



الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة وهران 2 محمد بن أحمد
University of Oran 2 Mohamed Ben Ahmed
كلية العلوم الاقتصادية، التجارية و علوم التسيير



Faculty of Economics Sciences, Business and Management Sciences

Faculty of Economics Sciences, Business and Management Sciences

كلية العلوم الاقتصادية، التجارية و علوم التسيير

Domain : Economics Sciences, Business and Management Sciences

Specialization : Common Core

COURSE HANDOUT

Subject :

STATISTICS I (M 111) :

*****LECTURES AND APPLICATION EXERCISES*****

Course for bachelor's degree students : First Year (L1)
Semester 01 (S1)

Prepared by : Dr. **SAIM Tahar**

University of Oran 2 Mohamed Ben Ahmed

Email: tahar.sam@hotmail.fr

saim.tahar@univ-oran2.dz

Academic Year :
2024 - 2025

Contact Sheet

- **Teaching Assistant for Tutorials** : Mr. SAIM Tahar
- **Rank** : MCB
- **Field** : Economic Sciences, Commerce, and Management Sciences
- **Specialization** : Common Core
- **Level** : First Year (L1)
- **Course Title** : Statistics I
- **Course Code** : F111
- **Semester** : S1
- **Target Audience** : This course is designed for first-year LICENCE (L1) students in the Common Core track.
- **Coefficient** : 2
- **Credits** : 4
- **Teaching Unit** : Methodological
- **Total Contact Hours** : 45 hours (15 weeks)
- **Weekly Contact Hours** : 2 hours (lecture) + 1.5 hours (tutorials)
- **Assessment Method** : 60% Final Exam, 40% Continuous Assessment
- **Contact Information** : tahar.sam@hotmail.fr

Prerequisites

To maximize the benefits of this course, students are expected to have prior knowledge of the following concepts:

- **Functions of a Real Variable** (Analysis I),
- **Integral Calculus and Differential Equations** (Analysis II),
- **Polynomials and Vector Spaces** (Algebra I).

General Objectives (Targeted Competencies)

The primary goal of this course is to provide students with foundational knowledge in descriptive statistics, specifically:

- Understanding the distinction between descriptive statistics and inferential statistics.
- Recognizing the importance of descriptive statistics in any preliminary data analysis.
- Graphically exploring data to characterize it and identify potential issues.
- Using descriptive statistics to summarize and describe data effectively.
- Performing the appropriate statistical test based on the objectives of their study.
- Processing statistical data more efficiently and effortlessly.
- Interpreting results with greater confidence and reliability.

Course evaluation criteria

The final grade for the module is composed of two components:

1. Continuous Assessment Grade (40% weight):

- Based on written tests, individual assignments, attendance, and participation.

2. Final Exam Grade (60% weight):

- Based on a written exam covering all units taught during the course.
- The exam includes various question types:
 - **Multiple-choice questions (MCQs)** to assess knowledge acquisition.
 - **Application exercises** to evaluate the practical application of acquired knowledge.
 - **Reflective questions** to test reasoning skills and the use of knowledge.

Passing criteria:

