80
[2.5 , 2.8[	2.65	5	7	78
[2.8 ,3.1[	2.95	20	27	73
[3.1 , 3.4[	3.25	19	46	53
[3.4 , 3.7[	3.55	20	66	34
[3.7 , 4.0[	3.85	8	74	14
[4.0 , 4.3[	4.15	4	78	6
[4.3 , 4.6[	4.45	2	80	2
Total		80		

2.2. Graphical representation.

Statistical tables are combined with graphical representations to provide a more comprehensive understanding of the data. Depending on your needs, you can use :

- Headcount and cumulative headcount.
- Relative frequencies and cumulative relative frequencies.

  Graphical representations differ according to the type of variable.

(a) Discrete variable.

- **Bar chart.** This is represented by plotting the values x_i taken by the v. s. on the x-axis and then, starting from each point x_i, drawing a bar whose length is proportional to n_i or f_i (see Figure 2.5).

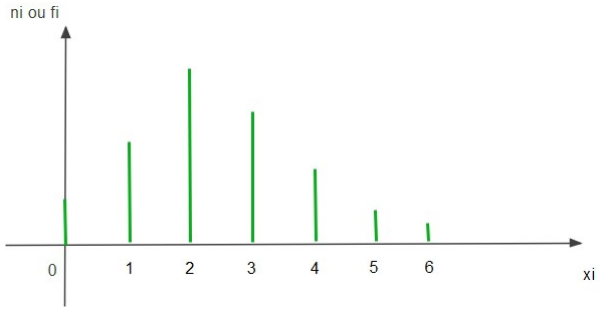


Figure 2.1: Bar chart

- The frequency polygon (or headcount polygon)

This representation is obtained by joining the stick vertices.

- The cumulative headcount curve

We define the function which associates with each value $x \in IR$, the sum of the numbers of all the $x_i < x$ and which we call **the distribution function of the numbers (or** the distribution function of the character X).

The graphical representation of this function is called the "cumulative frequency curve". The cumulative curve is a stepped curve representing relative cumulative frequencies.

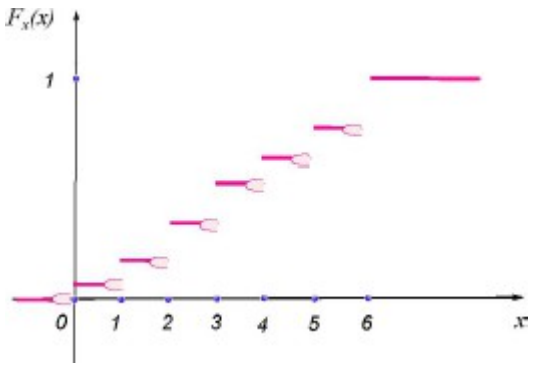


Figure 2.2: Representation of a discrete quantitative variable by the cumulative curve.

(b) Continuous variable.

b1) Case of classes with equal extents

- Histogram.

On the x-axis, the boundaries of the different classes are represented and each class is associated with a rectangle, the base of which is a part of the x-axis between the boundaries of that class, the length of which is proportional to n_i or f_i.

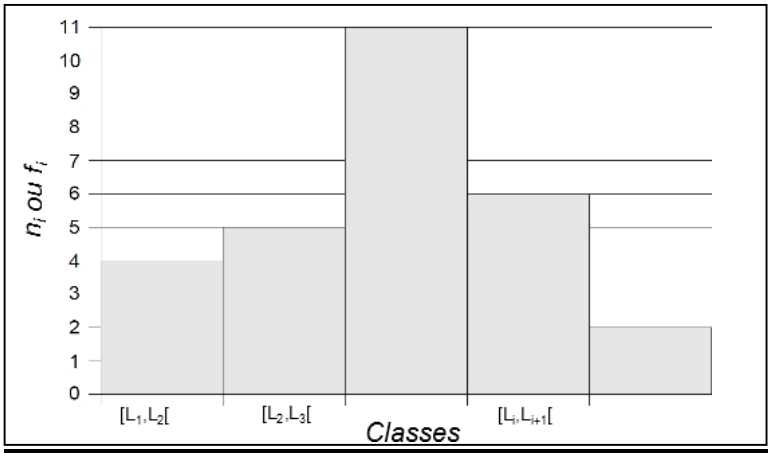


Figure 2.3: Histogram of frequencies or numbers...

- This representation is obtained by joining the points (x_i , n_i) by straight lines. It is completed by 2 extreme classes of the same amplitude.

Comments.

1- The area of all the rectangles is equal to 1 if we represent relative frequencies and n if we represent the workforce.
2- The area between the headcount polygon and the x-axis is equal to the area of the histogram.

b2) Case of classes with different extents.

Let's go back to example 2 and combine classes 1 and 2 into a single class, as well as classes 6, 7 and 8. The statistical table therefore becomes :

The classes	*xi*	*or*	The rectifiable workforce *ñi*
[2.2, 2.8 [	2.5	7	(2+5)/2 = 3.5
[2.8 ,3.1 [	2.95	20	20
[3.1, 3.4 [	3.25	19	19
[3.4, 3.7 [	3.55	20	20
[3.7, 4.6 [	4.15	14	4.66
Total		80	

In this case, the rectangles of the histogram have a length proportional to *ñi*, the rectifiable number in each class.

- The polygon of cumulative increasing numbers (or cumulative increasing relative frequencies)

The polygon of increasing cumulative numbers is obtained by joining, by straight segments, the points having the upper limits of the classes as abscissa and the increasing cumulative numbers (or cumulative relative frequencies) corresponding to the class in question as ordinates. The first point is (a_0, 0).

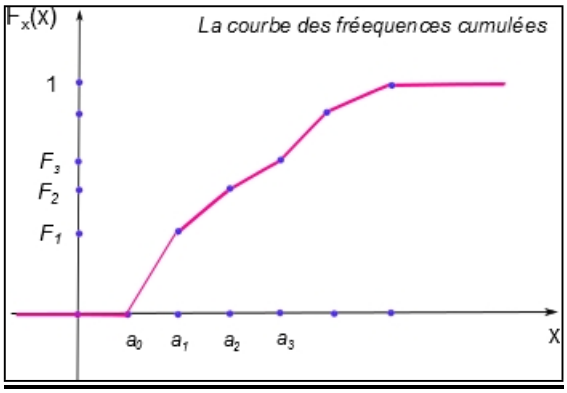


Figure 2.4: The cumulative frequency curve.

- The polygon of cumulative decreasing numbers

The polygon of decreasing cumulative numbers is obtained by joining, by straight segments, the points having as abscissa the lower limits of the classes and as ordinates the decreasing cumulative numbers (or cumulative relative frequencies) corresponding to the class in question. The first point is (a_k,0).

2.3. Characteristic parameters.

A statistical distribution can be summarised by a few values called **characteristic parameters**, classified into 3 categories.

- Characteristics with a central tendency (position).
- Dispersion characteristics.
- Shape characteristics.

2.3.1 Central tendency characteristics

a) The *Mo* mode. The mode is the value of the maximum size variable.

(i) If the s.v. is discrete, the mode represents the value corresponding to the largest number of individuals (or the highest partial frequency).

In example 1, Mo mode = 0.

(ii) If the s.v. is continuous, the mode is the centre of the modal class, i.e. the class corresponding to the largest number of individuals *(or the highest partial frequency)*.

- In example 2, we have 2 modal classes which are [2.8 , 3.1[and [3.4 ,3.7[, so there are 2 modes $_{Mo1}$ = 2.95 and $_{Mo2}$ = 3.55.

- Quantity:

$$M_0 = L_i + \frac{\Delta_1}{\Delta_{1+\Delta_2}} a_i$$

This is called the with mode (see Figure 3.7)

- L_i : the lower limit of the modal class.

– $a_i$modal class step.

– $\Delta_1 = n_0 - n_{,1}\Delta_2 = n_0 - n_2$ or$\Delta_1 = f_0 - f_{,1}\Delta_2 = f_0 - f_{.2}$

– n0 and f0 are the number and frequency associated with the modal class.

– n1 and f1 are the number and frequency of the class preceding the modal class.

– n2 and f2 are the number and frequency of the class following the modal class.

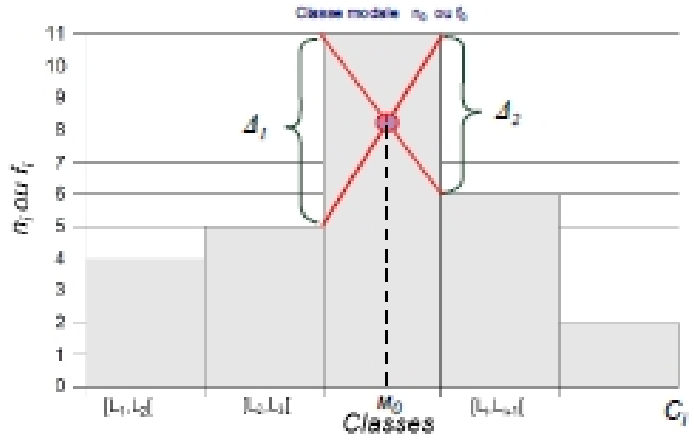


Figure 2.5: Graphical representation or determination of the mode (continuous case).

b) The arithmetic mean. The **mean** of a statistical series (x_1 , x_2 , . . . , x_k) of numbers (n_1 , n_2 , . . .) is the arithmetic mean.

., n_k) is the real value noted by$\bar{x}$ and given by the formula

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i.$$

To find the value of the mean, we add a column to the statistical table in which we calculate the product $n_i x_i$, for all *i*. If the s. v. is continuous, the values x_1 , x_2 , . . . , x_k represent the class centres.

Examples - For example 1 we have, $\bar{x} = \frac{1}{75}\sum_{i=1}^{6} n_i x_i = \frac{77}{75} = 1.027$

Number of defects x_i	Number of samples n_i	$n_i x_i$
0	38	0
1	15	15
2	11	22
3	6	18
4	3	12
5	2	10
Total	75	77

- For example 2 we have $\bar{x} = \frac{1}{80}\sum_{i=1}^{8} n_i x_i = \frac{266}{80} = 3.325.$

The classes	Class centres x_i	Staff numbers *and*	$n_i x_i$
[2.2 , 2.5[	2.35	2	4.7
[2.5 , 2.8[	2.65	5	13.25
[2.8 ,3.1[	2.95	20	59
[3.1 , 3.4[	3.25	19	61.75
[3.4 , 3.7[	3.55	20	71
[3.7 , 4.0[	3.85	8	30.8
[4.0 , 4.3[	4.15	4	16.6
[4.3 , 4.6[	4.45	2	8.9
Total		80	266

- **Properties of the average**

-The average does not change if a given number of values is replaced by their average multiplied by the sum of their numbers.

- The average preserves changes in the axis and the origin

$$X(x_i, n_i) \to Y(y_i = ax_i + b, n_i)$$
$$\bar{x} \quad \mapsto \bar{y} = a\bar{x} + b$$

c) The median *Me* is the value that divides the population into two parts of equal size. If the statistical series is ordered, there are as many observations ranked before as there are ranked after the value of the median.

(i) Discrete variable. Let $x_1, x_2, \ldots, x_N$ be an ordered series of s.v. values.

