Lebanese University

Faculty of Sciences

Lebanese University
Faculty of Sciences

S1101
Statistics

Department of Applied Mathematics

Fall Semester

2021/2022

Contents

Chapter 1

Basic Statistical Concepts

1.1 General Introduction

Definition 1 *Statistics is the science of collecting, analyzing, presenting, and interpreting data, as well as making decisions based on such analysis.*

Every day we make decisions that may be personal, business related, or of some other kind. Usually these decisions are made under conditions of uncertainty. Many times, the situations or problems we face in the real world have no precise or definite solution. Statistical methods help us make scientific and intelligent decisions in such situations. Broadly speaking, applied statistics can be divided into two areas: descriptive statistics and inferential statistics:

Definition 2 *Descriptive statistics consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.*

Definition 3 *Inferential statistics consists of methods that use sample results to help make decisions or predictions about a population.*

1.2 Basic Terminology

A) Population: A population is the set of all elements that are subject to a statistical study. For example, Students of the Lebanese University, Patients visiting the emergency clinic, Lebanese families, etc.

B) Statistical unit or individual or element: is a specific subject or object about which the information is collected. For example, a student of the LU, a patient, a family, etc.

C) SAMPLE: A sample is a representative set (part) of the reference population. The number of individuals in a sample is less than the population. For example, 100 Students of the LU, 20 patients, etc.

D) SAMPLE SIZE: The sample size of a statistical sample is the number of observations that constitute it. It is typically denoted by n and it is always a positive integer.

E) CENSUS AND SAMPLE SURVEY: A survey that includes every member of the population is called a census. The technique of collecting information from a portion of the population is called a sample survey.

F) REPRESENTATIVE SAMPLE: A sample that represents the characteristics of the population as closely as possible is called a representative sample.

G) SAMPLING: A sample is some part of a larger body specially selected to represent the population. Sampling is the process by which this part is chosen. Sampling is thus taking any portion of a population or universe that is instead of as representative of that population or universe .

H) STATISTICAL VARIABLE OR CHARACTER: A statistical variable is each of the qualities or characteristics that the individuals of a population have. For example, ; age of students is a statistical variable. It can take on many different values, such as 18, 21, 23, etc.

I) MODALITIES: this is the qualitative or quantitative value that can take the previously defined variable. Example: Gender (male and female), weight 45 kg, etc.

J) STATISTICAL SERIES: It is the whole of the results of a study, in other words it is the values of the character and the corresponding frequency. There are three types of series in Statistics and Data:

- Individual Series: is a statistical series in which all the observations are listed out and all the observations have a frequency of 1.
- Discrete Series: is a statistical series in which all the observations are listed out along with their corresponding frequency in the form of a table. The observations may not necessarily have the same frequency.
- Continuous Series: is a statistical series in which all the class intervals along with their corresponding frequency are listed out in the form of a table. The observations may not necessarily have the same frequency.

Example : A statistical evaluation of student grades at the LU is:
$10, 9, 12, 14, 14, 10, 12, 10, 10, 12$

x_i	9	10	12	14
n_i	1	4	2	2

Table 1.1: Distribution of grades

K) DATA SET: A data set is a collection of observations on one or more variables. For example The data of table below represents the height, weight and gender of 9 students at the Lebanese University

Height	Weight	Gender
183	80	M
182	75	M
173	66	F
178	78	M
192	77	M
158	57	F

Table 1.2: Distribution of Information about students

1.3 Type of Variables

A variable may be classified as **qualitative** or **quantitative**. These two types of variables are explained next

Qualitative or Categorical Variables: A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a qualitative or categorical variable. The data collected on such a variable is called qualitative data. For example, Gender, professional status, etc. When there is no natural ordering of the categories, we call them *nominal* categories. Examples might be gender, religion, or sport. When the categories may be ordered, these are called *ordinal* variables. Categorical variables that judge size (small, medium, large, etc.) are ordinal variables.

Quantitative Variables: A variable that can be measured numerically is called a quantitative variable. The data collected on a quantitative variable is called *quantitative data.* Such quantitative variables may be classified as either *discrete* variables or *continuous* variables.

1. Quantitative discrete variable: A variable whose values are countable is called a discrete variable. In other words, a discrete variable can assume only certain values with no intermediate values. Example: number of students at LU.

2. Quantitative continuous variable : A variable that can assume any numerical value over a certain interval or intervals is called a continuous variable. Example: Weight, Employee salaries.

Chapter 2

Univariate Statistical Analysis

We consider, here in this chapter, a unique Variable X. The objective is to explore the elementary tools, with respect to the type of the variable X, allowing to present this variable via an appropriate graphical representation and to summarize the principal characteristics of the distribution.
We will define some notions related to this chapter, then we will present the graphical and numerical representations for each type of variables.

2.1 Basic Notation

Definition 4 *The frequency, denoted by n_i, is the number of appearance of the value x_i in the statistical distribution.*

The frequencies will be grouped with the corresponding values of x_i in a table called *frequency table distribution* of the variable X (Table 2.1):

x_i	x_1	x_2	$\ldots$	x_i	$\ldots$	x_k
n_i	n_1	n_2	$\ldots$	n_i	$\ldots$	n_k

Table 2.1: Frequency distribution of the variable X

Definition 5 *(**Total Frequency**) The total frequency, denoted by n, is the total number of observations. We have:*

$$n = \sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k$$

Definition 6 *(**Relative Frequency**) A relative frequency f_i is obtained by dividing the frequency of the value x_i by the total number of observations. This ratio is smaller or equal than 1.*

Moreover:

$$f_i = \frac{n_i}{n} \quad \text{and} \quad \sum_{i=1}^{k} f_i = 1$$

Definition 7 *(**The Percentage**) The percentage, denoted by p_i is given as relative frequency multiplied by 100.*

We have

$$p_i = f_i \times 100 \quad \text{and} \quad \sum_{i=1}^{k} p_i = 100$$

Definition 8 *(**The Increasing Cumulative Frequency**) the increasing cumulative frequency (denoted by N_i) and the increasing cumulative relative frequency (denoted F_i) are defined as follows:*

$$N_i = \sum_{j=i}^{k} n_j \quad \textit{and} \quad F_i = \sum_{j=i}^{k} f_j$$

In others words, N_i represents the number of observations smaller or equal than x_i and F_i their relative frequencies. We denote that:

$$N_k = n \quad \text{et} \quad F_k = 1$$

Definition 9 *(**The Decreasing Cumulative Frequency**) the decreasing cumulative frequency (denoted by $N_i \downarrow$) and the decreasing cumulative relative frequency (denoted $F_i \downarrow$) are defined as follows:*

$$N_i \downarrow = \sum_{j=1}^{i} n_j \quad \textit{and} \quad F_i \downarrow = \sum_{j=1}^{i} f_j$$

In others words, $N_i \downarrow$ represents the number of observations greater than or equal to x_i and $F_i \downarrow$ their relative frequencies. We denote that:

$$N_1 \downarrow = n \quad \text{et} \quad F_1 \downarrow = 1$$

2.2 Graphs

When we observe a character describing individuals, the previously defined tables are not very speaking nor meaningful. However, they are very useful for constructing various graphs, which allow to have an idea of the way in which individuals are distributed.

2.2.1 Nominal and Ordinal Qualitative Variable

By definition, the values of a qualitative variable are not numerical but features, called modalities. When these terms are naturally ordered, the variable is called ordinal (for example, the grade in a student population takes three modalities: good, very good and excellent). In the opposite case i.e. when the values of the variables cannot be ordered then the variable is called nominal (for example, the profession of the students parents: doctor, professor, worker)

In the statistical study of a qualitative variable, one will thus be satisfied to make statistical tables and graphical representations. It should be noted that the notions of cumulative frequencies and cumulative relative frequencies only make sense for ordinal variables (they are not defined for nominal variables).

The graphical representations that we encounter with the qualitative variables are several. The most common, which are also the most appropriate, are:

- Bar chart
- Pie chart

Strip chart

A strip chart is a set of rectangles where each rectangle corresponds to a value (or observation).

- The modalities of the variable are placed on a horizontal line (note that if the variable is nominal, do not direct this line because the modalities do not have an order relation).
- The frequencies or relative frequencies are placed on a vertical axis. The height of the rectangle associated to the modality x_i is proportional to its frequency.
- The rectangles have a certain thickness so that there is no confusion with the bar chart in the case of a discrete quantitative variable.
- There must be a space between the rectangles to avoid confusing them with the histograms in the case of a continuous quantitative variable.

Pie chart

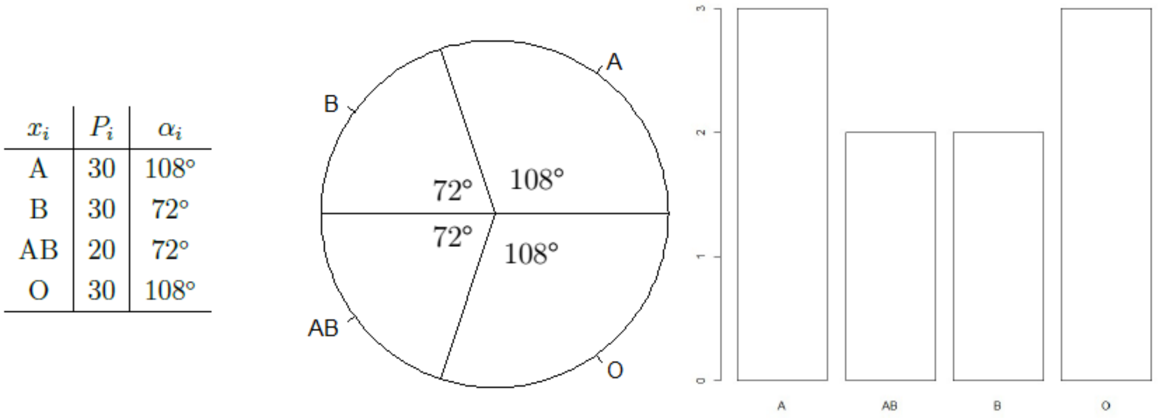
- The total frequency is represented by a circle.
- Each modality is represented by an angle proportional to its frequency.

- The angle α_i associated with the value x_i of the variable X is calculated as follows:

$$\alpha_i = f_i \times 360$$

Remark 1 *This type of graphs is widely used but this can be a too bad presentation since it presents a risk in the interpretation. Actually, it is more easy to compare heights of rectangles in a bar chart than angles in a pie chart specifically when the angles are not too different.*

Example 1 *We asked* 10 *students about their blood type and we obtain the following distribution:* $A, B, O, AB, A, O, O, B, AB, A$.



x_i	P_i	α_i
A	30	108°
B	30	72°
AB	20	72°
O	30	108°

Figure 2.1: Graphical representation of a nominal variable

Example 2 *The distribution of the scores (X) of 130 students in the Lebanese University are given in the following table, where La distribution de notes de 130 étudiants est donnée dans la table qui suit.*

x_i	n_i
Excellent (1)	10
very good (2)	25
Good(3)	40
Pretty good (4)	30
Fair (5)	25

Table 2.2: Frequency distribution of the variable X

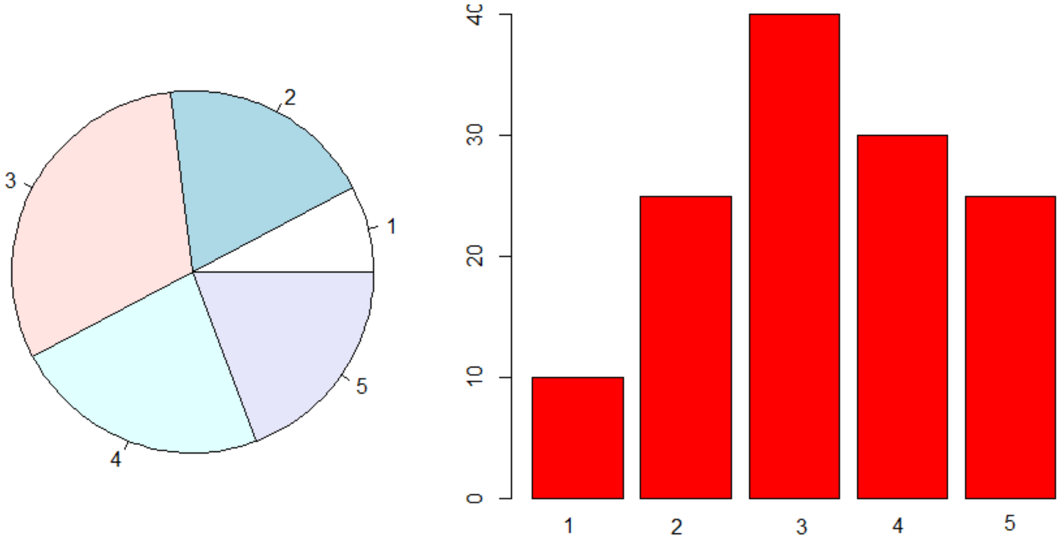


Figure 2.2: Graphical representation of an ordinal variable

2.2.2 Quantitative Discrete Variable

The basic graphs of this variable is the bar chart and the cumulative distribution function (for the cumulative frequency).