- If *N* is odd, i.e. $N = 2p + 1$, then the median is given by. $Me = x_{p+1}$

- If *N* is odd, i.e. $N = 2p$, then the median is given by $Me = \frac{x_{p+1} + x_p}{2}$

Examples:

1) Let be the results obtained by a student in the statistics module

10 9 12 10 13 14 18 13 15

The ordered series is: **9 10 10 12 13 13 14 15 15**

4 values **4 values**

$n = 9 = 2 \times 4 + 1$ then the median $Me = x_{k+1} = x_5 =$ **13 = *Me***.

2) We keep the same series and add the value **11**, so the ordered series becomes :

9 10 10 11 12 13 13 14 15 15

5 values5 values

$n = 10 = 2 \times 5$, then the median $Me = (x_k + x_{k+1})/2 = (x_5 + x_6)/2 =$ **$12.5 = Me$**.

3) In example 1, we have $n = 75 = 2 \times 37 + 1$, so $Me = (x_k + x_{k+1})/2$.

According to table -1- we have : $Me = (x_{34} + x_{34+1})/2 = Me = (0 + 0)/2 =$ **$0 = Me$**.

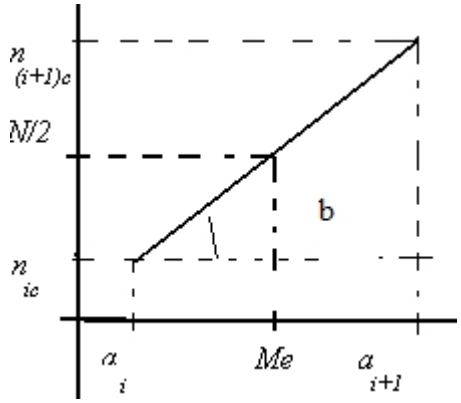
(ii) Continuous variable.

By reading the table, we can determine the class at which half the numbers are reached. The median therefore belongs to this class. This class is **called the median class**. The exact value of the median is calculated using **linear interpolation**. If $[a_i , a_{i+1}[$ is the median class, then we have,

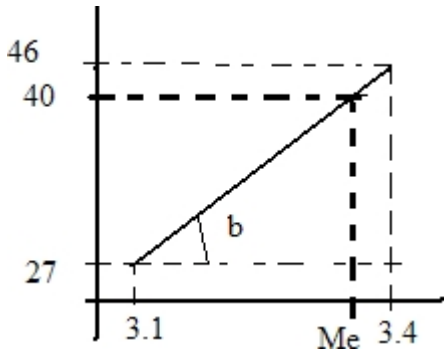
$$tgb = \frac{\sum_{j=1}^{i} n_j - \sum_{j=1}^{i-1} n_j}{a_{i+1} - a_i} = \frac{\frac{N}{2} - \sum_{j=1}^{i-1} n_j}{Me - a_i}$$, where $n_{ic} = \sum_{j=1}^{i-1} n_j$ and $n_{(i+1)} = \sum_{j=1}^{i} n_j$

So

$$Me = a_i + \frac{\frac{N}{2} - \sum_{j=1}^{i-1} n_j}{\sum_{j=1}^{i} n_j - \sum_{j=1}^{i-1} n_j}(a_{i+1} - a_i).$$



Returning to example 2, since N/2 = 80/2 = 40, we deduce from table -2- that $Me \in [3.1, 3.4[$, so using linear interpolation we obtain :



$$tg\alpha = \frac{46-27}{3.4-3.1} = \frac{40-27}{Me-3.1} \Leftrightarrow Me = 3.1 + \frac{46-27}{3.4-3.1}(3.4-3.1) = \underline{\quad 3.3053kg = Me \quad}$$

2.3.2. Dispersion characteristics.

Two statistical series can have the same mean, the same median, the same mode and yet be very different in the sense that the observations made can be more or less "***dispersed***" in relation to the same central value.

This phenomenon is highlighted by calculating numbers known as **dispersion characteristics.**

(i) The range of a distribution is the value given by .$E = x_{max} - x_{min}$

In the case of a continuous variable, the valuesx_{max}etx_{min} are respectively the centres of the last and first class.

(ii) Standard deviation - Variance.

The **variance** of a statistical series ($x_1, x_2, \ldots, x_k$) of numbers ($n_1, n_2, \ldots, n_k$) is the real value denoted by V_X or *var* (*X*) is given by

$$V_X = Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2.$$

When the v.s. is continuous, the x_i represent the class centres.

The **standard deviation** of the distribution is the positive square root of the variance, i.e.

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$$

.

The smaller the standard deviation, the more the distribution is clustered around the mean.

Note. Variance is often used in the form

$$V_X = \frac{1}{N}\left(\sum_{i=1}^{k} n_i {x_i}^2\right) - \bar{x}^2$$

To find the value of the mean, we add a column to the statistical table in which we calculate the product $n_i x_i^2$, for all *i*.

Example Returning to example 1, we have

xi	*or*	*ni xi* ^{2}
0	38	0
1	15	15
2	11	44
3	6	54
4	3	48
5	2	50
Total	75	211

So $V_X = \frac{1}{75}\left(\sum_{i=1}^{6} n_i x_i^2\right) - \bar{x}^2 = \frac{211}{75} - \bar{x}^2 = \underline{\quad 1.7593 = V_x \quad}$ and in this case $\sigma_X = 1.3264$.

Properties of variance.

1) The variance is always positive or zero.

2) Change of scale and origin

$$X(x_i, n_i) \to Y(y_i = ax_i + b, n_i)$$
$$V_X \quad \mapsto V_Y = a^2 V_X$$

(iii) Coefficient of variation.

The coefficient of variation C_{vx} is the ratio of the standard deviation to the mean, c - to - d :

$$Cv_X = 100 \times \frac{\sigma_X}{\bar{x}}.$$

It is expressed without units and is given as a percentage.

The coefficient is used to compare the dispersions of statistical series that are not expressed in the same units of measurement or series with very different averages.

(iv) Interquartile range.

Quartiles are the values which divide the ordered statistical series into 4 parts of equal size.

- The first quartile is the number Q_1 such that 25% of the values are less than or equal to Q_1.
- The third quartile is the number Q_3 such that 75% of the values are less than or equal to Q_3.
- The second quartile Q_2 is the median.

Remarks :

- The first quartile Q_1 is the median of the first half of the statistical series.
- The third quartile Q_3 is the median of the second half of the statistical series.
- The method for calculating quartiles is therefore identical to that for calculating the median.

The **interquartile range** is the number *IQR* such that $IQR = Q_3 - Q_1$. It gives the range of the central half of the observations.

Examples:

1) Let be the results obtained by a student in the statistics module

10 9 12 10 13 14 18 13 15

The ordered series is: **9 10 10 12 13 13 14 15 15**

4 values4 values

The 1er half of the series contains 4 ($= 2 \times 2 = 2 \times k$) values, so the median of this part is

$$Q_1 = (x_k + x_{k+1})/2 = (x_2 + x_3)/2 = \mathbf{10 = Q_1}.$$

The 2ème half of the series also contains 4 values, so the median of this half is

$$Q_3 = (x_{4+1+k} + x_{4+1+k+1})/2 = (x_7 + x_8)/2 = (14 + 15)/2 = \mathbf{14.5 = Q_1}$$

2) We keep the same series and add the value **11**, so the ordered series becomes :

9 10 10 11 12 13 13 14 1515 .

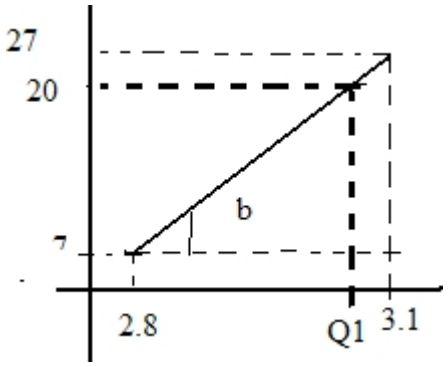
The 1er half of the series contains 5 ($= 2 \times 2 + 1 = 2 \times k + 1$) values, so the median of this part is

$$Q_1 = x_{k+1} = x_3 = \mathbf{10 = Q_1}.$$

The 2ème half of the series also contains 5 values, so the median of this half is

$$Q_3 = x_{5+k+1} = (x_7 + x_8)/2 = 14 \mathbf{= 14 = Q_1}$$

Taking example 2 again, N/4 = 20, then according to table -2- Q_1 is [2.8 , 3.1 [. So using linear interpolation we obtain



$$tgb = \frac{27-7}{3.1-2.8} = \frac{20-7}{Q_1-2.8} \Leftrightarrow Q_1 = 2.8 + \frac{27-7}{3.1-2.8}(3.1-2.8) = \underline{\quad 2.995kg = Q_1 \quad}.$$

Exercises :

Exercise 1

The manager of a shop selling everyday consumer goods recorded the number of items sold per day for a particular item that appears to be very popular. His report covered sales in March and April, corresponding to 52 days of sales. The observations were as follows:

7 13 8 10 9 12 10 8 9 10 6 14 7 15 9 11 12 11 12 5 14 11 8 10 14 12 8

5 7 13 12 16 11 9 11 11 12 12 15 14 5 14 9 9 14 13 11 10 11 12 9 15.

1. What type of statistical variable is being studied?

2. Determine the statistical table based on the numbers, frequencies, cumulative

numbers and cumulative frequencies.

3. Draw the bar chart for the variable X.

4. Let Fx be the distribution function. Determine Fx.

5. Calculate the Mo mode and the arithmetic mean x.

6. Determine the value of the median Me from the table and then from the graph.

7. Calculate the variance and standard deviation.

Exercise 2

A residential area comprises 99 housing units with an average rental value of 10,000 Da. Two new housing units have been built in the neighbourhood: one has a rental value of 7,000 Da and the other, a luxury villa, has a rental value of 114,000 Da.

- What is the new average rental value for the district?

Exercise 3

- A machine cuts 12 cm bars. Unfortunately, the machine is not properly adjusted and the lengths vary around the expected value. A study of 185 bars gave the following results:

Lengths in cm	11.5	11.6	11.7	11.8	11.9	12.0	12.1	12.2	12.3
Workforce	3	15	16	16	18	20	25	25	28

1. What type of statistical variable is being studied?

2. Determine the statistical table (ni, Ni cumulative increasing and decreasing, fi, Fi increasing and decreasing.

3. Draw the bar chart and cumulative frequency curve for the statistical variable.

4. Calculate the mean and standard deviation.

Exercise 4:

A sample of 100 tubes was taken from a plastic tube manufacturer and their diameter measured in decimetres. We divided the values into classes of amplitude a=0.15 :

Classes	[1.94,...[	[...,...[	[...,...[	[...,...[	[...,...[	[...,...[	[...,...[	[...,...[
or	3	9	18	29	25	6	6	4

1- Fill in the table above.
2- Calculate the quartiles (Q1, Q2 and Q3)
3- Calculate position and dispersion characteristics
4- Draw a histogram of the statistical variable.
5- What is the proportion (percentage) of tubes with a diameter greater than 2.69 dm.
6- What is the proportion of tubes whose diameter does not exceed 2.39dm (The distribution function F (2.39)).