Example 3 *We observe the number of people per household for* 50 *households in a district. The obtained results are showed in the Table 2.3.*

1	1	1	1	1	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	6	6	6	8	8

Table 2.3: The distribution of the number of people

The frequency table (Table 2.4) and tha associated bar chart of the variable number of persons are given here after:

x_i	n_i	N_i	f_i	F_i
1	5	5	0.100	0.100
2	9	14	0.180	0.280
3	15	29	0.300	0.580
4	10	39	0.200	0.780
5	6	45	0.120	0.900
6	3	48	0.060	0.960
8	2	50	0.040	1

Table 2.4: Frequency distribution of the *number of persons*

Bar Chart

A bar chart is a set of juxtaposed bars where each bar corresponds to a value x_k and the height of the bar corresponds to its frequency n_k (or relative frequency f_k or percentage p_k). Note that, in a bar chart:

- The discrete values of the variable X are ordered in the x-axis.
- Frequencies, relative frequencies or percentages are placed in the y-axis.

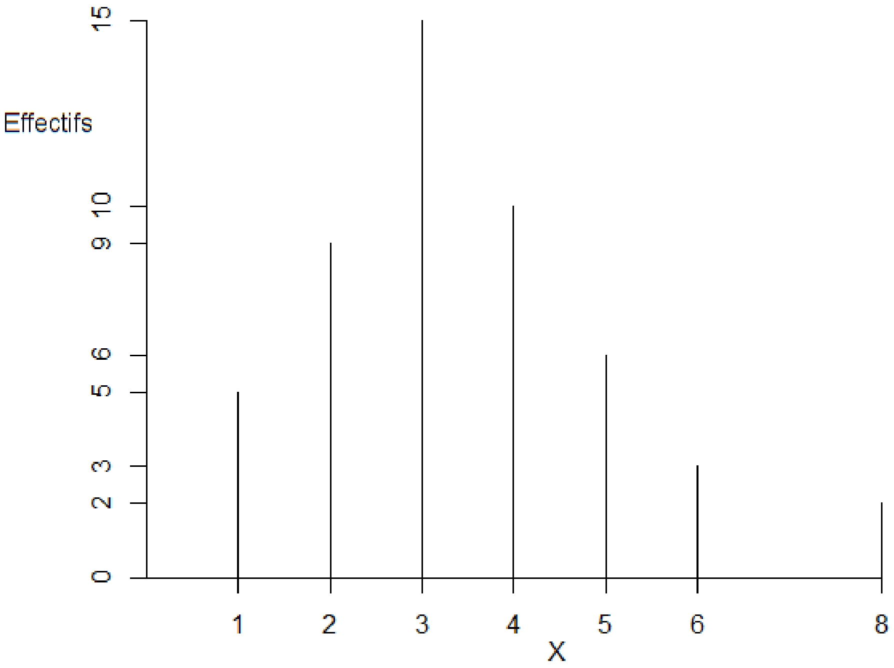


Figure 2.3: Bar Chart

Frequency polygon

The frequency polygon can be obtained by joining the tops of the bars.

Cumulative diagram or cumulative distribution function

This graph helps to visualize the cumulative frequencies (or relative cumulative frequencies or cumulative percentages). It also allows to determine the number, or the proportion, of observations smaller ot equal than a given value in the series. In a cumulative diagram:

- The values of the variables are presented in the x-axis.
- The cumulative frequencies (or relative cumulative frequencies or cumulative percentages) are presented in the y-axis.
- In front of each observation, a point is presented having an ordinate equal to the associated cumulative frequency. Then, to complete the graph, an horizontal segment is drawn at this point between the two observations. **SenTenCE To complete**.

Example 4 *Going back to the Example 3, the cumulative diagram is shown in the Figure 2.4.*

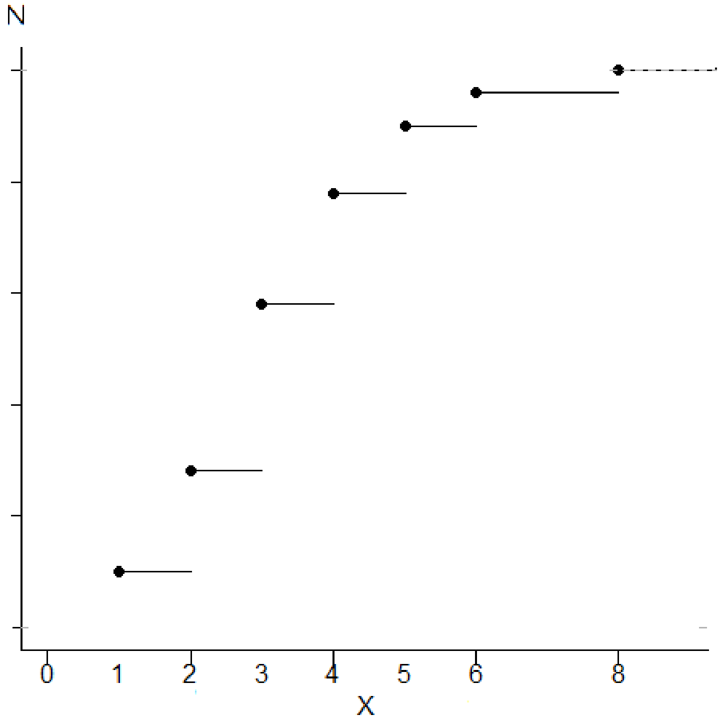


Figure 2.4: Cumulative Diagram

2.2.3 Quantitative Continuous Variables

A continuous variable is a variable taking an infinite number of values. Then the domain of definition is $\mathbb{R}$ or an interval of $\mathbb{R}$. For example, the height or the weight. To make

graphical representations and statistical tables, the observed values of the variable are regrouped in classes.
If $[C_i^-; C_i^+[$ represents the class i, let denote:

- C_i^- the lower bound of the class i,

- C_i^+ the upper bound of the class i,

- $C_i = \dfrac{C_i^- + C_i^+}{2}$ the midpoint or the center of the class i,

- $a_i = C_i^+ - C_i^-$ the width of the class i,

- n_i the frequency of the class i,

- N_i the cumulative frequency i,

- f_i the relative frequency of the class i,

- F_i the cumulative relative frequency of the class i.

The basic graphical that represents such variable are histogram, frequency polygon and cumulative distribution function.

Histogram

The histogram is a set of juxtaposed rectangles. Each rectangle, associated with each class, has a surface (an area) proportional to the frequency of that class. If the classes are of equal amplitude (width), then the height of the rectangles is proportional to the frequency of the class. Before any histogram construction, we must therefore look at whether the classes are of equal amplitude or not. A horizontal axis is used to represent the class boundaries of the variable under consideration. The case of classes of equal amplitude, the frequencies are placed in the y-axis. Otherwise, i.e. if the classes have different amplitudes, we report on the y-axis the density $d_i = \dfrac{n_i}{a_i}$.

Example 5 ***Equal amplitude***
We measured the surface (in Hectares) of 400 *farms in a certain region and the following distribution has been obtained:*

Surface	$[2,6[$	$[6,10[$	$[10,14[$	$[14,18[$	$[18,22[$
Number of farms	60	80	130	100	30

Table 2.5: Distribution of the farms - Equal amplitude

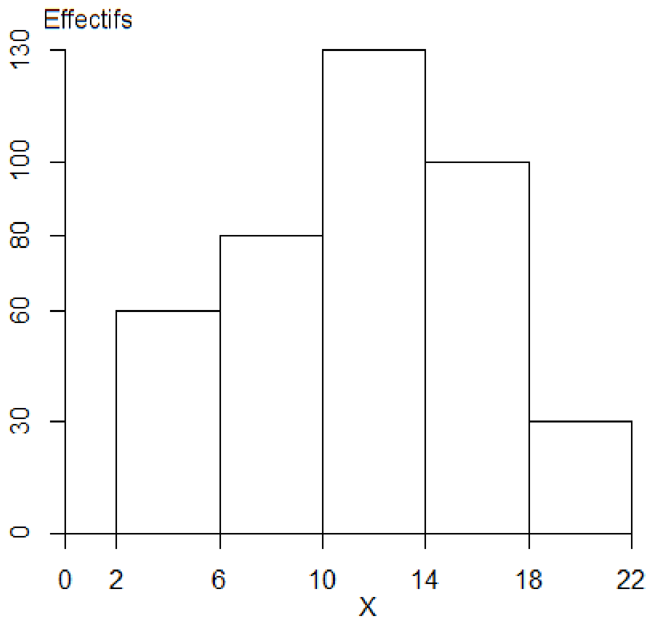


Figure 2.5: Histogram of the variable surface - Equal amplitudes

Example 6 ***Not Equal amplitude*** *We measured the surface (in Hectares) of* 400 *farms in a certain region and we get the following:*

Surface	$[2,3[$	$[3,4[$	$[4,8[$	$[8,16[$	$[16,22[$
Number of farms	60	80	130	100	30

Table 2.6: Distribution of the farms - Unequal amplitude

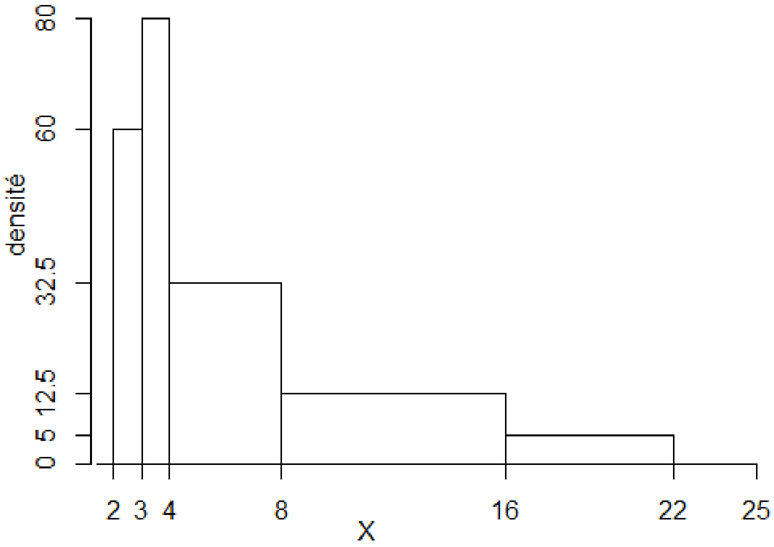


Figure 2.6: Histogram of the variable surface - Unequal amplitudes

Frequency polygone

If the amplitudes are equal, this polygon represents the line that joints the ordinates of the midpoints of the classes $[C_i^-; C_i^+[$. however, if the amplitudes are not equal, then a common amplitude is to be defined ($a'_i = pgcd(a_i)$ for all the classes) and then we follow the same manner as for the equal amplitude.

Cumulative frequencies curve

For a quantitative variable, each class is firstly represented by a single point whose abscissa is the upper bound of the class and the ordinate is the number (or the frequency, or the percentage). cumulative of this class. The cumulative curve is then the curve joining these points.

Example 7 *The cumulative (increasing) frequencies curve of the last example (Example 6) is provided in the Figure 2.7*

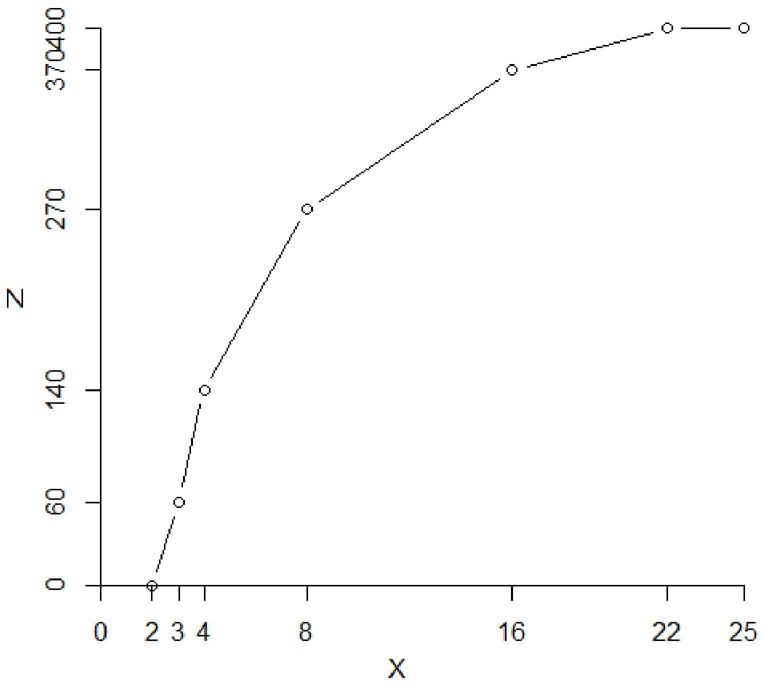


Figure 2.7: Cumulative frequencies polygone

Remark 2 *Regarding the cumulative curve of decreasing cumulative frequencies (or decreasing cumulative relative frequencies), each class must first be represented by a single point whose abscissa is the lower bound of the class and the ordinate is the cumulative frequency (or relative frequency, or percentage) of that class. The cumulative curve is then the curve joining these points.*

2.2.4 Grouped Data

The distribution of raw data in classes requires defining the number of classes k and therefore the amplitude of each class. There are several rules that allow us to compute the number of classes and the amplitude for a statistical series of n observations. In fact, the optimal number k of the classes satisfies the following property:

$$2^k \geq n \Longrightarrow k \ln 2 \geq \ln n \Longrightarrow k \geq \frac{\ln n}{\ln 2}$$

the width (amplitude) of the classes is given by:

$$a \geq \frac{x(n) - x(1)}{k}$$

Example 8 *We observed the salary of* 12 *employees and we obtained:*

$$315, 423, 300, 548, 374, 452, 463, 339, 324, 432, 315, 450$$

The optimal number of classes that we should take is 4 *with an amplitude equal* 62.

2.3 The Numerical Characteristics

Essentially, there are characteristics of central tendency (or of position or of location) and of characteristics of dispersion

2.3.1 Measures of Central Tendency

A measure of central tendency is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of central tendency:

- The mean
- The mode
- The median

These three parameters do not describe the same thing and are complementary in the description and the analysis of a distribution. These parameters can be calculated only in the case of quantitative variables.

The mean (The arithmetic mean)

The mean denoted by $\overline{x}$, is one of the fundamental parameters of central tendency but it is not sufficient to characterize a distribution. The mean is the measure the most calculated and most used when of the description of statistical series.
For a sample of n observations, $x_1, x_2, \ldots, x_n$,the mean is equal to the sum of all the values in the data set divided by the number of values in the data set, then

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

If the distribution is given as a table of frequencies, where n_i is the frequency of x_i , then:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \quad k \quad \text{is the number of the different values}$$

If the data are grouped into classes, c_i is the class midpoint, the mean is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i, \quad k \quad \text{is the number of classes}$$

Remark 3 *Note that the sum of the deviations from the mean is zero. Moreover, the mean is affected by the extreme values and the mean obtained after grouping the data into classes differs slightly due to a loss of information.*

Example 9 *The mean of the variable of the example 6 is* $\overline{x} = 2980/400 = 7.45$

Example 10 *The mean of the variable age of the example 3 is:* $2094/48 = 43.625$.

The Mode

The mode, denoted by M_o is the observation with the highest frequency. We can calculate the mode for a qualitative or quantitative variable. For the grouped data into classes we find first the modal class (it's the class that contains the mode) then we calculate the mode as the follows:

1) If the classes have the same amplitude:
   the modal class $[C_i^-, C_i^+[$ is the class with the highest frequency and we have:

$$Mo = C_i^- + \frac{(n_i - n_{i-1})}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \times a_i$$

2) If the classes have different amplitudes:
   the modal class $[C_i^-, C_i^+[$ is the class with the highest density and we have:

$$Mo = C_i^- + \frac{(d_i - d_{i-1})}{(d_i - d_{i-1}) + (d_i - d_{i+1})} \times a_i$$

Where:
C_i^- is the lower limit of the modal class,
$n_i(resp.d_i)$ is the frequency (resp. density) of the modal class,
$n_{i-1}(resp.d_{i-1})$ is the frequency (resp. density) immediately below the modal class,
$n_{i+1}(resp.d_{i+1})$ is the frequency (resp. density) immediately after the modal class,
a_i is the amplitude of the modal class.

Remark 4 *The mode is not influenced by the extreme values of the variable distribution but is very sensitive to variations of amplitude of classes. Moreover, the mode may not exist, and if it exists, it may not be unique.*

The Median

The median, denoted by Me, is the middle value of a statistical series in ascending order. The median therefore is the value that divides an ordered data set into two equal parts For a sample of n values of a discrete variable, the median is obtained as follows:

1) Order the data from smallest value to largest value: $x(1) \leq x(2) \leq \ldots \leq x(n)$

2) 2) Then the median is :

$$Me = \begin{cases} \dfrac{x(\frac{n}{2}) + x(\dfrac{n}{2}+1)}{2} & si\ n\ is\ even \\ x([\dfrac{n}{2}]+1) & si\ n\ odd \end{cases}$$

where $[\dfrac{n}{2}]$ is the integer part of $\dfrac{n}{2}$.

For a sample of n values grouped into classes, proceed as follows:

1) We find the median class , $[C_i^-, C_i^+[$, is the class that have $N(C_i^-) < \frac{n}{2} < N(C_i^+)$

2) We calculate $Me = C_i^- + \dfrac{\frac{n}{2} - N_{i-1}}{n_i} \times a_i$

Remark 5 *The median is the x-coordinate of the intersection point of the increasing cumulative frequency polygon (or increasing cumulative relative frequency) and the line* $y = \frac{n}{2}$ *(or* $y = \frac{1}{2}$*).*

Remark 6 *The median is also the point of intersection of the increasing cumulative frequency polygon and the decreasing cumulative frequency polygon.*

2.3.2 Measures of Dispersion

A parameter of dispersion aims to summarize the disparity of observations and their variability on both sides of the central tendency. There are four parameters

- Range of the data (range)
- Variance and standard deviation
- Coefficient of variation
- Interquartile interval

Range of the data

The range is the difference between the largest and smallest observations in a sample. The extension gives a first idea of the dispersion of the observations. It is unreliable and is affected by extreme values. We have

$$\text{Range} = \max(x_i) - \min(x_i) = x(n) - x(1)$$

Example 11 *the range of the set* $2, 3, 3, 5, 5, 5, 8, 10, 12$ *is* $12 - 2 = 10$.

Variance and standard deviation

The variance is a measure of the dispersion of a set of data points around their mean value. In other words, variance is a mathematical expectation of the average squared deviations from the mean $\overline{x}$.
Four a statistical series of n values $x_1, x_2, \ldots, x_n$, the variance is defined by :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

If we have a frequency table, the variance is:

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2, \quad k \quad \text{is the number of different values}$$

If we have a grouped data in classes then the variance is

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \overline{x})^2, \quad k \quad \text{is the number of classes}$$

The variance is always positive. If $V(X)$ is large then the data said to be dispersed.

The standard deviation denoted by σ is the radical of the variance $\sigma(X) = \sqrt{V(X)}$.

Property 1 *The variance can also be written as follows:*

$$V(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

or

$$V(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2, \quad \textit{kis the number of classes}$$

Example 12 *In the example 6, the variance is* $V(X) = \frac{31265}{400} - 7.45^2 = 22.66$ *and the standard deviation* $\sigma(X) = 4.7603$*.*

Coefficient of Variation

Although the variance and the standard deviation play an extremely important role in statistics, they are sometimes not the best choices as measures of spread A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows: (standard deviation) / (expected value). The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another. It is calculated as follows:

$$CV(X) = \frac{\sigma(x)}{\overline{x}}$$

The coefficient of variation used to compare the variability of two sets which have very different averages (means) average or even they are not expressed in the same units. The coefficient of variation provides the homogeneity of the series, if the coefficient of variation is less than 15%, considering that the data are homogeneous and conversely, if the variation coefficient exceeds 15%, it is said that the data is heterogeneous.

Example 13 *In the example 6, the coefficient of variation is* $CV(X) = \frac{4.7603}{7.45} = 63.896$*.*

Example 14 *The following table 2.7 provides the distribution of a sample of 10 individuals following the age and the salary.*

Num	Age	Salary
1	37	300
2	35	310
3	36	290
4	36	305
5	41	305
6	38	295
7	40	300
8	36	310
9	35	290
10	37	295

Table 2.7: Data table of age (in years) and the salary (in dollars)

	Age	Salary
Mean	37.1	300
Variance	3.69	50
Standard deviation	1.92	7.07
CV	0.05	0.02

Table 2.8: Characteristics of two variables

We remark that the standard deviation of the age is 1.92 *years and the standard deviation of the salary is* 7.07*. We cannot say that the salaries are more dispersed than the ages because the two variables do not have the same units. To compare the dispersion in the two variables, we calculate the coefficient of variation because this statistic has no unit. As a result, the dispersion is more important in the salaries than in the ages.*

Interquartile range

It's the difference between the third and the first quartiles, denoted by IQR and we have

$$IQR = Q_3 - Q_1$$

The IQR does not depend on the extreme values of a distribution.

Quantiles

The Quantiles are values that divide the distribution into several equal parts. We are interested in the quartiles the *quartiles* (dividing the distribution into 4 equal parts), the *deciles* (dividing the distribution into 10 equal parts), and the *percentiles* (dividing the distribution into 100 equal parts).

- The quartiles:

The quartiles Q_1, Q_2 and Q_3 are three values that divide the statistical distribution into four equal parts.Q_2 is the median, the first quartile Q_1 is the smallest data in the list such that at least 25% of the data is less than or equal to Q_1. The third quartile Q_3 is the smallest data on the list such that at least 75% of the data is less than or equal to Q_3.

A) To find the quartiles of a series of n values $x_1, x_2, \dots, x_n$:

1) Order the data from smallest to largest $x(1), x(2), \dots, x(n)$

2) Calculate Q_1, Q_2 and Q_3 associated respectively with the values $\alpha = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$ such that

$$Q_i = \begin{cases} \dfrac{x(n\alpha) + x(n\alpha + 1)}{2} & \textit{si } n\alpha \textit{ est entier} \\ x([n\alpha] + 1) & \textit{si } n\alpha \textit{ non entier} \end{cases}$$

where $[n\alpha]$ is the integer part of $n\alpha$.

B) To find the quartiles of a variable given by a frequency distribution

1) Determine the increasing cumulative frequencies N_i distribution

2) Calculate $n\alpha$ for $\alpha = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$ and we take the value of the variable that correspond to the increasing cumulative frequency more than or equal to $n\alpha$

C) For the grouped data in classes, we follow the following steps:

1) Determine the increasing cumulative frequencies N_i distribution

2) Calculate $n\alpha$ for $\alpha = \frac{1}{4}, \frac{1}{2}$ et $\frac{3}{4}$ and we take the class $[C_i^-, C_i^+[$ of the variable that correspond to the increasing cumulative frequency more than or equal to $n\alpha$

3) the quartile Q_i is within the quartile Class

$$Q_i = C_i^- + a_i \left(\frac{n\alpha - N_{i-1}}{n_i} \right)$$

Remark 7 *The quartiles Q_1, Q_2 and Q_3 are respectively the x-coordinate of the intersection points of the increasing cumulative frequency polygon (or increasing cumulative relative frequency) and the lines $y = \dfrac{n}{4}, \dfrac{n}{2}$ and $\dfrac{3n}{4}$ (respectivelly $y = \dfrac{1}{4}, \dfrac{1}{2}$ and $\dfrac{3}{4}$.*

Remark 8 *for the deciles (or percentiles), we take $\alpha \in \{\dfrac{1}{10}, \dots, \dfrac{9}{10}\}$ (ou $\alpha \in \{\dfrac{1}{100}, \dots, \dfrac{99}{100}\}$.*

2.3.3 Boxplot

A boxplot is a graphical representation summarizing some information contained in the statistical series. Tha basic form of a boxplot is provided in the Figure 2.8. The components of this diagram are as follows:

- A box (rectangle) going from the first to the third quartile of the distribution. The box is also divided by a line corresponding to the median of the distribution.
- This box is completed by two line segments whose ends correspond respectively to the minimum value (to the left of the box) and the maximum value (to the right of the box).

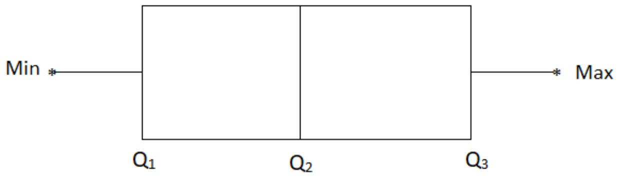


Figure 2.8: Boxplot without extreme values

If the data presents some extremes values, we plot a modified version of the boxplot. This latest diagram is composed of:

- A box going from the first to the third quartile. The box is also is divided by a line corresponding to the median.
- The box is then completed by two lines segments. To draw them, we compute firstly the bounds:

$$v_1 = Q_1 - 1.5\ IQR \quad \text{and} \quad v_2 = Q_3 + 1.5\ IQR$$

- We then identify the smallest and the largest values contained in the interval $[v_1, v_2]$. These observations are called *adjacent values.*
- The segments lines are obtained by joining these adjacent values to the box (to the left and to the right).
- All others values that are not in the interval $[v_1, v_2]$ are represented by points (or stars) and they are called *extreme values.*

The obtained boxplot in case of presence of extreme values take thus the form shown in the Figure 2.9. The extreme values are represented using points or others symbols outside of the class $[v_1, v_2]$.