Exercise 5:

A study of household budgets for summer holidays produced the following results:

Budget X	[800,1000]	[1000,1400]	[1400,1600]	[1600, β]	[β,2400]	[2400, α]
F.C.C (F_i)	0.08	0.18	0.34	0.64	0.73	1
Frequency f_i						

Some data are missing. Calculate the missing bound α knowing that the range of the series is equal to 3200.

– Calculate the frequencies in the table.

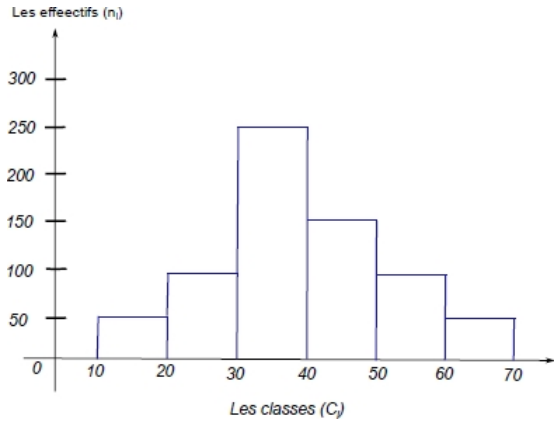
– Calculate the missing bound β in the following two cases:

1. The average budget is equal to 1995.

2. The median budget is 1920.

Exercise 6:

- In a bus station, the waiting time for passengers is measured in minutes. Here is the histogram of the absolute frequencies (Numbers) of this variable.



1. Determine the statistical variable X and its type and population.
2. Determine the number of passengers.
3. From the graph, determine the statistical table.
4. Draw the curve for the cumulative function.
5. Determine the mode graphically and explain what this value represents in relation to our study.
6. Calculate the median from the graph of the cumulative function.
7. Calculate the mean and standard deviation.

Exercise 7:

A taxi company is interested in the mileage of its vehicles. To this end, it recorded the mileage of 50 of its taxis for a morning's work.

Classes (in Km)	[10,20[	[20,30[	[30,40[	[40,60[	[60,90[	[90,130[
Number of taxis	7	10	20	6	3	4

1- Draw the histogram of this distribution.

2- Give the modal class, modal value (and median), mean and standard deviation of the distribution.

3- Now group the data into classes (of the same amplitude) [10;40[, [40;70[, [70;100[and [100;130[. Draw the histogram. Recalculate the parameters from question 2 and compare the results.

Exercise 8:

The marks (variable X) obtained by a class of 5 eme pupils in a French test give the following table:

x_i	n_i	$n_i \times x_i$	$n_i \times (x_i)^2$	N_i *(eff. cum.)*
4	*2*			
5	*3*			
6	*5*			
7	*3*			
8	*2*			
9	*2*			
10	*4*			
11	*4*			
12	*3*			
14	*2*			
TOTAL			**2432**	

1) Specify the variable studied and its type.

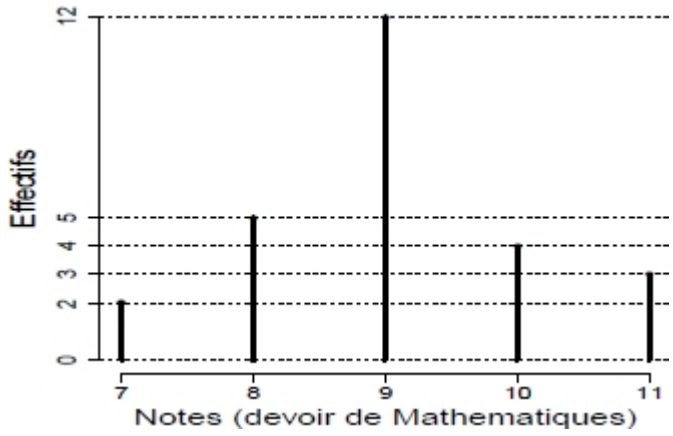
2) Complete the table above.

3) Draw a bar chart representing the distribution of X.

4) Calculate the mean and variance of X.

5) Determine the value of the modality that separates the sample into 2 sub-samples of the same size.

6) The figure below shows the distribution of marks obtained by the same class for a mathematics test (variable Y):



6.a) From the representations of the distributions of X and Y, without doing any calculations, which variable do you think has the smallest variance? For which of the 2 variables is the mean the most representative?

6.b) Deduce the value of the median from the graph

6.c) From the figure showing the distribution of Y, check by calculation that the mean of Y = 9.04.

Chapter 3: Bivariate statistical series

INTRODUCTION

3.1. Relationships and dependency

Let there be two quantitative characteristics X and Y, describing the same set of units. We say that there exists a relationship between X and Y if the allocation of the terms of X and Y is not random, i.e. if the values of X depend on the values of Y or if the values of Y depend on the values of X. Say that Y depends on X means that knowledge of the values of X predicts, to some extent, the values of Y. In other words, if Y depends on X, we can find a function f such that Y=f (X)

Example: there is a relationship between temperature and altitude. The dependence between temperature and altitude is expressed by the relationship :

Tz = -0.6*Z + T0

With

Tz: temperature at altitude z

Z: altitude in hundreds of metres T0:

temperature at sea level.

3.2. Types of relationship between two quantitative characteristics

Prior to any correlation measurement using appropriate coefficients, it is necessary to define the form of any relationship between two characteristics using an appropriate graphical representation. Depending on the form of the relationship observed, the same assumptions will not be made and the same measurement tools will not be used.

3.3. The correlation diagram

To find out if there is a relationship between two characteristics, we draw up a correlation diagram, i.e. a diagram crossing the modalities of *X* and *Y*. Each element i is represented by a point with coordinates (x_i, Y_i). All the points together form a scatterplot, the shape of which can be used to characterise the relationship using three criteria:

- intensity of the relationship

- form of relationship

- sense of relationship

3.3.1. The intensity of the relationship

A relationship is strong if units with neighbouring values on X also have neighbouring values on Y, i.e. if we have the following relationship

x_i close to x_j => y_i close to y_j

=> the point cloud then takes the form of a line or curve whose points are not far apart.

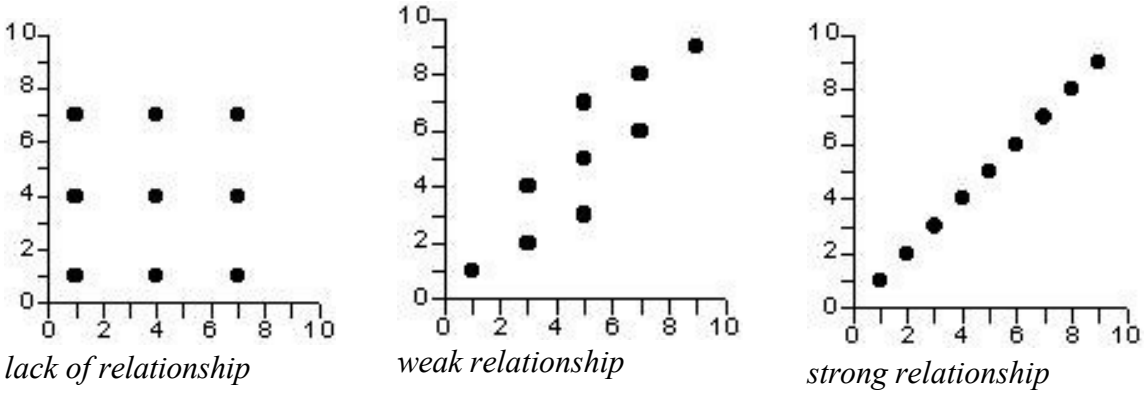


Figure 3.1: Graphical representation of the correlation diagram (Scatter plot).

A relationship is weak if units with neighbouring values in X can have distant values in Y, i.e. if two values close to X can correspond to two very different values in Y.

=> the point cloud does not have the shape of a line or curve, or only very roughly.

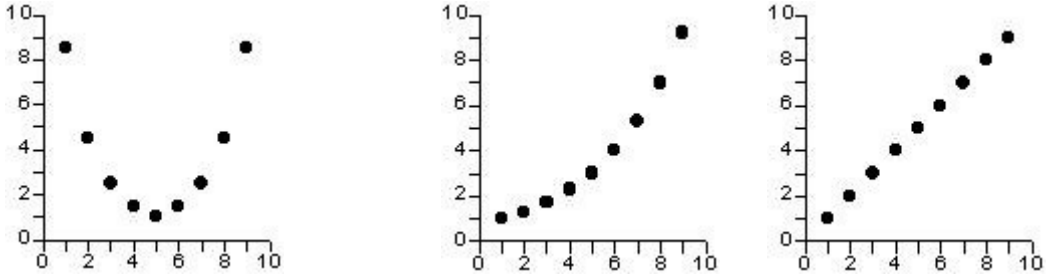
A relationship is null if the values of X do not predict the values of Y

=> the point cloud is shaped like a square, a circle, a "potato" with no real guidelines.

3.3.2. The form of the relationship

A relationship is linear if we can find a relationship between X and Y of the form Y=aX+b, i.e. if the point cloud can be fitted correctly to a straight line.

A relationship is non-linear if the relationship between X and Y is not of the form Y=aX+b, but of a different type (parabola, hyperbola, sinusoid, etc). The point cloud then has a complex shape with curves. A non-linear relationship is **monotonic** if it is strictly increasing or strictly decreasing, i.e. if it has no minima or maxima. All linear relationships are monotonic.



non-linear and non-monotonic relationship

non-linear and monotonic relationship

linear relationship (always monotonic)

Figure 3.2: Scatter plot showing the form of the relationship between X and Y

3.3.3. The meaning of relationships

A monotonic relationship (linear or non-linear) is positive if both characteristics vary in the same direction, i.e. if we generally observe that :

$$x_i > x_j \Rightarrow y_i > y_j$$

- strong values of X generally correspond to strong values of Y.
- the average values of X generally correspond to the average values of Y.
- low values of X generally correspond to low values of Y.