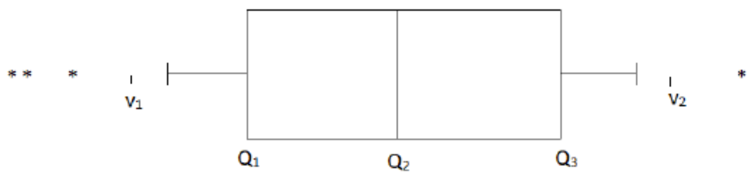


Figure 2.9: Modified version of a boxplot with extreme values

2.3.4 Skewness

The asymmetry (skewness) of a distribution of a variable X is measured using the Fisher asymmetry coefficient:

$$\gamma_1 = \frac{m_3}{\sigma^3}$$

where $m_3 = \frac{1}{n}\sum(x_i - \overline{x})^3$ and σ is the standard deviation of X.

Remark 9 *There is also others parameters indicating the skewness such as the Yule coefficient (A_Y) and the Pearson coefficient (A_P) defined as follows.*

$$A_y = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{IQR}$$

and

$$A_P = \frac{\overline{x} - Mo}{\sigma}$$

the coefficient of skewness can be either positive, negative or zero. Figure 2.10 shows the 3 types of skewness. If the coefficient is null then the distribution is said to be symmetric. If the coefficient is positive then the distribution is said to be positively skewed or skewed right.If the coefficient is negative then the distribution is said to be negatively skewed or skewed left.

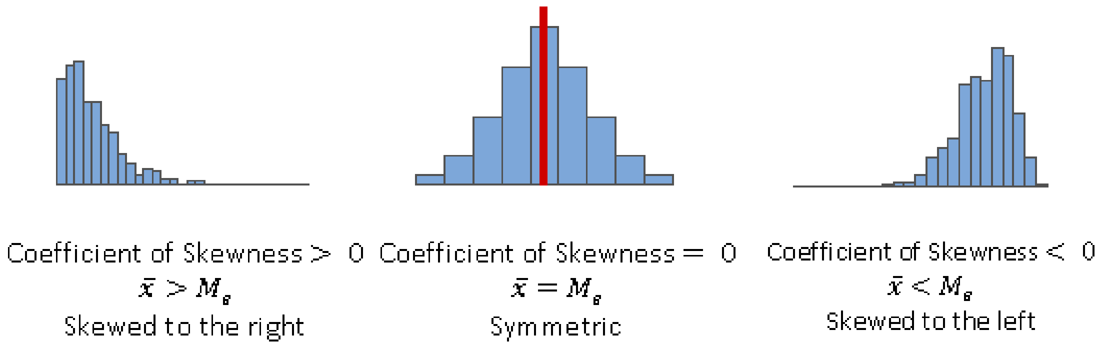


Figure 2.10: The three types of Skewness

Remark 10 *Note that the mean, the median and the mode of a symmetric distribution are equal. However, if the distribution is right skewed then $\overline{x} > Q_2$. If the distribution is left skewed than we have $\overline{x} < Q_2$.*

2.3.5 Kurtosis

The kurtosis of a distribution indicates how much the distribution is flat. In other words, it indicates whether a distribution has a peack or not. The kurtosis is measured using Fisher coefficient defined by:

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

where $m_4 = \frac{1}{n}\sum(x_i - \overline{x})^4$ and σ is the standard deviation of X. This can be either positive, negative or zero. Moreover,

- A distribution with a positive kurtosis value indicates that the distribution has heavier tails than the normal distribution which is a bell shaped distribution and it's called **leptokurtic**.
- A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution and it's called **platykurtic**
- A distribution with null kurtosis is a bell shaped normal distribution and it's called **mesokurtic**.

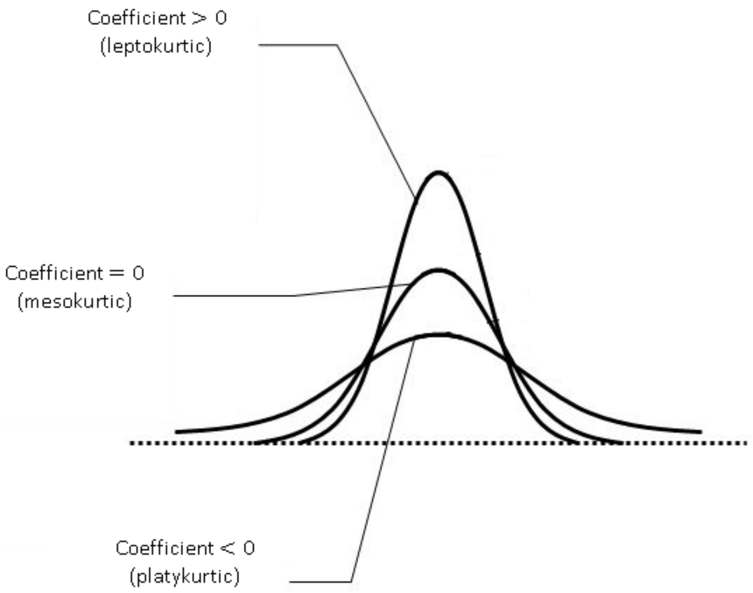


Figure 2.11: Difference between null, positive and negative kurtosis

2.4 Tranformation of variables

Definition 10 *A change of a variable X is a transformation of this variable into a new variable Y by multipying X by a constant and/or adding a constante to the variable X.*

Definition 11 *A change of variables consists of transforming all observations in the statistical series. Generally, it can be represented as follows:*

$$y_i = ax_i + b; \ \forall \ i = 1, \dots, n$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}$.

If $y_i = ax_i + b$, then:

- the mean of Y is $\overline{y} = a\overline{x} + b$ where $\overline{x}$ is the mean of X.

- the variance of Y is $V(Y) = a^2V(X)$ and the standard deviation is $\sigma(Y) = |a|\sigma(X)$.

Remark 11 *Note that:*

- *The location measures are all affected by a transformation, i.e. $Me(y) = aMe(y) + b$ and $Mo(y) = aMo(x) + b$.*

- *The skewness and the kurtosis are not affected by a transformation of the variable.*

2.5 Means and Variances of two Groups

We assume that the n observations are divided into two groups: Group G_A and Group G_B. The first n_A observations are in the group G_A and the last n_B observations belong to the group G_B, with

$$n_A + n_B = n$$

The statistical series is then of the form:

$$\overbrace{x_1, x_2, \dots, x_{n_A}}, \overbrace{x_{n_A+1}, \dots, x_{n_A+n_B}}$$

agregated mean

The means of thwo groups are respectively defined by the following:

$$\overline{x}_A = \frac{\sum_{i=1}^{n_A} x_i}{n_A}, \quad \text{et} \quad \overline{x}_B = \frac{\sum_{i=n_A+1}^{n_A+n_B} x_i}{n_B}$$

The total or the agregated mean is a weighted average of the two means groups. This is written as follows:

$$\overline{x} = \frac{n_A\overline{x}_A + n_B\overline{x}_B}{n_A + n_B}$$

Agregated variance

The variances of the two groups are respectively defined as follows:

$$\sigma_A^2 = V_A(X) = \frac{\sum_{i=1}^{n_A}(x_i - \overline{x}_A)^2}{n_A}, \quad \text{et} \quad \sigma_A^2 = V_B(X) = \frac{\sum_{i=n_A+1}^{n_A+n_B}(x_i - \overline{x}_B)^2}{n_B}$$

Here after the total or the agregated variance of the variable X:

$$V(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

is decomposed as follows:

$$V(X) = \frac{n_A\sigma_A^2 + n_B\sigma_B^2}{n} + \frac{n_A(\overline{x}_A - \overline{x})^2 + n_B(\overline{x}_B - \overline{x})^2}{n}$$

2.6 Exercises

Exercise 1 The table here after provides the distribution of the monthly salary, in Dollars of the employees in a certain company:

Salary	$[800; 900[$	$[900; 1000[$	$[1000; 1050[$	$[1050; 1150[$	$[1150; 1300[$
Frequency	42	49	74	19	16

a) Identify the population, the sample and its size, the individual, the variable and its type.

b) Compute the average salary in this company.

c) In this company, how many employees receive less than 1050 Dollars?

d) Compute the standard deviation of this distribution. Comment.

e) Represent this distribution by an appropriate graph.

f) Plot the graph of the cumulative frequency and deduce the quartiles.

g) Represent this data using a boxplot.

h) We add an additional category of employees with salary in the class $[1300; 1500[$. What is the frequency of this class knowing that the average salary is 1200 Dollars.

Exercise 2 The distribution of distance (in Kilometers) between the home and workplace of 100 employees in a certain city is given in the following table.

Distance	0	1	2	3	4	5	8
Increasing cumulative frequency	5	26	50	65	85	98	100

a) Give the frequency of this distribution.

b) Determine the mean of the average distance between the home and workplace.

c) What is the standard deviation of this distribution? Comment your result.

d) Evaluate the median of this distribution and determine the interquartile range. Comment your result.

Exercise 3 We observe the diameters (in centimeters) of trunk of 400 trees and we found the following results:

Diamètre en cm	25	26	27	28	29	30
Pourcentage	10 %	15 %	$\cdots$	35 %	5 %	5 %

a) Complete the missing value in the table.

b) How many trees have a trunk diameter greater than or equal 27 cm?

c) Among the trees having a diameter than 26 cm, what is the percentage of trees with a diameter less than or equal 27 cm ?

d) What is the average diameter of these trees ?

Exercise 4 Studenst in a class are divided into two groups A and B. The mean of the age of 12 students in the group A is 23 years old. While the mean of the age of 15 students in the group B is 25 years old.
Compute the mean of the age of all students of this class.

Exercise 5 In a sample of 120 patients, we observe the quantity of a certain molecule in the blood (in micrograms per liters). The collected data are summarized in the following table :

Quantity($\mu g/l$)	$[130, 140[$	$[140, 150[$	$[150, c[$	$[c, 180[$	$[180, b[$
Frequency	0.1	a	0.3	0.25	0.2

a) Identify the population, the sample and its size, the individual, the variable and its type.

b) Find the value of a.

c) Compute the value of b knowing that the range is 70.

d) Compute the value of c knowing that the average quantity is equal to 165 micrograms per liters.

e) Give the frequency distribution.

f) Hereafter we consider $c = 170$. Plot the histogram of this distribution and deduce the mode.

g) Determine the median of this distribution and comment the obtained value.

h) Compute the percentage of patients having a quantity less than 145 $\mu g/l$.

Exercise 6 We observe the number of defective items produced by two machines A and B in each hour during 100 hours.
Machine A :

number of defective items	0	1	2	3	4	5	6	7
Number of hours	13	42	38	2	2	1	1	1

Machine B :

number of defective items	0	1	2	3	4	5
Number of hours	35	40	1	1	10	13

a) Compute the average number of defective items produced by A. Compute then Var_A.

b) Compute the average number of defective items produced by B. Compute then Var_B.

c) Determine the quartile for both distributions.

d) Represent in the same graph, the two boxplots corresponding to these two distributions. Then compare your results.

Exercise 7 We observe the residence place of 50 inhabitants. The obtained distribution is:

Residence place	Downtown	Suburban	Town	City	Other
Angles	156.6	73.8	57.6	$\cdots$	18

a) What is the type of this variable?

b) Complete the table and compute the corresponding frequencies.

c) Represent graphically this distribution.

d) Determine the mode of this distribution and comment.

Exercise 8 We give the following distribution:

Classes	[2,4[	[4,8[	[8,a[	[a,16]
Frequency	10	30	35	25

Compute a given that the median is equal to 10.

Exercise 9 The distribution of the salary is given by:

Salary	[5,10[	[10,20[	[20,a[	[a,40]
Number of employees	10	35	45	10

a) What is the value of a if the mean of the salary is 22.125.

b) Hereafter $a = 35$, plot the histogram of this distribution.

c) Calculate the median and the mode of this distribution.

d) Determine the standard deviation of this distribution.

Exercise 10 The weight of 50 students is given in the following:

43	43	43	47	48
48	48	48	49	49
49	50	50	51	51
52	53	53	53	54
54	56	56	56	57
59	59	59	62	62
63	63	65	65	67
67	68	70	70	70
72	72	73	77	77
81	83	86	92	93

a) Give the frequency distribution of weight in the appropriate classes.

b) Give the distribution of the increasing cumulative frequencies.

c) Plot the histogram and deduce the mode.

d) Plot the graph of the cumulative frequencies and deduce the median.

Exercise 11 The distribution of the weight of 40 students is given by:

Classes	40-50	50-60	60-75	75-85
Frequency	n_1	12	15	n_4

Complete the table given above knowing that the sample mean is 60.55 kg.

Chapter 3

Two-dimensional Descriptive Statistics

The bivariate statistic analysis consists in studying the variables taken in couples via descriptive technical. The objective of this chapter is to highlight a relation or an absence of relation between two variables. Moreover, in the case of existence of the relation, look for and study this relation between the two variables.

In this chapter, we consider X and Y two variables quantitative (discrete or continuous). These two variables are measured on the n individuals. Then we have a double (bivariate) series which is written in the following form:

$$(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

Example 15 *We are interested in pocket money given to young people aged between 11 and 16. The age (X) and the average weekly amount (Y) expressed in dollars were measured for a sample of 10 individuals. The joint series observed is as follows:*

Table 3.1: joint series of the age (X) and of amount (Y)

x	12	12	15	14	16	14	12	13	11	11
y	4.1	3.4	11.3	10.2	11.5	7.2	6	7.8	3.5	3

3.1 Data Representation

In the case of quantitative variables, the statistical series can be represented by a *scatter plot*. On the other hand, the statistical series may in some cases give rise to a *contingency table* to define a two-dimensional distribution.

3.1.1 Scatter plot

A simple way of visualizing the data is to represent each i individual by a point defined in an orthogonal coordinate system, with its coordinates x_i and y_i. The graph obtained is called a *scatter plot*.

Example 16 *take the example 15, the scatter plot is given in the Figure 3.1.*

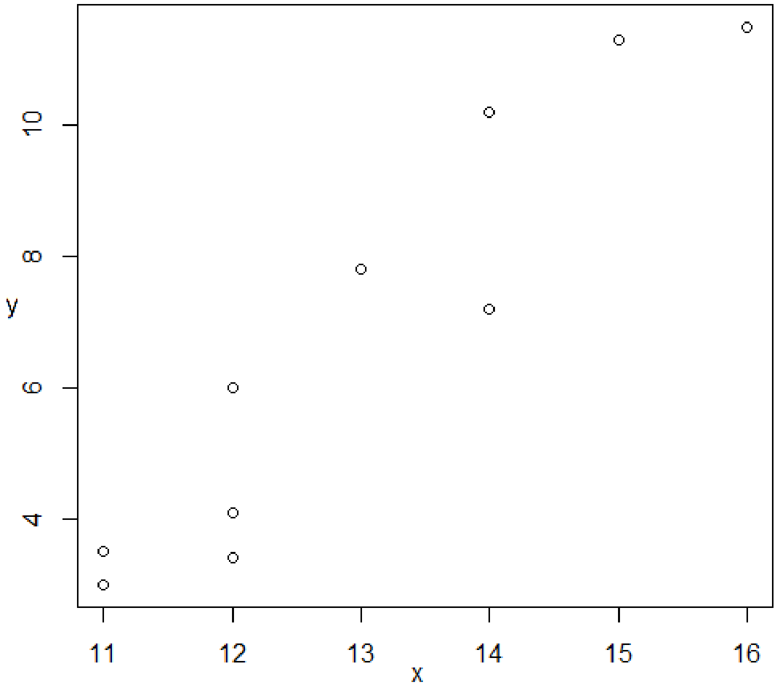


Figure 3.1: Scatter plot

The scatter plot gives how the points are dispersed in the plane. It also allows to take an idea on the form of the relation and its sense in the case of its existence.

3.1.2 Contingency Table

In the case where n is higher, each couple of observations may appear several times. This case arises especially when the variables are discrete and the number of distinct values of each variable is small. Then we can make a similar operation to that introduced in the previous chapter can then be performed by constructing an observed distribution associating with each couple of distinct values observed a frequency representing the number of times it has appeared.

Represent the distinct values of X and of Y by $\{x_1, \ldots, x_p\}$ on the one hand, $\{y_1, \ldots, y_q\}$ on the other hand. Denote by n_{ij} The **frequency** of the couple (x_i, y_j). Then we can define **the distribution of two-dimensions** by the set

$$\{(x_i, y_j, n_{ij}), i = 1, \ldots, p; j = 1, \ldots, q\}$$

This distribution can be represented as a table called a *contingency table*, as shown in Table 3.2.

Y X	y_1	$\dots$	y_j	$\dots$	y_q
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iq}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\dots$	n_{p1}	$\dots$	n_{pq}

Table 3.2: Table de contingence

The frequencies verify the relation

$$\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = n$$

Remark 12 *It is also possible to associate with each couple* (x_i, y_j) *its* ***relative frequency*** f_{ij} *defined by:*

$$f_{ij} = \frac{n_{ij}}{n}, i = 1, \dots, p; j = 1, \dots, q$$

the relative frequencies verify the relationship

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1$$

Remark 13 *The graphical representation of a contingency table is more difficult to construct. We can associate a ”point” of coordinates* (x_i, y_j) *with the couple defined by the* i*th value of* X *and the* j*th value of* Y *by endowing it with a surface equal to* n_{ij}.

Example 17 *The joint distribution of number of the children* (X) *and number of brothers and sisters* (Y) *of a sample of* 80 *women is provided in the Table 3.3.*

Y X	0	1	2	3	4
1	4	4	2	0	0
2	9	16	4	0	0
3	4	12	9	2	0
4	1	6	1	1	2
5	0	1	0	1	1

Table 3.3: Joint distribution of (X) and of (Y) of the Example 17

3.2 Association between two quantitative variables

In this section, we present the description of each separate variable by introducing the definition of marginal and conditional distributions.

3.2.1 Series and marginal distributions

The study of a bivariate series $\{(x_i, y_i), i = 1, \ldots, n\}$ includes firstly the analysis of univariate marginal series obtained by considering only one variable:

- **Marginal series of** X: $\{x_i, i = 1, \ldots, n\}$;
- **Marginal series of** Y: $\{y_i, i = 1, \ldots, n\}$.

The marginal series are univariate series and can therefore be analyzed in detail using one or other of the graphical representations presented in Section 2.2 and by use of different measures of position, of dispersion, and of form introduced in Section 2.3 of Chapter 2. If we have a distribution provided in a contingency table $\{(x_i, y_j, n_{ij}), i = 1, \ldots, p; j = 1, \ldots, q\}$, then we can by a similar approach to define the **marginal distributions**.

a) **Marginal distribution of** X: the marginal distribution of X is defined by the set of couples $\{(x_i, n_{i.}), i = 1, \ldots, p\}$ where we associate with each value x_i the marginal frequency $(n_{i.})$ defined by:

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \quad i = 1, \ldots, p$$

The distribution of X could also be in the form of a frequency table (see Table a) as we saw in Chapter 2.

X	x_1	$\ldots$	x_i	$\ldots$	x_p	Total
marginal frequencies of X	$n_{1.}$	$\ldots$	$n_{i.}$	$\ldots$	$n_{p.}$	n

b) **Marginal distribution of** Y: the marginal distribution of Y is defined by the set of couples $\{(y_j, n_{.j}), j = 1, \ldots, q\}$ Where we associate with each value y_j the marginal frequency $(n_{.j})$ defined by:

$$n_{.j} = \sum_{i=1}^{p} n_{ij} \quad j = 1, \ldots, q$$

The marginal distribution of Y could also be in the form of a frequency table (see Table b) as we saw in the chapter 2.

Y	y_1	$\ldots$	y_j	$\ldots$	y_q	Total
Marginal frequencies of Y	$n_{.1}$	$\ldots$	$n_{.j}$	$\ldots$	$n_{.q}$	n

Remark 14 *The sum of the marginal frequencies is equal to n:*

$$\sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} = n$$

Example 18 *The marginal distributions of X and of Y are easily obtained from the contingency table. In the example 3.3, we have:*

Y X	0	1	2	3	4	$n_{i.}$
1	4	4	2	0	0	10
2	9	16	4	0	0	29
3	4	12	9	2	0	27
4	1	6	1	1	2	11
5	0	1	0	1	1	3
$n_{.j}$	18	39	16	4	3	80

Table 3.4: Contingency table with the marginal distributions

Remark 15 *These distributions are univariate and can be synthesized graphically and/or numerically*

Remark 16 *We can also introduce the notion of the* ***marginal relative frequency****:*

$$f_{i.} = \frac{n_{i.}}{n} \quad ; \quad f_{.j} = \frac{n_{.j}}{n}$$

Remark 17 *The sum of the marginal relative frequencies is equal to* 1*:*

$$\sum_{i=1}^{p} f_{i.} = \sum_{j=1}^{q} f_{.j} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{ij} = 1$$

3.2.2 Conditional distributions

A conditional distribution consists to fix the value for one variable and describing the distribution of the values of the other. We have the conditional distribution of Y given that $X = x_i$ (distribution of $Y|X = x_i$) and the conditional distribution of X given that $Y = y_j$ (distribution of $X|Y = y_j$).

a) **Conditional distribution of Y given X** or distribution of Y given that $X = x_i$
   We fix a value of X, for example $X = x_i$. We would like to know how the values $\{y_1, y_2, \ldots, y_q\}$ of Y are distributed among the individuals in whom the variable X has the value x_i?
   The answer to this question is given by the line of the contingency table associated

with the value x_i of X: when $X = x_i$, the value y_j of Y ($j = 1, \dots, q$) is observed n_{iq} times.

Y X	y_1	$\dots$	y_j	$\dots$	y_q	Marginal of X
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\dots$	n_{p1}	$\dots$	n_{pq}	$n_{p.}$
Marginal of Y	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.p}$	n

Table 3.5: The conditional distribution of Y given X

The conditional distribution of Y given X is the set of couples

$$\{(y_j, n_{ij}); i \text{ fixé}; j = 1, \dots, q\}$$

Example 19 *The conditional distribution of Y given that $X = 2$ in the example 3.3 is*

Y	0	1	2	3	4	$n_{i.}$
$X = 2$	9	16	4	0	0	29

Table 3.6: conditional distribution of $Y|X = 2$

b) **b) Conditional distribution of X given** Y or distribution of X given that $y = y_j$
We fix a value of Y, for example $y = y_j$. We would like to know how the values $\{x_1, x_2, \dots, x_p\}$ of X are distributed among the individuals in whom the variable Y has the value y_j?
The answer to this question is given by the line of the contingency table associated with the value y_j of Y: when $y = y_j$, the value x_i of X ($i = 1, \dots, p$) is observed n_{jp} times.

Y X	y_1	$\dots$	y_j	$\dots$	y_q	Marginal of X
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\dots$	n_{p1}	$\dots$	n_{pq}	$n_{p.}$
Marginal of Y	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.p}$	n

Table 3.7: The conditional distribution of X given Y

The conditional distribution of X given Y is the set of couples

$$\{(x_i, n_{ij}); j \text{ fixé}; i = 1, \dots, p\}$$

Example 20 *The conditional distribution of X given that $Y = 3$ in the example 3.3 is*

X	1	2	3	4	5	$n_{.j}$
$Y = 3$	0	0	2	1	1	4

Table 3.8: Conditional distribution of $X|Y = 3$

Remark 18 *we can also define the conditional relative frequencies:*

a) Conditional relative frequency of Y given that $X = x_i$:

$$f_{y_j|x_i} = f_{j|i} = \frac{n_{ij}}{n_{i.}}, \quad j = 1, \dots q; \quad i \textit{ fixé}$$

b) b) Conditional relative frequency of X given that $Y = y_j$:

$$f_{x_i|y_j} = f_{i|j} = \frac{n_{ij}}{n_{.j}}, \quad i = 1, \dots p; \quad j \textit{ fixé}$$

3.2.3 Analysis of marginal and conditional distributions

The marginal and conditional distributions are univariate distributions and can therefore be analyzed in detail using one or other of the graphical representations presented in Section 2.2 and by use of different measures of position, of dispersion, and of form introduced in Section 2.3of Chapter 2.

In this part, we define the marginal means and the marginal variances. The other parameters of position or of dispersion of the conditional distributions will be made in a similar way.

Marginal means and variances

If we have a bivariate series, the averages and variances of the two variables X and Y are calculated as mentioned in Section 2.3 of Chapter 2. If we have a contingency table, we determine the marginal distribution of X and that of Y, likewise for the conditional distribution of $Y|X = x_i$ and that of $X|Y = y_j$. If we have a contingency table, the marginal means and variances of the two variables X and Y are defined as follows:

a) Marginal means and variances of X:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i$$

and

$$\sigma_X^2 = V(X) = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2 - \overline{x}^2$$

b) Marginal means and variances of Y:

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j$$

and

$$\sigma_Y^2 = V(Y) = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j^2 - \overline{y}^2$$

Example 21 *Take the example 3.3, we have* $\overline{x} = 2.6$*,* $V(X) = 0.99$ *and* $\overline{y} = 1.1875$*,* $V(Y) = 2.139844$*.*

Example 22 *In the example 19, the conditional mean of* $Y|X = 2$ *is* 0.8275862 *and sa variance is* 0.4185493*.*

Example 23 *In the example 20, the conditional mean of* $X|Y = 3$ *is* 3.75 *and sa variance is* 0.6875*.*

3.2.4 Independence

Two variables are independent if the marginal relative frequency of one of the variables does not depend on the other. Then we have the following definition:

Definition 12 *Two variables* X *and* Y *are independent if and only if one of the following properties is true:*

1. $f_{i|j} = f_{i.}$, *for all* i
2. $f_{j|i} = f_{.j}$, *for all* j
3. $n_{ij} = \dfrac{n_{i.}n_{.j}}{n}$, *for all* i *et* j
4. $f_{ij} = f_{i.}f_{.j}$, *for all* i *et* j

3.3 Relation between two variables

In Section 3.1.1, we have seen how to represent a statistical series relating to two quantitative variables using a scatter plot. The examination of the scatter plot enable to decide whether there is a particular structure of association between the two variables. Two questions arise:

1. Can we quantify the strength of this association?

2. 2. If the association between the two variables actually translates a statistical dependence between Y and X, how can define a representation of this particular relation?