A monotonic relationship is negative if the two characteristics vary in opposite directions, i.e. if we generally observe that :

$x_i > x_j \Rightarrow y_i < y_j$

- high values of X generally correspond to low values of Y
- the average values of X generally correspond to the average values of of Y
- low values of X generally correspond to high values of Y

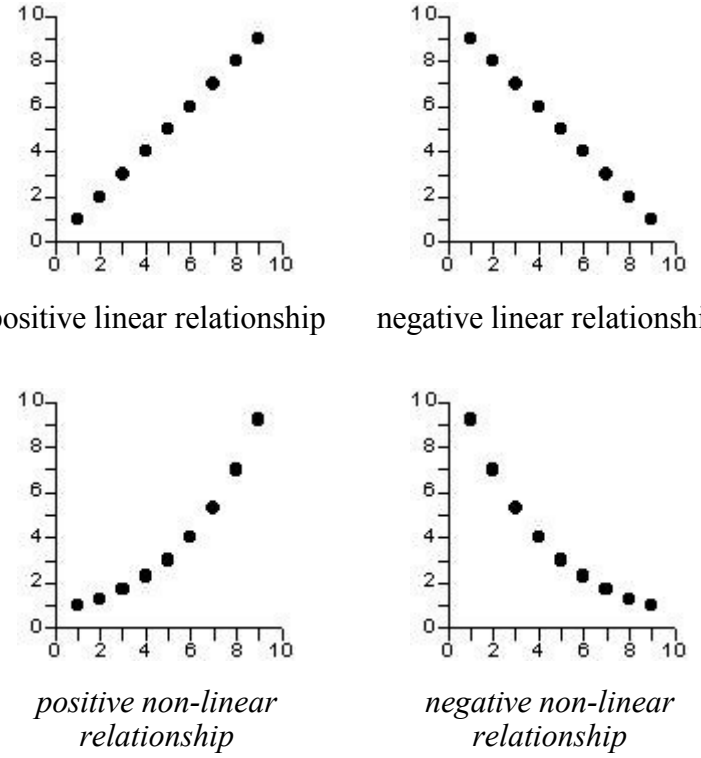


Figure 3.3: Scatter plot showing direction of relationship

3.4. Calculating the correlation coefficient

Correlation coefficients provide a summary measure of the intensity of the relationship between two characteristics and its meaning when the relationship is monotonic. Pearson's correlation coefficient is used to analyse linear relationships and Spearman's correlation coefficient is used to analyse monotonic non-linear relationships. There are other coefficients for non-linear and non-monotonic relationships, but they will not be studied in this course. To illustrate the use of these coefficients, we will start with the (fictitious) example of a psycho-sociobiology study to examine whether there is a relationship

between the size of children's feet and their intelligence. Using a sample of 10 children (labelled A, B, ...J) we will examine whether or not there is a linear correlation between their shoe size (X) and their IQ (Y). The data from the analysis are shown in table 1 below.

child	X	Y
A	31	50
B	31	55
C	32	52
D	33	56
E	33	63
F	34	65
G	35	69
H	36	90
I	37	110
J	38	150
average	34	76
standard deviation	2.4	32

Table 1: Shoe size (X) and intelligence quotient (Y) of 10 school-age children (fictitious data)

We propose to examine whether there is a relationship between spelling ability and foot size: there are four possible answers:

- The greater the size of the feet, the greater the ability to spell (POSITIVE RELATIONSHIP).
- The larger the feet, the lower the ability to spell (NEGATIVE RELATION).
- Foot size is linked to IQ by a complex relationship comprising at least one maximum and one minimum (NON-MONOTONE RELATIONSHIP).
- Foot size is not related to spelling ability (NULL RELATION)

3.4.1. Bravais-Pearson linear correlation coefficient

This coefficient is used to detect the presence or absence of a linear relationship between two continuous quantitative characteristics. To calculate this coefficient, first calculate the covariance. The covariance is the average of the product of the deviations from the mean.

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right) . \left(Y_i - \overline{Y}\right)$$

ou

$$Cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{N} X_i \ . \ Y_i\right) - \left(\overline{X} \ . \ \overline{Y}\right)$$

The linear correlation coefficient of two characteristics X and Y is equal to the covariance of X and Y divided by the product of the standard deviations of X and Y.

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X . \sigma_Y}$$

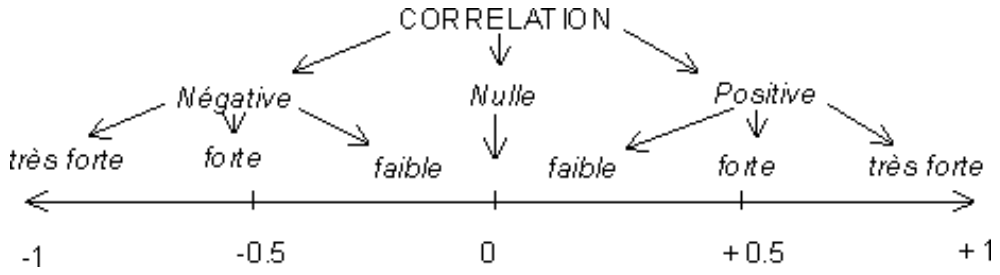
Note: when two characteristics are standardised, their correlation coefficient is equal to their covariance since their standard deviations are equal to 1.

3.4.2. Properties and interpretation of r(XY)

It can be shown that this coefficient varies between -1 and +1. Its interpretation is as follows:

- if *r* is close to 0, there is no linear relationship between X and Y
- if *r* is close to -1, there is a strong negative linear relationship between X and Y
- if *r* is close to 1, there is a strong positive linear relationship between X and Y

The **sign** of *r* therefore indicates the direction of the relationship, while the absolute value of *r* indicates the **strength of** the relationship, i.e. the ability to predict the values of Y as a function of those of X.



Example: Calculation of the linear correlation between foot size and intelligence in 10 school-age children (table 1).

child (i)	Xi	Yi	(Xi -mX)	(Yi -mY)	(Xi-mX)(Yi-mY)
A	31	50	-3	-26	78
B	31	55	-3	-21	63
C	32	52	-2	-24	48
D	33	56	-1	-20	20
E	33	63	-1	-13	13

F	34	65	0	-11	0
G	35	69	1	-7	-7
H	36	90	2	14	28
I	37	110	3	34	102
J	38	150	4	74	296
average	34	76	0	0	64.1
standard deviation	2.4	32			

Table 2: Example of calculation of the Bravais-Pearson correlation coefficient

Since the covariance of X and Y is 64.1, we obtain the correlation coefficient of X and Y by dividing the covariance by the product of the standard deviation of X and the standard deviation of Y :

*r(X,Y) = 64.1 / (2.4 * 32) = +0.83*

We find a strong positive correlation, which seems to indicate that there is a linear relationship (of the type Y=aX+b) between children's IQ and the size of their feet. However, the correlation coefficient does not tell us (1) whether the relationship observed is significant (the result of chance or not) and (2) whether it corresponds to a cause and effect relationship between the two factors X and Y studied. Furthermore, the significance of the linear correlation does not prejudge the existence of a better fit, which would be non-linear.

3.4.3. Limits of the Pearson coefficient

In principle, Pearson's coefficient can only be used to measure the relationship between two variables X and Y with a Gaussian distribution and no exceptional values. If these conditions are not met (a frequent occurrence), the use of this coefficient can lead to erroneous conclusions about the presence or absence of a relationship.

It should also be noted that the absence of a linear relationship does not mean that there is no relationship between the two characteristics studied.

3.5. Linear regression

Introduction

In the particular case where we have been able to demonstrate the existence of a significant linear relationship between two continuous quantitative characteristics X and Y, we can try to formalise the average relationship between these two variables using one of the following three equations:

(1) a.X + b.Y + c = 0: equation of the mean line linking the X and Y characters
(2) Y = a.X + b: regression line for Y as a function of X
(3) X = a.Y + b: regression line for X as a function of Y

The three equations proposed above correspond to three different lines, three different summaries of the point cloud (X,Y). The difference between the three lines comes from the fact that the three equations correspond to three different objectives:

(1) **The mean line** is a summary of the relationship between X and Y which does not introduce any particular hypothesis about the direction of the causal dependence that may exist between the two variables. The aim is therefore to draw the line that is closest to all the points, i.e. the residuals defined by the perpendicular of each point to the mean line (shortest path).

(2) **The regression line of Y as a function of X** introduces the hypothesis that the values of Y depend on those of X, i.e. postulates that knowledge of the values of X makes it possible to predict the values of Y. It is therefore a forecasting model and the objective is to minimise the forecasting error, i.e. the distance between the observed values Yi and the values Y*i estimated by the relationship Y*=aX+b. The residuals will therefore be the distance to the right of the Oy axis.

(3) **The regression line for X as a function of Y** introduces the reverse hypothesis that the values of X depend on those of Y, i.e. it postulates that knowledge of the values of Y makes it possible to predict the values of X. The aim this time is to minimise the forecast error on X, i.e. the distance between the observed Xi values and the X*i values estimated by the relationship X*=aY+b. The residuals

will therefore be the distance to the line in relation to the Ox axis and no longer in relation to the Oy axis as in the previous case.

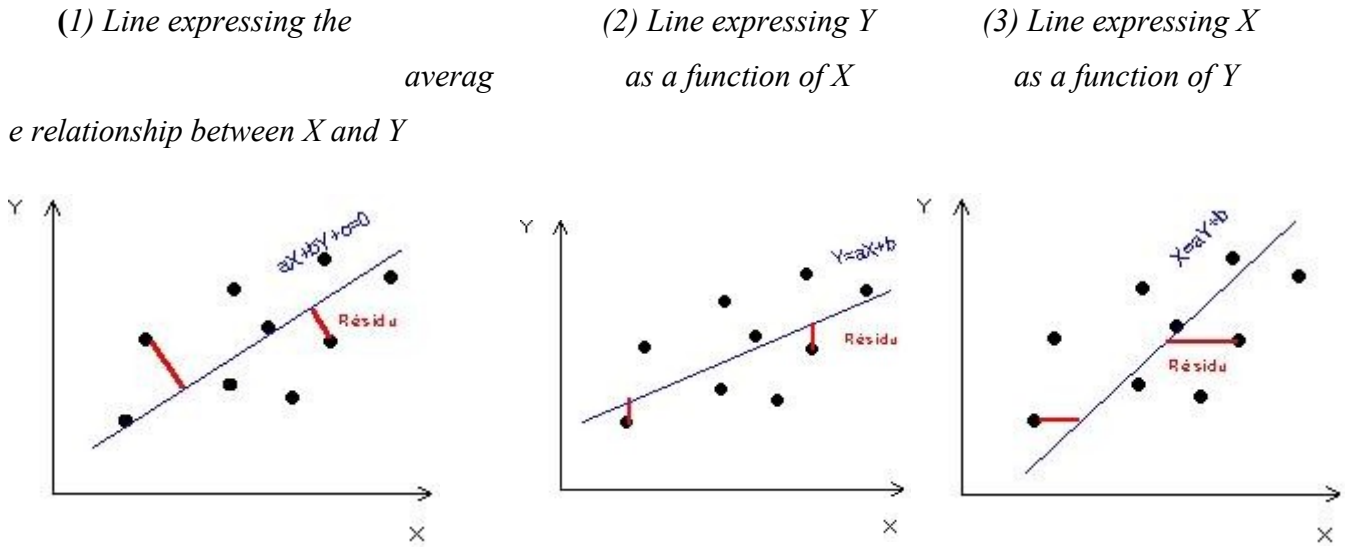


Figure 4: Three different ways of summarising a point cloud

As can be seen in Figure 4, the linear regression lines obtained will differ according to the assumption made about the relationship between X and Y and the presence or absence of dependency between the two characteristics. It is therefore important to always specify the assumption being made before calculating a regression line.

For the purposes of this chapter, we will confine ourselves to the last 2 cases, i.e. situations where we are seeking to express not the relationship between the two characteristics X and Y but the dependence of one characteristic on another (X as a function of Y or Y as a function of X). It is therefore predictive modelling, since it is assumed that knowledge of one of the variables (called the independent variable) enables the value of the other variable (called the **dependent variable**) to be estimated.

3.5.1. Calculating the regression line Y=aX+b

To make things clearer, we will start with a simple and very classic example, that of the relationship between altitude (X) and temperature (Y) within a region of sufficiently small size that we can neglect the macroscopic temperature variation factors (distance from the sea, latitude, etc.). The data presented in Figure 5 are imaginary, but they could correspond to the situation of a north-south alpine valley for which temperatures were recorded at midday at eight stations situated at different altitudes and located on each side of the valley.