In this section, we examine these two problems. In the case where the relationship between the two variables is linear, we use the *linear correlation coefficient* to answer the first question. Then we introduce the *regression line* to answer the second question.

3.3.1 Coefficient of linear correlation

Definition 13 *We call a covariance of a bivariate series (X,Y) with the variables X and Y are quantitative, the quantity denoted $cov(X,Y)$ or σ_{XY} defined by:*

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

where $\overline{x}$ and $\overline{y}$ denote the respective means of X and of Y.

Properties of the covariance

1. 1. The covariance between two variables X and Y may be positive, negative or null.
2. If the covariance is positive then the relation is positive increasing. On the other hand, if the covariance is negative, then the relation is negative decreasing. Finally, if the covariance is zero then there is no relation between the two variables X and Y.
3. $cov(X,Y) = \frac{\sum\limits_{i=1}^{n} x_i y_i}{n} - \overline{xy}$
4. $cov(aX,bY) = ab\ cov(X,Y)$ for all real numbers a and b
5. $cov(X,Y) = cov(Y,X)$
6. $cov(X,X) = V(X)$
7. $cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$

If we have a contingency table, the covariance of the joint distribution is given by:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \overline{x})(y_i - \overline{y})$$

where $\overline{x}$ and $\overline{y}$ denote the respective marginal means of X and of Y.

Remark 19 *If we have a grouped data, the values of the variables will be replaced by the class mark.*

Definition 14 *The linear correlation coefficient of two variables X and Y, denoted $\rho(X,Y)$, is defined by:*

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x.\sigma_y}$$

where σ_x and σ_y the standard deviations of X and Y. This coefficient "measures" the degree of linear relationship between the variables X and Y.

Properties of the coefficient ρ

1. $\rho(X,Y)$ is independent of the units in which the variables are expressed.
2. $\rho(X,Y)$ is invariant fo all change of variables ($\rho(aX+b, cY+d) = \rho(X,Y)$).This result comes from the properties of variance and covariance.
3. The linear correlation coefficient and the covariance take the same sign since the standard deviations are positive

4. $-1 \leq \rho \leq 1$
5. The coefficient of linear correlation is independent of the order in which the two variables are expressed. Then we have

$$\rho(X, Y) = \rho(Y, X)$$

Interpretation of ρ

1. If $\rho = 1$, X and Y are said to be perfectly positively correlated, which means that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases.
2. If $\rho = -1$, X and Y are said to be perfectly negatively correlated, which means that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y decreases as X increases.
3. If $\rho = 0$, then there is no linear correlation between the two variables.
4. If $|\rho| \to 0$ $(| \rho |\leq 0,75)$, X and Y are weakly correlated.
5. If $|\rho| \to 1$ $(| \rho |> 0,75)$, X and Y are strongly correlated and we can consider the existence of a linear relationship between the two variables.

Remark 20 *If the two variables X and Y are independent then the covariance and the coefficient of linear correlation between X and Y are null but the inverse is not true.*

3.3.2 Simple Linear Regression

The analysis of the regression can be defined as the search for the relationship that can link two or more variables. In this section, we will introduce the simple linear regression model since the model contains only one explanatory (independent) variable. The regression line will be de fined by estimating the parameters of the regression model from the sample data by the ordinary least squares method.
Consider a sample of n observations

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots (x_n, y_n)$$

The simple linear regression model of Y on X, for all $i = 1, \ldots, n$, is written as

$$y_i = ax_i + b + \epsilon_i$$

where the ϵ_i are random quantities not observed that we call the *errors.* The ordinary least squares method consists to minimize the errors in order to find an estimate for the two parameters a and b. Then we obtain

$$\hat{a} = \frac{Cov(X, Y)}{V(X)} = \frac{\sigma_{xy}}{\sigma_x^2}$$

and

$$\hat{b} = \overline{y} - \hat{a}\overline{x}$$

The regression line of Y on X has an equation:

$$y = \hat{a}x + \hat{b}$$

In a similar way, we can also define the simple linear regression model of X on Y, given by the equation:

$$x_i = cy_i + d + u_i$$

where u_i are the errors of the model. The parameters of the model are estimated by:

$$\hat{c} = \frac{Cov(X,Y)}{V(Y)} = \frac{\sigma_{xy}}{\sigma_y^2}$$

and

$$\hat{d} = \overline{x} - \hat{c}\overline{y}$$

Regression Validation. The regression model is validated using the coefficient of determination R^2. This coefficient indicates the part of the variance explained by the model. The coefficient of determination is written in the form:

$$R^2 = (\rho(X,Y))^2$$

We have $0 \leq R^2 \leq 1$. An R^2 close to 1 means that the model is useful for the prediction.

Remark 21 *a) The regression line always passes through the center of gravity $(\overline{x}, \overline{y})$.*

b) The regression line of Y in term of X introduces the hypothesis that the values of Y depend on those of X, This means that, the knowledge of the values of X makes it possible to predict the values of Y.

c) The regression line of X in term of Y introduces the inverse hypothesis that the values of X depend on those of Y, This means that, the knowledge of the values of Y makes it possible to predict the values of X.

d) d) The sum of errors is equal to zero.

3.3.3 Coefficient of correlation and Regression

In the simple regression model of Y on X:

$$y = ax + b + \epsilon$$

The parameter a is estimated by

$$\hat{a} = \frac{\sigma_{xy}}{\sigma_x^2}$$

On the other hand, the coefficient of linear correlation is given by:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

By combining the two formulas, we obtain:

$$\hat{a} = \rho(X, Y)\frac{\sigma_y}{\sigma_x}$$

This means that the linear correlation coefficient and the slope of the regression line have the same sign.
Moreover, if $\hat{c}$ denotes the slope of the regression line of X on Y then:

$$\rho(X, Y) = \hat{a}\hat{c}$$

Remark 22 *a)* $\overline{ax + by} = a\overline{x} + b\overline{y}$

b) $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abcov(X, Y)$

c) $cov(aX + c, bY + d) = abcov(X, Y)$

3.4 Exercises

Exercise 12 Prove the following properties :

a) $\overline{ax+by} = a\overline{x} + b\overline{y}$

b) $V(aX+bY) = a^2V(X) + b^2V(Y) + 2abcov(X,Y)$

c) $cov(aX,bY) = abcov(X,Y)$

Exercise 13 If $D_1 : y = \hat{a}x + \hat{b}$ and $D_2 : x = \hat{c}y + \hat{d}$ represent the two linear regression lines of y on x and of x on y. Prove that

$$\rho^2 = \hat{a}\hat{c}$$

Exercise 14 The joint distribution of the relative frequencies of the two variables X and Y is given in the following contingency table:

y x	-1	0	2
1	2/9	1/9	2/9
3	1/9	2/9	1/9

a) Give the distributions of the marginal relative frequencies of X and Y; are these two variables independent?

b) Calculate $\overline{x}, \overline{y}, \sigma_x^2, \sigma_y^2, cov(x,y)$ and $\rho(x,y)$

Exercise 15 The following table gives the size T (in cm) and the weight P (in kg) of a sample of 12 pigs chosen at random from a farm:

T	70	63	72	60	66	70	74	65	62	67	65	68
P	155	150	180	135	156	168	178	160	132	145	139	152

a) Determine the equation of the regression line of the weight in terms of the size.

b) Plot the data points and the regression line

c) Conclude the weight of a pork that measures 63 cm ?

Exercise 16 The correlation coefficient of two variables X and Y is $\rho = 0.98$ and the mean of Y is $\overline{y} = 32$. In addition, the equation of the regression line of y on x is given by:

$$D_{y/x} : y = 10.17x - 3.6$$

a) Determine the equation of the regression line of x on y

b) Estimate x for $y = 100$, and estimate y for $x = 10$

Exercise 17 The joint distribution of two variables X and Y is given in the following table:

y x	2	5	8
$[0; 2[$	10	0	0
$[2; 4[$	2	5	0
$[4; 6[$	5	0	3

a) Determine the marginal distributions of X and Y

b) Calculate the marginal means and variances of X and Y

c) Calculate the linear correlation coefficient between X and Y. Interpret the result.

d) Determine the equation of the regression line of y on x

e) Give the estimation of x if $y = 10$

f) Calculate the coefficient of determination of this distribution. Comment the obtained result.

Exercise 18 The following table presents the results of the effect of the temperature of (x) to the efficiency (y) of a chemical process

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

We want to study the relationship between X and Y.

a) Write the equation that models this relationship

b) Calculate the least square estimates of this model

c) What is the efficiency of the process if the temperature is 0.5 ?

Exercise 19 We want to study the relation between the rent L (in dollars) of an apartment and its surface S (in m^2) in a certain city. We suppose that this relation have the following form:

$$L = aS + b + \varepsilon$$

a) What can you say about the term a ? Can you interpret the terms b and ε ?

b) What is the expected sign of a, and the one of the linear correlation coefficient r ?

c) With a sample of $n = 100$ apartments, we find $\rho = 0.82$, $\overline{L} = 722$ and $\overline{S} = 63$. In addition, the variances of L and S are respectively 276094 and 2140. calculate $\hat{a}$ and $\hat{b}$.

d) If you are looking for an apartment of $45m^2$, How much you expect to pay?

Exercise 20 In a laboratory, we observe the growth of 100 plants according to the quantity of water used. The distribution of the quantities X (in ml) and the number of the plant's leaves (Y) are given in the following table:

Y X	1	2	3	c
[100,200[	8	4	16	12
[200,400[	4	a	8	6
[400,600[	b	4	16	12

a) Determine the marginal distributions of X and Y in terms of a, b and c.

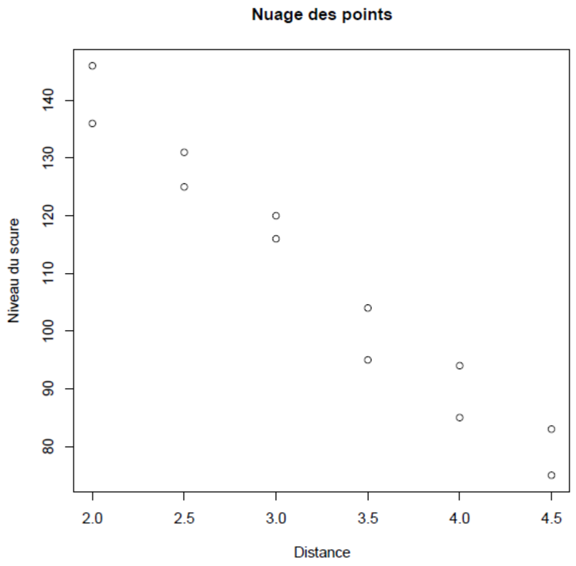
b) Find the values of a, b and c if the marginal mean of Y is 4 and the two variables are independent.

c) We suppose that $c = 4$, determine the conditional distribution of $Y/X \in [100, 200]$ and calculate its mean.

Exercise 21 We observe the sugar level (y) in the blood of a diabetic after a walk of a distance (x). 12 observations are registered in the following table:

x_i	2	2	2.5	2.5	3	3	3.5	3.5	4	4	4.5	4.5
y_i	136	146	131	125	120	116	104	95	85	94	83	75

The data points are plotted in the following figure:



In addition, a preliminary calculus gives:

$$\sum_{i=1}^{12} x_i = 39 \ , \ \sum_{i=1}^{12} y_i = 1310 \ , \ \sum_{i=1}^{12} x_i^2 = 135.5 \ , \ \sum_{i=1}^{12} y_i^2 = 148870 \ , \text{et} \ \sum_{i=1}^{12} x_i y_i = 4035.5$$

a) According to the plot, can you conclude a relationship between x and y? Justify your answer.

b) Write the equation that represents the relation between these two variables and calculate its coefficients.

c) Interpret the values of the coefficients obtained in b).

d) Calculate the correlation coefficient between x and y. Does the obtained value confirms your answer in a)? Comment.

e) Give an estimation of the sugar level after a walk of 3.1 km.

Exercise 22 The distribution of 25 married couples according to the age of the husband (X) and the age of the wife (Y) is given in the following table:

Y X	]12,20]	]20,25]	]25,30]
[20,25[	4	2	0
[25,30[	5	6	0
[30,35[	0	4	1
[35,40[	0	1	2

a) Determine the population, the sample, the variables and their types.

b) Give the marginal distribution of X and the one of Y.

c) Calculate the marginal mean and the marginal variance of Y.

d) Give the conditional distribution of $X/Y \in]20, 25]$. Calculate the its mean and variance.

e) Are these two variables independent? Justify your answer.

Exercise 23 A survey is done to 500 enterprises to determine the price of a certain material that should be bought. For each price, we observe the number of enterprises capable of buying the material.
In addition, the data points and their plot are given below:

Prix x	Nombre y
40	60
36	70
32	130
28	210
24	240
20	340
16	390
12	420
10	440
8	500

Table 3.9: Data table

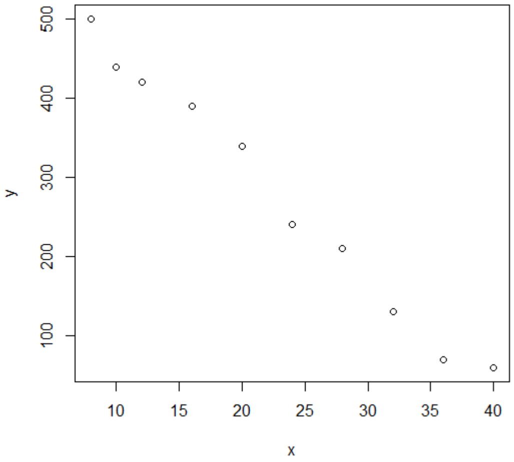


Table 3.10: Data points plots

We give the following information:

$$\sum_{i=1}^{10} x_i = 226 \ , \ \sum_{i=1}^{10} y_i = 2800 \ , \ \sum_{i=1}^{10} x_i^2 = 6244 \ , \ \sum_{i=1}^{10} y_i^2 = 1014800 \ , \text{et} \ \sum_{i=1}^{10} x_i y_i = 47200$$

a) According to the plot, can you conclude a relationship between x and y? Justify your answer.

b) Write the equation that represents the relation between these two variables and calculate its coefficients. Interpret their values.

c) Calculate the correlation coefficient between x and y. Does the obtained value confirms your answer in a)? Comment.

d) Give an estimation of the number of enterprises capable of buying this material if its costs 4000 dollars.

Exercise 24 In a city, we choose 48 women between 20 and 64 years old for a cooking contest. A committee had then tasted the dishes prepared by the women and had given a score between 1 and 6. The joint distribution of scores (X) and the woman's age (Y) is given in the following table:
In addition, we have:

$$\sum_i \sum_j n_{ij} x_i y_j = 6732.5$$

a) Determine the population, the sample and its size, the individual, the variables and their types.

b) Give the marginal distribution of X and the one of Y.

c) Calculate the marginal mean and the marginal variance of Y.

Y X	]20,30]	]30,35]	]35,40]	]40,50]	]50,65]
1	3	2	0	0	0
2	1	3	1	0	1
3	3	1	6	2	2
4	3	1	2	2	3
5	0	2	2	2	4
6	0	1	0	1	0

d) Give the conditional distribution of $X/Y \in]50, 65]$. Calculate the its mean and variance.

e) Are these two variables independent? Justify your answer.

Exercise 25 We observed the evolution of the Lebanese population (in million inhabitants) between 1960 and 2000 and we noted the following results:

Year X	1960	1970	1980	1990	2000
Population Y (in million inhabitants)	2.5	3	3.6	4.4	5.2

Let

$$Z_i = \frac{X_i - 1900}{10}$$

a) Plot the data points corresponding to (z, y) and interpret the obtained graph.

b) Calculate the linear correlation coefficient between z and y and interpret its value. Does the obtained value confirm the result in b)?

c) In this part, we will study the evolution of the population in terms of the year.

c.1) Write the equation of the linear regression line that describes Y in terms of Z.

c.2) Find the coefficients of this équation.

c.3) Deduce the equation of the line the describes Y in terms of the year X.

d) What is the prediction of the population in 2010?

e) In what year the population might exceed 15 millions inhabitants?

Exercise 26 The following table gives the distribution of 50 basketball players according to their size (X) and the number of blocks (Y) per game.

Y X	**0**	**1**	**2**	**3**	**Total**
[185; 190[	6	3	1	0	
[190; 195[	4	6	4	1	
[195; 200[	0	1	3	1	
[200; 205[	0	0	2	8	
[205; 215[	0	0	0	10	
Total					50

a) Determine the marginal distribution of X and of Y

b) Evaluate the mean and the variance of X and Y.

c) Determine the conditional distribution of $Y/(X \in [185; 190[)$ and $Y/(X \in [205; 215[)$. Deduce their means and comment the obtained results.

d) We suppose that the number of players blocks is independent of the size of the players. Complete the following table:

Y X	**0**	**1**	**2**	**3**	**Total**
[185; 190[					10
[190; 195[					15
[195; 200[				2	
[200; 205[			2		10
[205; 215[					
Total		10		20	50

Exercise 27 We summarize the size (in cm) of 20 plants during an experiment and we obtained the following :

2	3	405	5	5	6	8.5	9	9.5	11
13	13	15	15.5	16.5	17	18	19	19.5	19.5

In addition, we know that :

$$\sum_{i=1}^{20} x_i = 630 \text{ et } \sum_{i=1}^{20} x_i^2 = 167298.5$$

a) Determine the mean and the variance of this size series.

b) Determine the quartiles of this series.

We group the data into 4 classes : $[0,4[, [4,8[, [8,10[$ et $[10,20[$

c) Give the table of the corresponding frequencies.

d) Give the histogram and the polygon of the frequencies of this new distribution.

e) Calculate the mean, the variance and the quartiles of this new distribution.

f) Compare the obtained results with the ones obtained in a) and b).

Exercise 28 A) We observed the distance (Y in feet) that a car takes to stop. We observed 50 cars and obtained the following distribution :

distance	[0-20[	[20-40[	[40-80[	[80-140[
frequencies	8	18	n_3	n_4

a) Precise the population, the sample, the variable, its type and nature.

b) Calculate the values of n_3 and n_4 given that the mean distance is 47.2.

c) Calculate the variance of Y.

d) What is the number of cars that takes 30 to 60 feet to stop.

e) Plot the histogram of this distribution and calculate its mode.

f) Calculate the second decile.

B) In this part, we consider the same cars and we observe the speed (X) at which the car was moving before stopping. The joint distribution of 50 cars is :

X \ Y	[0,20[	[20,40[	[40,80[	[80,140[
[0,5[	2	0	0	0
[5,15[	6	12	2	1
[15,20[	0	5	9	1
[20,30[	0	1	7	4

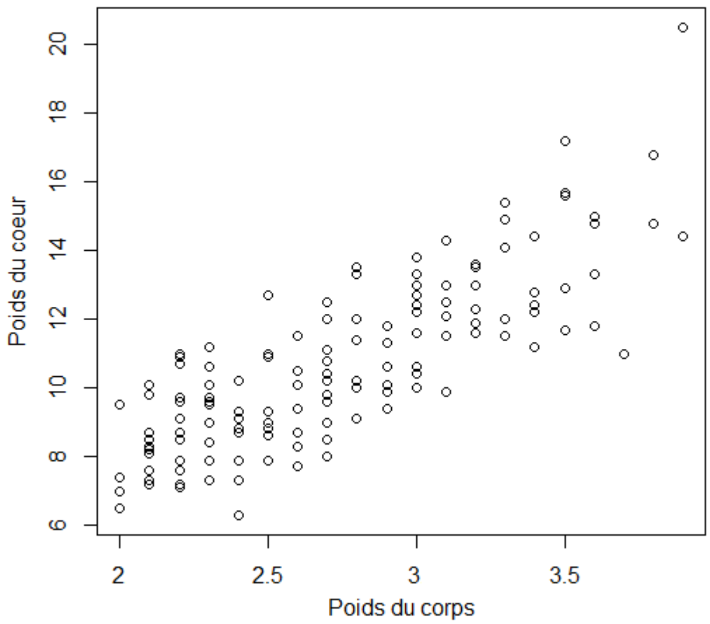
a) Determine the marginal distribution of X and calculate its mean.

b) Are these two variables independent? Justify your answer.

c) Calculate the covariance between X and Y given $\sum_i \sum_j n_{ij} x_i y_j = 42800$.

d) Determine the conditional distribution of $Y/X \in [15, 20]$ and calculate its mean.

Exercise 29 We choose 144 cats and we measure their weight (x in kg) and their heart's weight (y in g). The data points plot is given below :



In addition, we know that :

$$\sum_{i=1}^{144} x_i = 392.2\,,\ \sum_{i=1}^{144} y_i = 1530.8\,,\ \sum_{i=1}^{144} x_i^2 = 1101.88\,,\ \sum_{i=1}^{144} y_i^2 = 17120.88\,,\text{et}\ \sum_{i=1}^{144} x_i y_i = 4305.17$$

a) According to the plot, can you conclude a relationship between x and y? Justify your answer.

b) Calculate the correlation coefficient between x and y. Does the obtained value confirms your answer in a)? Comment.

c) We want to study the effect of the body weight to the heart's weight for the cats.

c.1) Write the linear relation that models y in terms of x.

c.2) Calculate the coefficients of this relation. Interpret the obtained results.

d) Estimate the weight of a cat's heart that weigh 2 kilos.

Exercise 30 The productivity Y a variety of trees is evaluated according to its population density X (in plants per m^2). The following table gives the values (x_i, y_i).

X	1.11	1.22	1.49	2.01	2.46	3	3.22	3.67	4.02
Y	1.73	1.49	1.1	0.7	0.52	0.39	0.31	0.2	0.17

a) Plot the data points and explain why we do not consider a linear adjustment.

b) We want to estimate the productivity in terms of the density using a relation of the form $Y = \alpha X^{\beta}$. In order to do this, we pose, $Z = \ln(X)$ et $T = \ln(Y)$. Some results are given below.

Z	0.104	0.199	0.399	0.698	0.900	1.099	1.169	1.300	1.391
T	0.548	0.399	0.095	-0.357	-0.654	-0.942	-1.171	-1.609	-1.772

De plus, $\sum z_i = 7.259, \sum t_i = -5.463, \sum z_i^2 = 7.706, \sum t_i^2 = 9.011, \sum z_i t_i = -7.624$.

i) Plot the data points that are associated with (Z, T). What do you remark?

ii) Determine to the nearest 1000 (10^{-3}), the linear correlation coefficient between Z et T.

iii) Estimate to the nearest 1000 (10^{-3}), the coefficients of the adjustment line of T on Z.

c) Deduce the relation that allow to calculate the productivity in terms of the density.

d) Can you predict the productivity if the density is 5 m^2?

Chapter 4

Combinatorial analysis

The objective of this chapter is to learn the techniques of combinatorial analysis or enumeration. These techniques are thus used for calculating probability.

4.1 Combinatorial analysis

It is a set of methods and techniques that consist in choosing, enumerating objects, counting the different ways of classification, grouping elements in one or more sets. This part of mathematics is also called ”enumeration”. It is widely used in probability and statistics, particularly in the enumeration of favorable cases and possible cases of the classical ratio of probability.
The combinatorial analysis is not the enumeration of all the possibilities but the enumeration of this one by a computation.
The combinatorial analysis allows to answer the following questions: ”*How many different numbers of 4 digits can be formed?*” or ”*In a class of 15 students, we have to elect two class delegates, how many different pairs are there?*”

4.1.1 Tools

Several tools make it possible to present and list the possible outcomes of an experiment. We distinguish: the table, the list and the tree of classification. The tables only show two-sequence of experiments (see Example 24). The list is a long tool to present the issues especially that we may forget elements or put them several times (see Example 25). A classification tree is a schema for describing and counting all possible outcomes of a given experiment (see Example 26).