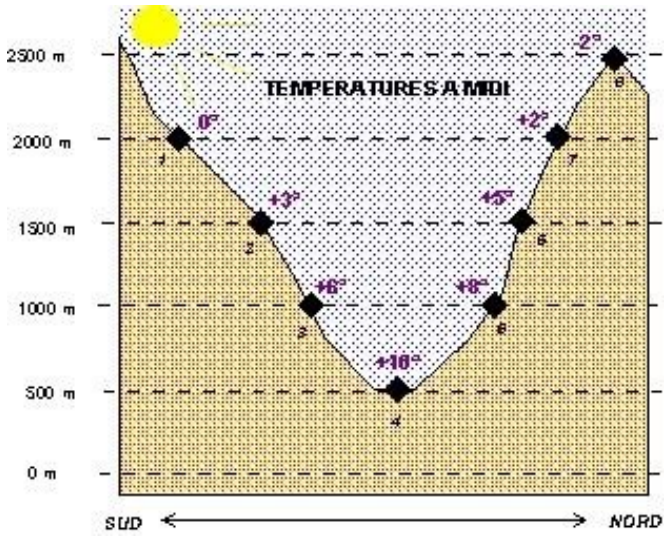


Figure 5: Temperature and altitude at 8 stations in an Alpine valley (imaginary data)

The altitude and temperature data for the 8 stations can be compiled in a table (Table 3), from which the characteristic parameters of each variable (mean and standard deviation) and their covariance can be calculated.

i	(Xi)	(Yi)	(Xi-mX)	(Yi-mY)	(Xi-mX)(Yi-mY)
1	2000	0	500	-4	-2000
2	1500	3	0	-1	0

3	1000	6	-500	2	-1000
4	500	10	-1000	6	-6000
5	1000	8	-500	4	-2000
6	1500	5	0	1	0
7	2000	2	500	-2	-1000
8	2500	-2	1000	-6	-6000
average	mX = 1500	mY = 4	0	0	-2250
standard deviation	612	3.8	-	-	-

Table 3: Characteristic parameters for temperature (Y) and altitude (X) at 8 weather stations in an Alpine valley (imaginary data)

The covariance (-2250) and the two standard deviations (612 for X and 3.8 for Y) indicate a very strong negative linear correlation between the two variables:

$$r(X,Y) = Cov(X,Y) / [\sigma(X) * \sigma (Y)] = -2250 / (612 * 3.8) = -0.97.$$

Even taking into account the small number of observations (8 weather stations, i.e. 7 degrees of freedom), this correlation appears highly significant: there is less than one chance in 1000 that chance could have generated such a strong correlation between the two variables X and Y. The shape of the point cloud crossing the values of X and Y is perfectly linear (Figure 7), which justifies the search for a straight-line fit. What remains to be determined is the **direction of the relationship**, i.e. the hypothesis made about the explanatory variable (independent) and the variable to be explained (dependent). In the example chosen, it seems quite natural to assume that temperature (Y) depends on altitude (X) and not the other way round, so we will look for temperature Y as a function of altitude X. But determining the inverse relationship would not be totally absurd, and we could imagine ... a mountaineer using a thermometer to determine the altitude at which he is (assuming that the

weather conditions are "normal" and there is no thermal inversion on that day).

3.5.1.1 Determining the regression line using the least squares criterion

In the very simple example given here, it is easy to guess the regression line that will give the best fit of temperatures as a function of altitude (Figure 7), but an objective criterion is needed to demonstrate that the proposed solution is indeed the optimum solution, a criterion that can then be applied to more complex point clouds where determining the optimum regression line is less straightforward.

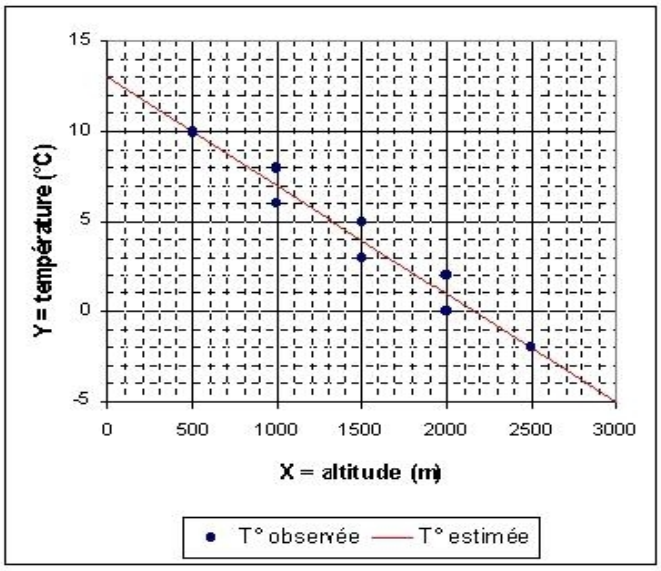


Figure 7: Regression line expressing temperature as a function of altitude for 8 weather stations in an Alpine valley (imaginary data)

We saw in the introduction that when we want to express Y as a function of X, we can assign to each observed value y_i a value estimated by the regression line $y^*_i = aX_i+b$. The estimation error for individual i is therefore equal to the residual εi defined by :

$$\varepsilon_i = (Y_i - Y^*_i) = Y_i - (aX_i+b)$$

Since we want to obtain a global fit that is optimal for all the stations, we need to define a general criterion that defines how well all the values fit the proposed line.

(a) The first solution (ERR1) that comes to mind is to **minimise the sum of the residuals**:

$$\mathbf{ERR1 = \Sigma\, \varepsilon_i}$$

However, this criterion is clearly questionable, as positive and negative residuals can offset each other (temperatures under- or over-estimated by the model) and we could obtain an optimal fit (ERR1=0) even though the straight line does not pass through all the points in the cloud.

(b) The second solution (ERR2) obviously consists of **minimising the sum of the absolute values of the residuals:**

$$\mathbf{ERR2 = \Sigma\, |\varepsilon_i|}$$

This is a good criterion, but it has the disadvantage of not having an analytical solution and requiring an iterative search on all the lines in the plane.

(c) The third solution (ERR3), which is most often used in statistics, is called **the least squares criterion** and consists of **minimising the sum of the squares of the residuals:**

$$\mathbf{ERR3 = \Sigma\, (\varepsilon_i)^2}$$

As in the previous case, the criterion is correct because there is no compensation between positive and negative residuals and the ERR3 value only cancels out if all the points in the cloud are aligned along a straight line. But this criterion has the immense advantage of leading to a very simple analytical solution. The equation of the regression line Y=aX+b which minimises the square of the differences between the observed and estimated values is obtained very simply using the following formulae:

The optimal values for fitting the parameters of the Y=aX+b line **for the least squares criterion** are given by the relationships :

$$\mathbf{a = Cov(X,Y) / (\sigma_X)^2}$$

$$\mathbf{b = m(Y) - a.\, m(X)}$$

Applied to the data in Table 3, these equations give the optimum parameters for fitting the temperature versus altitude regression line:

a = -2250 / (612*612) = -0.006 (°C / m)
b = 4 - (-0.006 * 1500) = 13 (°C)

From this we can deduce that the general equation giving temperature as a function of altitude in the example studied is as follows:

Temperature (°C) = -0.006 * altitude (m) + 13

- **Meaning of the parameters of the regression line**

The parameter *a* on the regression line indicates how much the value of *Y* varies on average when the value of *X* increases by one unit. In our example, the value of *a* is equal to -0.006 and indicates that the temperature decreases by an average of 6°C each time the altitude increases by 1000 metres. The parameter *a* therefore corresponds to what climatologists call the *thermal gradient* in a stable atmosphere (no thermal inversion). From a geometric point of view, the value of *a* corresponds to the slope of the regression line in relation to the Ox axis.

Parameter b on the regression line corresponds to the theoretical value of Y when the value of X is equal to 0. In our example, this is the estimated temperature at zero altitude, i.e. what climatologists call the temperature at sea level. From a geometric point of view, the value of b corresponds to the vertical coordinate of the intersection between the regression line Y=aX+b and the Oy axis.

The empirical interpretation of the parameters a and b obviously depends on the nature of the variables X and Y put in relation, but the principles defined above remain valid in any case: *a* is the rate of change of Y as a function of X and *b* is the value of Y for X =0.

Thus, in the case of a time regression of the type $Y(t)=a.t+b$, the parameter *a* corresponds to the average rate of growth (variation in Y per unit of time) and *b* to the value of Y at time t=0.

.

3.6. Contingency table and chi-square test :

Itroduction

We saw in the previous explanations how it was possible to test the existence of a relationship between two continuous quantitative characteristics using correlation coefficients and then to model this relationship using linear regression.

We will now examine the statistical procedures that should be used to test the existence of a **relationship between two discrete characteristics (quantitative or qualitative).** The nature of these characteristics precludes the use of correlation and regression procedures, and different tools need to be used to determine the form of the relationship (contingency table) and its significance (chi-2 test).

3.6.1 Description of a relationship between two discrete characteristics

Consider a set of n individuals denoted 1...n described by two discrete characteristics X and Y. We will note 1...k the different possible modalities of X (k<n) and 1...p the different possible modalities of Y (p<n). If we cross the possible modalities that an individual can take on X and Y simultaneously, we see that there are k*p possible crossings

Id	X	Y
1	X1	Y1
.		
.		
.		
.		
n	Xn	Yn

Example*:* The 36 students in the class of 2019 in the electrical engineering and automation department of the st - univ de relizane faculty are described by a set of variables relating to gender, age and group (there are two groups). We would like to know whether men and women are randomly distributed between the two groups or whether there is a stronger representation of men or women in one of the groups.

Attributes of auto 2019 students.

Code	Group (X)	Sex (Y)
015	1	m
beep	1	m
cms	1	f
dar	1	m
kas	1	m
lll	1	m
ma2	1	m
mik	1	m
phi	1	m
rai	1	m
rom	1	f
squ	1	m

xxx	1	f
yar	1	m
zic	1	m
zor	1	m
zzz	1	m
ab2	2	f
beb	2	f
can	2	m
coy	2	m
eca	2	m
wire	2	f
flu	2	f
fma	2	f
fre	2	m
goo	2	m
ho1	2	f
hug	2	m
joo	2	m

ply	2	f
sni	2	m
yza	2	f
yzc	2	m
zo2	2	f
zou	2	m

The Group variable (X) has two modalities (k=2) and the Sex variable (Y) has two modalities (p=2). There are therefore 4 possible crossovers: male from group 1, male from group 2, female from group 1, female from group 2 . ..

3.6.2. From the elementary table to the contingency table

To determine whether there is a relationship between the two characteristics under study, we construct a contingency table, i.e. a table counting the cross-tabulated modalities of the two characteristics X and Y. This table will therefore have k rows (number of modalities of X) and p columns (number of modalities of Y). This table will therefore have *k* rows (number of states of *X*) and *p* columns (number of states of *Y*). **Margins** will be added where row totals (the number of individuals in each mode of X), column totals (the number of individuals in each mode of Y) and finally the grand total (the number *n* of individuals studied) will be calculated.