Example 24 *the table 4.1 presents the possible outcomes when we launch successively two dice. There are* $6 \times 6 = 36$ *possible results.*

	1	2	3	4	5	6
1	(1;1)	(1;2)	(1;3)	(1;4)	(1;5)	(1;6)
2	(2;1)	(2;2)	(2;3)	(2;4)	(2;5)	(2;6)
3	(3;1)	(3;2)	(3;3)	(3;4)	(3;5)	(3;6)
4	(4;1)	(4;2)	(4;3)	(4;4)	(4;5)	(4;6)
5	(5;1)	(5;2)	(5;3)	(5;4)	(5;5)	(5;6)
6	(6;1)	(6;2)	(6;3)	(6;4)	(6;5)	(6;6)

Table 4.1: Table of outcomes by throwing two dice

Example 25 *The list of "words" that can be written with the letters A, B, C and D (without repetition) is:*

A B C D
A B D C
A C B D
A C D B
A D B C
A D C B
B A C D
B A D C
B C A D
B C D A
B D A C
B D C A
C A B D
C A D B
C B A D
C B D A
C D A B
C D B A
D A B C
D A C B
D B A C
D B C A
D C A B
D C B A

We can make with the letters A, B, C and D exactly 24 "words"

Example 26 *The Figure 4.1 shows the possible words that we can make with the letters A, B, C and D.*

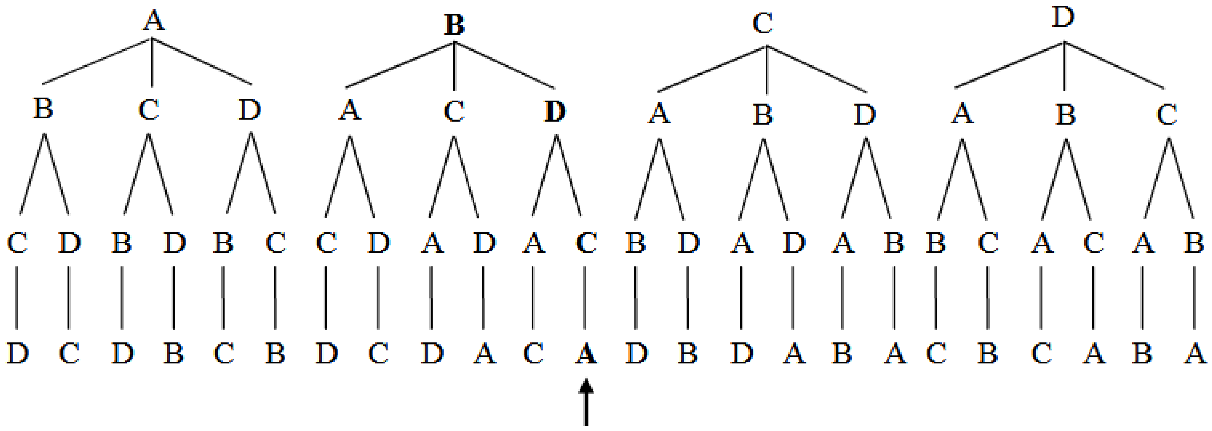

A
B
C
D
B C D A C D A B D A B C
C D B D B C C D A D A C B D A D A B B C A C A B
D C D B C B D C D A C A D B D A B A C B C A B A

Figure 4.1: classification tree

The tree reads vertically, for example, the arrow indicates the "word" B D C A. It can also be completed partially or schematically depending on the question that interests us.

we can make with the letters A, B, C and D exactly 24 ”words”. Note that the tree can be in horizontal form and is a safer tool than the list.

Remark 23 *The most often the trees are difficult to achieve. The model shown in Figure 4.2 will often be used to count the number of possibilities to fill them.*

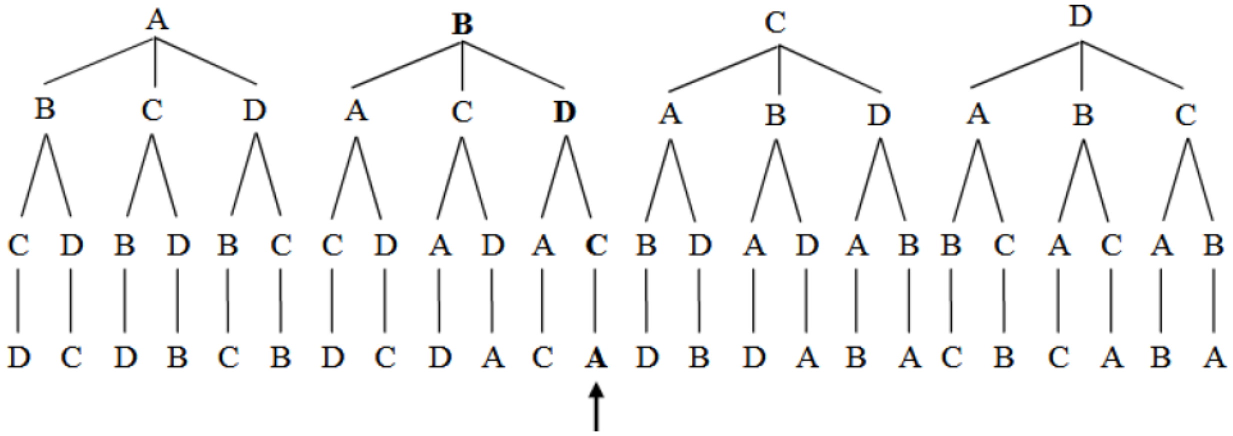

A B C D
B C D A C D A B D A B C
C D B D B C C D A D A C B D A D A B B C A C A B
D C D B C B D C D A C A D B D A B A C B C A B A

Figure 4.2: Model used in enumeration

4.1.2 The multiplication principle

If a first operation can be performed in n_1 different ways, then a second operation can be performed in n_2 different ways, then a third operation can be performed in n_3 different ways and so on until a k-th operation that can be performed in n_k different ways. So all of these operations can be can be performed in

$$n_1 \times n_2 \times n_3 \times \cdots \times n_k$$

different ways. .

4.1.3 The addition principle

If a first operation can be performed in n_1 different ways, then a second operation can be performed in n_2 different ways, then a third operation can be performed in n_3 different ways and so on until a k-th operation that can be performed in n_k different ways. At least one of these operations can be performed in

$$n_1 + n_2 + n_3 + \cdots + n_k$$

different ways if the operations can not be performed simultaneously. subsectionFactorial concept

Definition 15 *For any integer $n > 0$, we call factorial n, denoted $n!$, the product of all positive integers less than or equal to n:*

$$n! = 1 \times 2 \ldots \times n$$

The value of $0!$ *is* 1*, according to the convention for an empty product.*

4.2 Enumeration methods

There are three types of enumeration: permutations, arrangements and combinations.

4.2.1 Permutations

Definition 16 *A* ***permutation*** *is an ordering of all the objects of a set of objects.*

Definition 17 *The number of* **permutations without repetition** *of n different objects is given by $n!$*

Example 27 *The number of permutations of letters of word BLUE is* $4! = 24$.

Remark 24 *The number of permutations of n objects around a roundtable is* $(n-1)!$.

Definition 18 *Given a set of n objects such that there are n_1 identical objects of type* 1*, n_2 identical objects of type* 2*,... , and n_k identical objects of type k,then the number of* ***permutations with repetition*** *is*

$$\frac{n!}{n_1! \times \ldots \times n_k!}$$

with $n_1 + n_2 + \ldots + n_k = n$.

Example 28 *The number of words that can be formed with the letters of the word statistic is* 15120.

4.2.2 Arrangements

Definition 19 *An* ***arrangement with repetition*** *An arrangement with repetition is an ordering of k objects not necessarily distinct selected from n. The number of arrangements with repetition is given by n^k.*

Example 29 *The number of different sequences that can be read on a* 6*-digit car odometer is* 10^6.

Definition 20 *An* ***arrangement without arrangement*** *is an ordering of k distinct elements selected from n. The number of arrangement of k objects among n, denoted A_n^k, is given by*

$$A_n^k = \frac{n!}{(n-k)!}, \qquad (1 \leq k \leq n)$$

Example 30 *After the extra time of a football game the number of ways to choose the* 5 *penalty shooters among the* 11 *players is $A_5^{11} = 55440$.*

4.2.3 Combination

Definition 21 *A* ***combination without repetition*** *of n different elements taken k in k are the different groups of k elements not necessarily distinct that can be formed by these n elements, so that two groups differ only if they have different elements (*that is to say, the order does not matter*). the number of combinations with repetition of k objects among n, denoted by C_n^k, is given by*

$$C_n^k = \frac{n!}{(n-k)!k!}, \qquad (1 \leq k \leq n)$$

Example 31 *the number of ways of choosing 3 winners from 10 contestants is $C_{10}^3 = \frac{10!}{3!7!} = 120$.*

Definition 22 *A* ***combination with repetition*** *of n elements taken k in k are the different groups of k elements that can be formed by these n elements, so that two groups differ only if they have different elements (*that is to say, the order does not matter*). the number of combinations without repetition of k objects among n, is given by:*

$$C_{(n+k-1)}^k = \frac{(n+k-1)!}{(n-1)!k!}$$

Example 32 *The number of groups of* 3 *letters, with repetition, that can be formed with the* 4 *letters a,b, c and d is $C_{(4+3-1)}^3 = C_6^3 = 20$.*

Remark 25 *a) An object is observed only once in an arrangement without repetition.*

b) An object can be observed several times in an arrangement with repetition.

c) A combination is characterized only by the choice of objects.

d) A combination is not characterized by the order of the objects.

e) The model in Figure 4.2 can not be used because it induces an orderly choice, which is not the case in combinations.

f) The number of ways of choosing k among n (without order) is the number of ways of choosing $n-k$ among n.

g) Simultaneously: there is neither order nor repetition. Simultaneously: the possible drawings are combination of K of n objects.

h) Successively without the replacement (repetition) of the drawn object before the next drawing: there is an order but there is no repetition. Successively without replacement: the possible drawings are arrangements of k of n objects (if $k=n$, that is; if you draw all of the objects, they will be permutations of n objects).

i) Successively with replacement (repetition): there is an order and a repetition. Successively with replacement: the possible drawings are k-lists of k-uplets formed with n objects; their number is n^p.

4.3 Exercises: Enumerations

Exercise 31 In a horse racing course of 24, how many ways the horses can be placed?

Exercise 32 In an exam, a student has to answer 5 out of 8 questions.

a) How many choices he has?

b) Same question but he must answer the first three questions?

c) Same question but he must answer 4 out of the first 5 questions?

Exercise 33 What is the capacity of a telephone network whose numbers consist of 8 digits?

Exercise 34 A private television decides to opt for the system of "pay-program" using the decoders ordered by 8-digit codes.

a) Find the number of potential subscribers and the number of subscribers with code consisting of 8 different digits.

b) Calculate the number of codes with 2 different digits; one used 1 time and the other 7 times.

c) Same question with 3 different digits, 2 of them used 1 time and the 3^rd used 6 times.

Exercise 35 a) How many ways can we choose a committee of 3 persons from 20?

b) Suppose that the committee consists of a president, a secretary and a cashier. How many ways we can form this committee?

Exercise 36 a) How many ways can 5 persons sit in a row?

b) Same question if 2 of these 5 people have to sit beside each other.

c) Solve the 2 previous questions if the persons are sitting in a circle.

Exercise 37 Give the number of words of nine letters obtained with sets of letters: AEIOLNRST and EEEEBRRRV.

Exercise 38 Twelve persons have at their disposal 3 cars of 6, 4 and 2 seats respectively. How many ways can we assign these 12 people to three cars, supposed:

a) Any one can drive?

b) Only 4 of 12 can drive?

Exercise 39 The twelve volumes of a medical encyclopedia that are placed randomly.

a) How many ways can be placed?

b) How many ways can be placed if the volume 1 and 2 still beside each other, with this order.

Exercise 40 a) How many ways can we form a committee of 4 women and 5 men chosen from 8 women and 7 men?

b) How many forms of committees we have if a couple refuses to sit be together?

c) How many forms of committees we have if 2 women refuse to sit together?

Exercise 41 With the letters of the word FIBER how many words of 5 letters can be formed, that makes sense or not:

a) Overall.

b) Beginning and ending with a vowel.

c) Beginning and ending in a consonant.

d) Beginning with a vowel and ending in a consonant.

e) Beginning with a consonant and ending with a vowel.

Exercise 42 How many pieces in a game of dominoes knowing that on each side there are two numbers between 0 and 6.

Exercise 43 A group of five persons chosen from among 10 women and 5 men.

a) How many ways can we choose this group with 3 women and 2 men?

b) How many ways can we choose a group with the same sex?

Exercise 44 A keyboard of 11 keys $\{1; 2; 3; 4; 5; 6; A; B; C; D; E\}$ is used to dial the access code for a building , using a letter followed by a number from 3 digits (different or not).

a) How many different codes can be formed?

b) How many codes can be formed beginning with a vowel?

Exercise 45 A company sends 23 employees to an office of 13 people in China, and two offices of five people each, one in France and the other in USA. How many groups could it be formed?

Exercise 46 Ten children are to be divided into an A team and a B team of 5 each. The A team will play in one league and the B team in another. How many different divisions are possible?

Exercise 47 In a digital computer, a bit is one of the integers $\{0, 1\}$, and a word is any string of 32 bits. How many different words are possible?

Exercise 48 a) In how many ways can 12 teachers be divided among 3 schools if each school must receive 4 teachers?

b) In how many ways can this be done if two of the teachers, Rana and May, want to be together in the same school?

Exercise 49 What is the number of solutions of the equation $x_1 + x_2 + \cdots + x_{12} = 5$, with each $x_i = 0$ or 1?