The various boxes are abbreviated using a variable *N* with appropriate indices:

- N_{ij}: number of individuals in the cell corresponding to the $i^{ème}$ row and the $j^{ième}$ column of the table, i.e. the number of individuals having as attribute the $i^{ème}$ modality of X and the $j^{ième}$ modality of Y.
- $N_{i.}$ sum of the $i^{ème}$ line, i.e. the number of individuals with the $i^{ème}$ modality of X as an attribute.
- $N_{.j}$: sum of the $j^{ème}$ column, i.e. number of individuals with the $j^{ème}$ modality of Y as an

attribute.

- $N_{..}$: overall sum of the table, i.e. total number of individuals studied

	Y1	.	.	**Yj**	.	.	**Yp**	**Total**
X1	N11	.	.	N1j	.	.	N1p	$N1_{.}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Xi	Ni1	.	.	Nij	.	.	Nip	Ni.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Xk	Nk1	.	.	Nij	.	.	Nkp	$Nk_{.}$
Total	$N_{.1}$	.	.	$N_{.j}$	.	.	$N_{.p}$	N

Example*:* construction of a contingency table cross-referencing the group and gender of students in the auto 2019 class.

Nij	Gender = "f	Gender = "m	Total
Group = "1	3	14	17
Group =" 2"	9	10	19
Total	12	24	36

This contingency table allows us to count all the possible cases of single modalities (one character) or crossed modalities (two characters). We can say that there are 14 male students in group 1 (box N_{12}), that there are 19 students in group 2 (box $N_{2.}$), that there are 24 male students (box $N_{.2}$) and that there are 36 students in all (box N).

3.6.2.1 Analysis of online and column profiles

As the contingency table shows the raw numbers, it cannot be used to compare the proportions of students of a particular type, nor can it be used to directly answer questions such as "Is the proportion of men higher in group 1 than in group 2?) We therefore generally construct two profile tables showing the percentages in rows or columns.

- **The row profile table** is constructed by dividing the number of employees in each cell by the total for the corresponding row:

Construction of line profiles : $N_{ij} \Rightarrow N_{ij} / N_{i.}$

- **The table of profiles in columns** is constructed by dividing the number of employees in each cell by the total for the corresponding column:

Construction of column profiles : $N_{ij} \Rightarrow N_{ij} / N_{.j}$

The interpretation of the two tables is obviously different, since the ratios are not based on the same reference base. When commenting on the results, care must be taken not to confuse the two percentages describing the same cell in a contingency table.

Example: Construction of the row and column profiles of the contingency table cross-referencing the group and gender of auto students in the class of 2019.

Line profiles

N_{ij} / $N_{i.}$	Gender = "f	Gender = "m	Total
Group = "1	18 %	82 %	100 %
Group =" 2"	47 %	53 %	100 %
Total	33 %	67 %	100 %

=> This table shows that the proportion of women in the graduating class as a whole is 33%, but that it is significantly higher in group 2 (47%) than in group 1 (18%).

Column profiles

N_{ij} / $N_{.j}$	Gender = "f	Gender = "m	Total
Group = "1	25 %	58 %	47 %
Group =" 2"	75 %	42 %	53 %
Total	100 %	100 %	100 %

=> This table shows that group 1 accounts for only 47% of the students in the year. However, it accounts for 58% of all men and only 25% of all women in the year.

Note that the same box in the contingency table can always be described in two different ways. If we take box N_{12} , it indicates that the 9 women in group 2 represent 47% of the students in group 2 and 75% of the women in the class of 1996.

3.6.2.2. Calculation of theoretical numbers and deviations from independence

Another way of approaching the study of a contingency table is to compare the observed numbers in each of the cells (N_{ij}) with the theoretical numbers (N_{ij*}) that would be obtained if there were no link.

between the two modalities X and Y, i.e. if each modality were assigned independently.

To reconstitute the theoretical distribution of the k*p cells of the contingency table, we will use the margins of the table, which define the conditional probabilities that an individual will receive a given mode of X or Y.. .

- The probability of an individual receiving mode i of X is equal to Ni. / N..

- The probability of an individual receiving mode j of Y is equal to N.j/N..

- The probability that an individual will simultaneously receive mode i of X and mode j of Y is therefore equal to (Ni.* N.j) / (N.. * N..)

1- The theoretical number of individuals in the Nij cell (N*ij) is obtained by multiplying the probability of an individual receiving this mode by the number of individuals (N..). This gives the following general formula:

- Calculation of theoretical headcount : $\mathbf{N^*ij = (_{Ni.} * N_{.j)} / N_{..}}$

This theoretical headcount is that which would be obtained if there were perfect independence between the assignment of the X and Y modalities. However, there can obviously be deviations between the theoretical distribution and the observed distribution, either because of random fluctuations or because of the existence of a dependency between the two characteristics X and Y. Before testing the significance of this relationship, we can calculate the **deviations from independence** in order to be able to describe **the form of any relationship between the modalities of X and Y**.

Calculation of deviations from independence *:* $\mathbf{Dev_{ij} = (N_{ij} - N^*_{ij})}$

Example: Construction of the theoretical distribution of the contingency table crossing the group and gender of AUTO students in the class of 2019.

- Theoretical profile :

N*ij	Gender = "f	Gender = "m	Total
Group = "1	5.7	11.3	17
Group =" 2"	6.3	12.7	19
Total	12	24	36

=> This table shows, for example, that if students had been assigned to a group regardless of their gender, there should have been 5 or 6 girls in group 1 (theoretical value = 5.7) and not 3 as observed in the actual distribution.

- Departures from independence.

Nij - N*ij	Gender = "f	Gender = "m	Total
Group = "1	-2.7	+2.7	0
Group =" 2"	+2.7	-2.7	0
Total	0	0	0

=> This table shows that women are over-represented in group 2 and men in group 1. Conversely, women are under-represented in group 1 and men under-represented in group 2. All these deviations are in relation to the reference distribution, which is the one that would have been obtained if the groups had been constructed randomly (i.e. without taking into account the gender of the students). Knowing that an empirical distribution can never coincide exactly with a theoretical distribution, the question that arises is whether the differences observed are the effect of chance or whether they reveal a significant correlation between the two variables X and Y (a correlation that we could then try to explain, for example by asking the person who made the groups how they did it).

3.6.3 Chi-2 test

There are a large number of tests that can be used to measure the degree of significance of the relationship between two qualitative characteristics. Some of these tests are adapted to particular situations (contingency tables crossing two variables with 2 modalities each) while others are more general (contingency tables with any number of rows or columns). For the purposes of this lesson, we will confine ourselves to presenting the most frequently used test, which is best suited to most situations: the chi-2 test.

3.6.3.1 Determining the observed Chi-2 and the number of degrees of freedom

The general idea of the Chi-2 test is to quantify the sum of the deviations between the observed and theoretical numbers present within a contingency table using a single quantity (statistic) and then to compare the value of this statistic with its probability of occurrence in the case of a series of random draws, taking into account the size of the table (number of degrees of freedom).

To eliminate the sign of deviations from independence, we calculate a measure of deviation from independence for each cell that is always a positive quantity. This quantity is called the local Chi-2, or Chi-2 of a cell, and is equal to the square of the difference between the observed value and the theoretical value, divided by the theoretical number in the cell. It is therefore a relative deviation which takes into account the fact that a deviation of +3 does not have the same meaning depending on whether it refers to a theoretical population of 5 individuals or 100 individuals.

Calculation of local Chi-2 : $\text{Chi-2}_{ij} = (N_{ij} - N^*_{ij})^2 / N^*_{ij}$

The higher the local Chi-2 of a cell, the more statistically significant the deviation between observed and estimated values (i.e. the more it corresponds to a rare event that would be unlikely to occur if the X and Y variables were independent).

We then summarise the overall amount of deviation present in the table by calculating the Chi-2_{Obs} value, which is the sum of all the local Chi-2s for the k*p cells in the table.

Finally, we determine the number **z** of **degrees of freedom**, which depends on the number of rows and columns in the contingency table and expresses the number of cells that can produce mutually independent deviations. In the case of a contingency table with 2 rows and 2 columns, this number of

degrees

of freedom is equal to 1 since, since the sum of the marginal deviations must be equal to zero, it is sufficient to know the deviation of one cell to find the deviations of all the others by difference. More generally, the number of degrees of freedom is equal to the number of columns minus one multiplied by the number of rows minus one, i.e. :

Determining the number of degrees of freedom : $z = (k-1)*(p-1)$

Example: Determination of the local Chi-2 and global Chi-2 of the contingency table crossing the Group and gender of AUTO students in the class of 2019.

Chi-2ij	Gender = "f	Gender = "m	Total
Group = "1	1.255	0.628	-
Group =" 2"	1.129	0.561	-
Total	-	-	3.567

=> The most significant deviation concerns the under-representation of women in group 1. The total

Chi-2 value of the table (sum of the four local Chi-2s) is 3.567.

The number of degrees of freedom in this table is equal to (2-1) (2-1) i.e. 1 degree of freedom.

3.7. Cramer's V coefficient :

Cramer's V is a statistic used to measure the strength of the association between two categorical variables and takes values between 0 and 1. Values close to 0 indicate low

association between the variables and values close to 1 indicate a strong association between the variables.

The Cramer V statistic is a symmetrical measure, in the sense that it does not matter which variable is placed in the rows and which is placed in the columns.

The Cramer V statistic is calculated using the following formula:

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n * (min(l,c) - 1)}}$$

Hence :

- V varies from 0 to 1

- The khi-2 is calculated according to the following formula :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Où n* es l'effectif théorique c'est-à-dire l'effectif que l'on aurait eu si les variables étaient indépendantes :

$$n^*_{ij} = \frac{n_{i.} n_{.j}}{n}$$

Principle :

1- Calculation of theoretical headcount

2- Calculation of the difference in relation to the theory

3- Overall indicator of difference from independence Example :

Statistical study of blood alcohol levels and their relationship to gender.

We have the following data:

	Pas Saoul	Un peu Saoul	Saoul	Total
Homme	5	17	4	26
Femme	1	7	16	24
Total	**6**	**24**	**20**	**50**

The theoretical headcount is therefore as follows:

(numbers assuming independence: 3.12=(26*6)/50)

	Pas Saoul	Un peu Saoul	Saoul	Total
Homme	3,12	12,48	10,4	26
Femme	2,88	11,52	9,6	24
Total	**6**	**24**	**20**	**50**

We can therefore calculate the distance to this theoretical headcount :

(1,13=(5-3,12)²/3,12)

	Pas Saoul	Un peu Saoul	Saoul	Total
Homme	1,13	1,64	3,94	26
Femme	1,23	1,77	4,27	24
Total	**6**	**24**	**20**	**50**

The result is :

$\chi^2 = 13{,}98$

$$V = \sqrt{\frac{13,98}{50}} = \sqrt{\frac{13,98}{50 * (min(2,3) - 1)}} = 0,53$$

3.8. Mayer line and affine adjustment

- **Definition. Representation :**

✓ A statistical series with two quantitative characters, x_i and y_i, is **a double series** whose values are given by the pairs $(x_i; y_i)$.

✓ This series is represented in an orthogonal frame of reference by points with coordinates $(x_i; y_i)$ which form a point cloud.

Together, these points form a point cloud. This cloud can be elongated, curvilinear or highly dispersed.

***Note**:*

If the values of one of the two characters are measures of time, the series is said to be **chronological.**

- **Average cloud point**

The **mean point G(*x*; *y*)** is the point whose coordinates are the averages of the x_i and y_i values in the series.

x_G = **Error!**; y_G = **Error!**

- **Affine adjustment**

An elongated point cloud representing a double series (***xi; yi***) can be fitted by a straight line called the **affine fitting line.**

- **Affine adjustment method (Mayer method)**

In the case of an elongated scatterplot, and in order to make it easier to study the series, it is possible to replace this scatterplot with a straight line called the affine fitting line.

This line is drawn using Mayer's method.

✓ The cloud is divided into two clouds of equal size according to the increasing values of x_i :

- we calculate the coordinates of the mean points G_1 and G_2 of these two clouds;

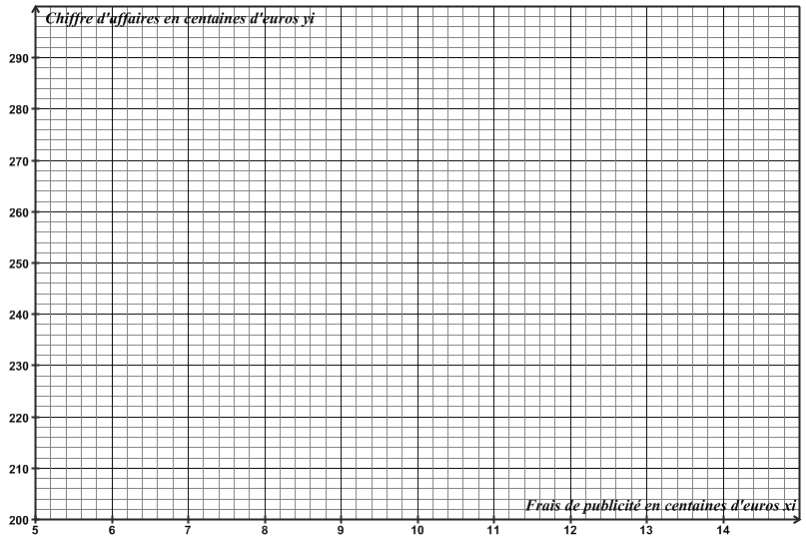
- the equation of the straight line (G_1G_2) is determined. This line is called **the Mayer line.** It passes through the mean point G.

For example:

A shop sales manager is analysing the trend in his sales over the last period. To do this, he records the amount of advertising costs incurred over the same period. He draws up the following table (amounts are expressed in hundreds of euros)

Advertising costs x_i	10	6	6,5	11,5	11	8	7	6,5	11	9
Sales y_i	250	220	228	262	268	244	240	222	259	246

1- Show this double series in the orthogonal reference frame below, by placing the 10 points whose coordinates are the pairs $(x_i; y_i)$.



Possible adjustment methods

1- The manager will look for a link between sales figures and advertising costs: the elongated shape of the scatterplot in the figure above indicates a preferred direction.

It is possible to draw a straight line in this direction, without it deviating greatly from the points in the cloud.

The manager will look for an affine adjustment of this cloud and will be able to determine a future estimate of turnover.

1- To fit a straight line to a set of points, the designer has a choice of methods:

- it can make a judgment adjustment ;

- or draw a straight line through the midpoint of the cloud.

Calculate the mean point of the example series

G : **Error!**

1- To fit the line to the set of points, the designer can also use a more precise method, as follows:

a- Divide the cloud into two groups of points

- the first formed by the 5 points with the smallest abscissas;

- the second group made up of the 5 points with the largest abscissas. To do this, complete the following table

1er group	2e group

Advertising costs x_i	6	6,5			8	9		11		11.5
Sales y_i	220									262

b- Calculate the coordinates of G_1, the mean point of the first group.

$G_1 \left\{ x_1 \text{ ----------}; y_1 = \text{----------} \right.$.

c- Calculate the coordinates of G_2, the midpoint of the second group.

$G_2 \left\{ x_2 = \text{----------}; y_2 = \text{----------} \right.$

d- Place the points G_1 and G_2 in the reference frame and draw the line

(G_1G_2). e - Determine the equation of the line (G_1G_2).

f- Check that the mean point of the cloud G(8.65; 243.9) belongs to the straight line (G_1G_2).

How do I use an affine fit?

Based on the above fine-tuning, the sales manager can estimate the sales he expects to achieve if he incurs advertising costs of €1,300.

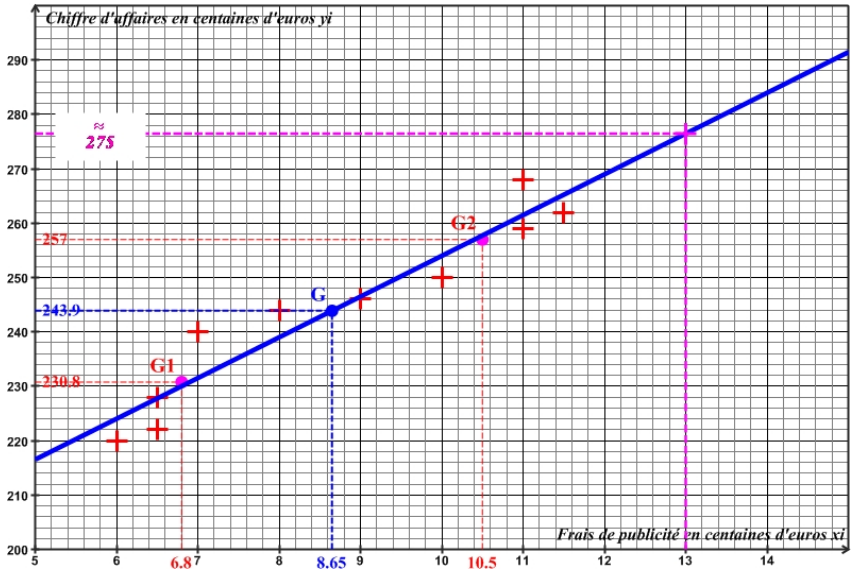
1. Determine the expected sales figure graphically. 2- Calculate the sales figure.

Remarks

The expression "linear fit" is sometimes misused. In fact, the adjustment line does not always pass through the origin of the reference frame;

If the cloud contains an odd number of points, there are two possible splits.

- The graphical representation above is called a **scatterplot**



- The coordinates of G, G1 and G2 are :

- G (8.65, 243.9).

- G1 **Error!**

- $G_2 \begin{cases} x = 10 \\ 5; y = 257 \end{cases}$

- The equation is of the form: y = ax+

b We have : G1 (6.8; 230.8) and G2 (10.5; 257)

hence: a = **Error!** = **Error!** ≈ 7,08

and: $\boldsymbol{b} = y_{G1} - ax_{G1} = 230.8 - 7.08 \times 6.8 = 182.7$

The coordinates of point G_2 can also be used to calculate ***b***.

The equation of the line (GlG2) is: y = 7.08 x+ 182.7 For

$\boldsymbol{x} = 8.65$, we have: $\boldsymbol{y} = 7.08 \times 8.665 + 182.7 = 243.9$

- The coordinates of point G satisfy the equation of the line ($_{GlG2}$). Point G is a point on the line ($_{GlG2}$).

How do I use an affine fit?

a) The graph shows the ordinate of the point on the straight line with abscissa 13 (hundreds of euros).

Sales were €27,500.

b) Using the equation for the line,

we obtain y = 7.08 × 13 + 182.7 = 274.7

The manager can expect sales of around €27,500.

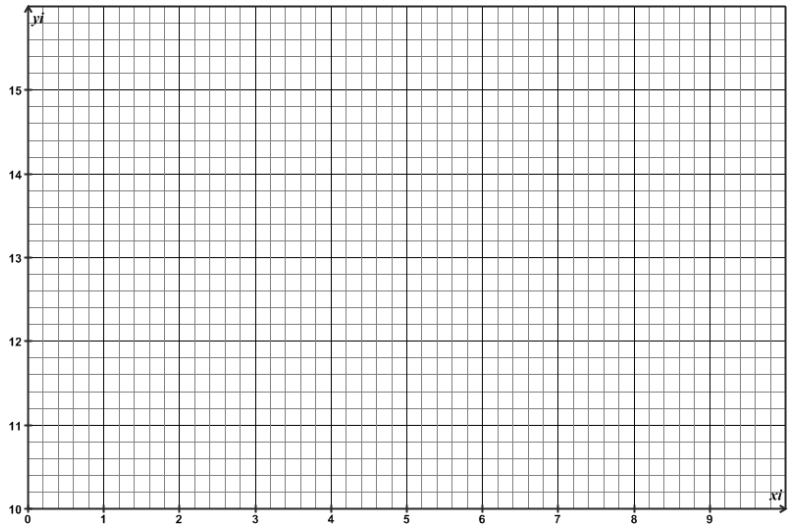
This value is only an estimate: greater precision would be meaningless.

Exercises

Exercise 1 :

Show the point cloud of the following double series in the orthogonal reference system:

xi	1	2,5	3	3,5	4	4	5	5,5	5,5	6	8	9
yi	15	14	13	13,5	13	12,5	12	11,5	12	11,5	11	11



Exercise 2:

Consider the following double series:

xi	1	2,5	3	3,5	4	4	5	5,5	5,5	6	8	9
yi	15	14	13	13,5	13	12,5	12	11,5	12	11,5	11	11

1. Divide the points $(x_i; y_i)$ into two groups: the first with the 6 points with the smallest abscissas, the second with the 6 points with the largest abscissas.
   Calculate the coordinates of the mean points G_1 and G_2

2. Determine the equation of the line (G_1G_2).

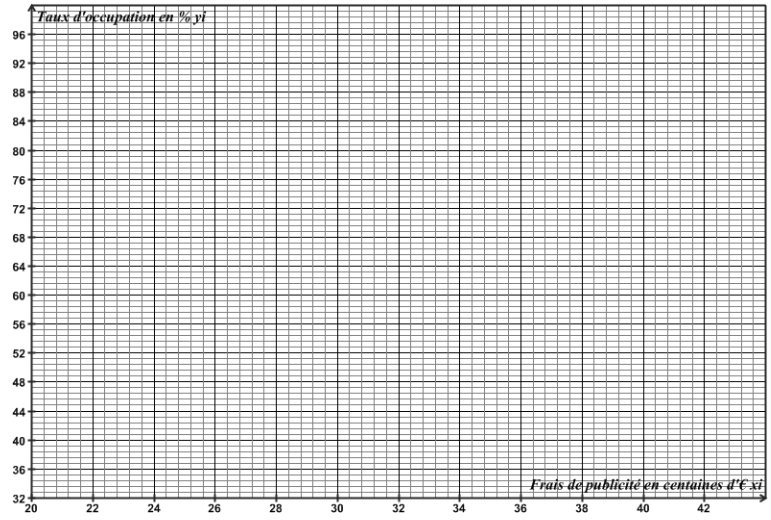
Exercise 3:

In order to guide its investments, a hotel chain carries out analyses of room occupancy rates.

An analysis establishes a link between the occupancy rate, expressed as a %, and the amount of advertising costs (in thousands of euros).

Advertising costs x_i	30	27	32	25	35	22	24	35
Occupancy rate y_i	52	45	67	55	76	48	32	72

1. Show the point cloud $M(x_i; y_i)$ in the orthogonal diagram below.



2- Determine the coordinates of the mean point ***G*** of this cloud, rounded to the nearest whole number. Place this point in the previous reference frame.

3- The straight line passing through the mean point ***G*** and the point ***P*** with coordinates (35; 72) is chosen as the adjustment line for this scatter plot.

a. Locate point ***P*** and draw this line in the previous reference frame.

b- Determine graphically the amount of advertising costs required to achieve an 80% occupancy rate. The construction lines should be shown on the diagram.

Exercise 4:

Of the 6 dictations carried out by the CM2 pupils, 3 took place in a noisy environment and 3 in a quiet environment. The table below shows the number of accumulated errors (grouped into three categories) according to the type of environment:

X \ Y	*bruyant*	*silencieux*	Marge de X
moins de 25	*37*		**241**
de 25 à 30		*86*	**288**
plus de 30		*21*	
Marge de Y			N=691

1. Specify the variables X and Y studied in this study and their type, then complete the table above.

2. Determine the distribution of X conditional on Y. Graph this distribution. What can you say about the relationship between X and Y for this sample?

3. Calculate the Cramer index. Comment on the result.

Exercise 5:

The following table gives the braking distance for a vehicle travelling on a dry road as a function of its speed.

Vitesse en km /h (xi)	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**110**
Distance en m (yi)	**8**	**12**	**18**	**24**	**33**	**41**	**48**	**58**

a) Represent this statistical series by a scatter plot. Calculate the average speed and average distance.

b) Using the method of least squares, determine the equation of the straight line representing distance as a function of speed.

c) Calculate the parameter that allows us to measure the strength of the linear relationship between the two variables. Interpret your result.

d) What is the estimated braking distance for a vehicle travelling at 120km/h?

Exercise 7:

Example of an affine fit

In this activity, all the numerical results will be given as their approximate decimal value to the nearest 10-3, obtained directly with a calculator.

The table below shows the number of passengers carried annually on an airline over the last five years:

Année	*Rang x_i de l'année*	*Nombre p_i de passagers*
1992	1	7550
1993	2	9235
1994	3	10741
1995	4	12837
1996	5	15655

1) Let $y = \ln p$ where ln denotes the natural logarithm.

a) Complete the following table after reproducing it:

b) Show the point cloud Mi(*xi, yi*) in an orthogonal reference frame. Is it possible to make an affine fit of this cloud?

2) a) Using the method of least squares, determine the equation of the regression line D from *y* to *x.*

b) Determine the correlation coefficient r between the two variables *y* and *x*. Does the result confirm the observation made in l) b)*?*

c) From a), derive an expression for *p* as a function of *x.*

d) Assuming that the trend observed continues in subsequent years, use the relationship obtained in c) to estimate the number of passengers carried in 1998.

Exercise 8:

Example of the use of smoothing by the moving average method before affine adjustment. The quarterly sales figures, for the last twelve quarters, of a company manufacturing electronic equipment are given in the following table

Rang du trimestre x_i	Chiffre d'affaires (en MF) : y_i
1	300
2	450
3	130
4	200
5	280
6	410
7	200
8	250
9	320
10	500
11	210
12	250

1) Graphically represent the point cloud Mi(*xi, yi*) in a plane with an orthogonal datum. orthogonal reference point.

The units will be 1 cm on the x-axis and 2 cm for 100 MF on the y-axis.

2) The cloud obtained in 1) shows more or less regular deviations on either side of a fitting line drawn by guesswork. In this case, the cloud is often **smoothed** by replacing the points with mean points.

Rang du trimestre x_i	4	5	…
Chiffre d'affaires (en MF) y_i	270	265	…

b) Show the point cloud *Ni*(*xi, zi*) on the figure in 1). The irregularities in the point cloud Mi have been reduced. Use two different conventions to represent the points *Mi and Ni*.

3) Consider the sales series obtained after smoothing in question 2' a).

a) Determine the linear correlation coefficient to the nearest 10-2 for the dual statistical series of variables *x* and *z*.

b) Using the method of least squares, determine the equation of the regression line *D* from z to *x*. Give an equation of the form $z = ax + b$ *where a* is a value approximated to within 10-2 and *b* is a value approximated to within one unit.

c) It is assumed that the trend observed in quarters 4 to 12 continues. Give the forecast sales for quarters 13 and 14.

Exercise 9: *Quality control*

A machine tool automatically produces cylindrical parts. Initially set for a diameter of 8 mm, it goes wrong during use. The aim of the exercise is to determine the number of parts that can be produced before their diameter reaches 8.1 mm. In order to monitor production and make any necessary adjustments, the diameter of the last part in each series of ten parts produced is measured. The results are as follows:

Numéro x_i de la pièce	10	20	30	40	50	60	70	80	90	100
Diamètre y_i de la pièce (en mm)	8,00	8,00	8,01	8,01	8,02	8,03	8,03	8,04	8,05	8,06

1) Represent the scatterplot Mi(*xi, yi*) associated with the above statistical series in the plane with an orthogonal reference point. Take the origin as the point with coordinates (0, 8), the unit as 1 cm per ten pieces on the abscissa and 1 cm per 0.01 mm on the ordinate.

2) Calculate the coordinates of the mean point G in the cloud. Place point G on the diagram.

3) a) Calculate the coordinates of the mean point *G1* associated with the points in the cloud with the five smallest abscissas and the coordinates of the mean point *G2* associated with the other five points in the cloud.

b) Take the line (*G1G2*) as the adjustment line. Draw it.

c) Determine an equation of *(G1G2)* in the form $y = ax+b$.

4) The parts produced must have a diameter of 8 mm, with a tolerance of 0.l mm. Determine graphically the number of parts that can be produced before the diameter reaches 8.1 mm, then calculate this number using the equation found in 3)c). (Round off to the nearest whole number).

Exercise 10: *Laboratory tests*

The following table gives the results obtained from 10 laboratory tests on the breaking load of a steel as a function of its carbon content.

Teneur en carbone x_i	70	60	68	64	66	64	62	70	74	62
Charge de rupture y_i (en kg)	87	71	79	74	79	80	75	86	95	70

1) Plot the cloud of points with coordinates (*xi, yi*).

The x-axis will be 1 cm for one unit, representing the x-axis from value 60. On the ordinate, take 1

cm for 2 kg, representing the ordinates from 70 upwards.

2) Calculate the coordinates of the mean point of this cloud.

3) Determine the linear correlation coefficient of the statistical series of variables *x* and *y* to the nearest 10-3. Interpret the result.

4) Determine an equation of the form y = *ax* + *b* of the regression line D from *y* to *x* by the method of least squares. Give the approximate values of the coefficients *a* and *b to the* nearest 10-3. Plot the line D on the graph in 1).

5) A steel has a carbon content of 77. Give an estimate of its breaking load.

Exercise 11: Car fuel consumption

The fuel consumption of a car, *z*, is given as a function of its speed, *x*, by the following table:

x (en km/h)	80	90	100	110	120
z (en litres/ 100 km)	4	5	6,5	8	10

1) Is fuel consumption proportional to speed? Quickly justify your answer.

2) After reproducing the table above, complete it with a line: $y = \ln z$, giving approximate values to

6 decimal places (the best possible).

3) In a reference frame with origin 0 *(xo* = 70; *yo* = 1.30 using 1 cm for 1 0 km/h as the abscissa

and 1 cm for 0, 1 0 as the ordinate, plot the cloud of 5 points *(x,* $y = \ln z$).

4) Give the equation of a least squares adjustment line for the five points with coordinates *(x, y)* in the cloud. Give this equation in the form y = *Bx* + *A*, with the best possible approximate values of *B* and *A* to 3 decimal places obtained using a programmable calculator.

5) Estimate *y* for a speed of 140 km/h.

Estimate the fuel consumption per 100 km for this speed of 140 km/h, to the nearest 0.5 L as in the table initially given.

Exercise 12

The table below gives the joint numbers for the joint distribution of the two variables X = "Mother's country of birth" and Y = "Father's country of birth":

Y (père) / *X (mère)*	*né en France*	*né à l'étranger*
né en France	*129*	*17*
né à l'étranger	*13*	*30*

1. Specify the nature of the variables studied.

2. Show the joint distribution of (X; Y).

3. Complete the joint headcount table by giving the marginal distributions

(i.e. margins of X and Y); plot the marginal distribution of Y.

4. Plot the distribution of X conditional on Y. What can you tell from this

graph about the relationship between X and Y?

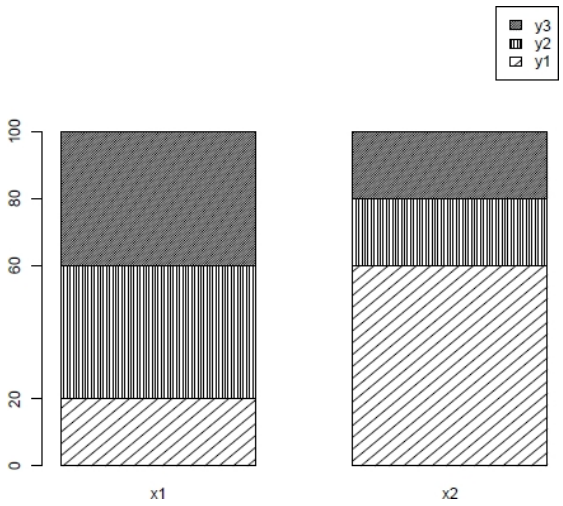
Exercise 13:

We have two variables X and Y for which we have obtained the graph shown in the figure below (the vertical axis is expressed in %).

1. Is it the: joint distribution of (X; Y)? marginal distribution of X? marginal distribution of Y? conditional distribution of X to Y? conditional distribution of Y to X?

2. Construct the table corresponding to this graph.

3. Knowing that the marginal distribution of X is given by $n1_{.} = 50$ and $n2_{.} = 60$, draw up the corresponding contingency table, explaining the margins (i.e. marginal laws).



Exercise 14:

Some of the pupils at a secondary school were surveyed to find out the distance, grouped into three categories (short, medium, long), that they had to travel to get t o school (i.e. home/school distance). We are also interested in the variable Y = school level. The aim is to

to study the possible impact of distance from home on school results. This gives us the following table:

X \ Y	*faible*	*moyen*	*élevé*
courte	23	25	79
moyenne	83	85	55
longue	102	21	27

1. Give the theoretical headcount table.

2. Give the table showing the contributions; deduce the Chi-2 value.

3. Calculate Cramer's V; what can you say about the strength of the link between the distance separating the pupil from the college and the level of education?

Bibliography

[1] J. Blard-Laborderie*, L'essentiel des outils de statistique descriptive pour aborder des études en sciences humaines et sociales, 2015.*

[2] G. Calot*, Cours de statistique descriptive, Dunod, 1969.*

[3] G. Chauvat and J.-P. Reau*, Statistiques descriptives, Armand Colin, 2002.*

[4] M. Tenenhaus*, Statistics: Methods for describing, explaining and forecasting, Dunod, 2006.*

[5] J.-J. Droesbeke*, Éléments de statistiques, Ellipses, 2001.*

[6] L. Leboucher and M.-J. Voisin*, Introduction à la statistique descriptive, 2013.*

[7] F. Mazerolle*, Descriptive Statistics, 2009.*

[8] B. Oukacha and M. Benmessaoud*, Descriptive statistics and probability calculus, 2013.*

[9] J. Vaillant*, Eléments de Statistique descriptive, 2015.*