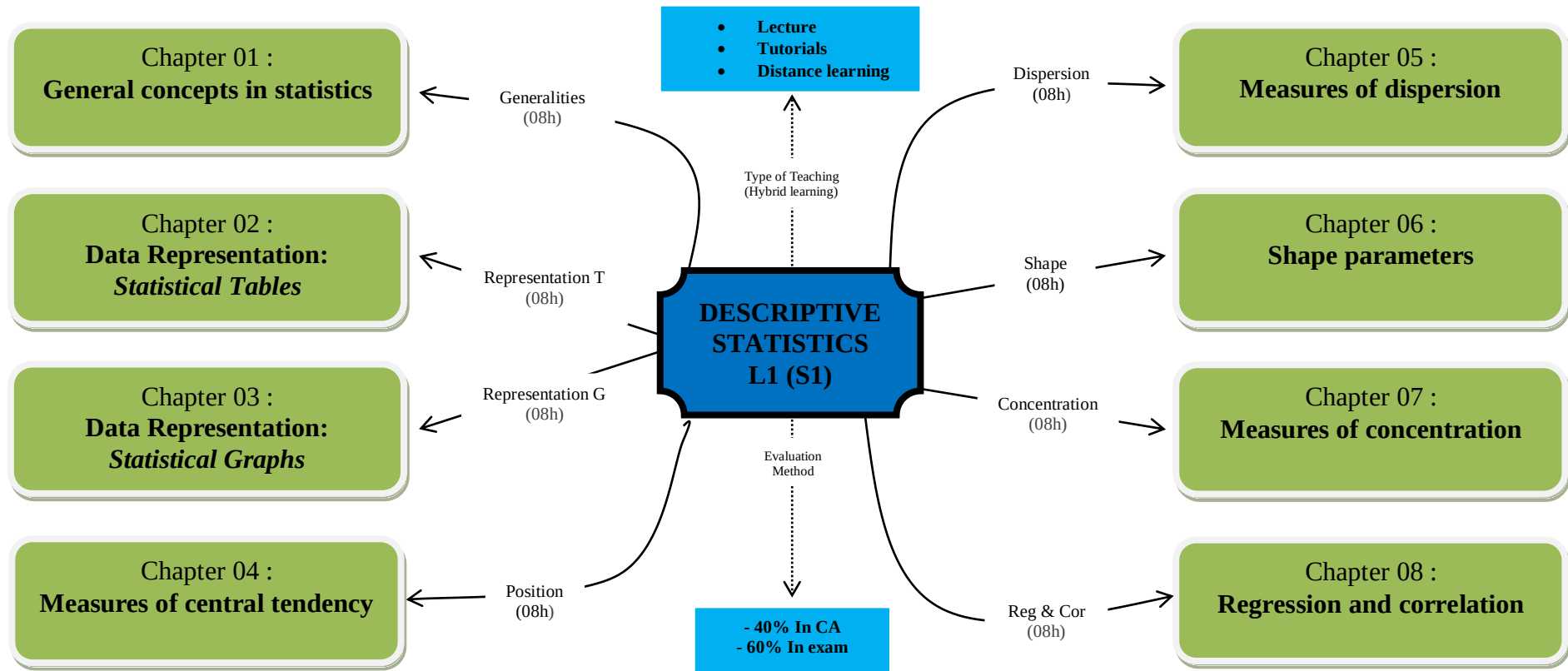
- A minimum overall grade of **10/20** is required to pass the module.

Assessment Type	Grade	%
1. Continuous Assessment (CA)	20/20	40%
- Written Test	8/8	16%
- Individual Assignment	7/7	14%
- Attendance	3/3	6%
- Participation	2/2	4%
2. Final Exam (FE)	20/20	60%

Calculation Formula for the Final Grade:

Final grade (M) = (Continuous Assessment Grade×0.4) + (Final Exam Grade×0.6).

Concept map of the course (C_M)



University of Oran 2 Mohamed Ben Ahmed
Faculty of Economic, Commercial, and Management Sciences

Department : Common Core
Level : First-Year Bachelor's Degree (L1)
Teaching Unit : Methodological
Credits : 04
Coefficient : 02
Weekly hourly volume :
*** Lecture: 02h00
*** Tutorial (TD): 1h30
*** Annual Course

Mr SAIM Tahar
(MCB)
Section : 03 + 04

Teaching Objectives :

- This concept map is designed for students in Economics, Commerce, and Management Sciences.
- To teach students the fundamental principles of statistics, such as sampling, frequencies, graphical representations, etc.
- To provide an overview of the key areas related to this discipline (see above).
- To enable students to apply statistical tools commonly used in economics, management, and related fields.

Table of contents

Chapter 01 : Introduction to Statistics	10
Chapter 02 : Data Representation: Statistical Tables	23
Chapter 03 : Data Representation: Statistical Graphs	35
Chapter 04 : Measures of Central Tendency	49
Chapter 05 : Measures of Dispersion	72
Chapter 06 : Measures of Shape	91
Chapter 07 : Measures of Concentration	101
Chapter 08 : Regression and Correlation	119
Chapter 09 : Applications	138

Symbols and Notations

Symbol	Meaning
[]	The integer part.
Card()	The cardinal: the number of elements in the set.
:=	Is defined as (assignment symbol).
N	Population.
Σn_i	Sample size.
n_i	Absolute frequencies.
f_i	Relative frequencies.
$f_i \%$	Frequencies in percentage.
$NCI \uparrow$	Ascending cumulative frequencies.
$NCI \downarrow$	Descending cumulative frequencies.
$FCI \uparrow$	Ascending cumulative relative frequencies.
$FCI \downarrow$	Descending cumulative relative frequencies.
$FCI \uparrow \%$	Ascending cumulative frequencies in percentage.
$FCI \downarrow \%$	Descending cumulative frequencies in percentage.
X_i	Values of variable X.
R	Set of real numbers.
R^2	Set of pairs of real numbers.
$\Sigma_{i=1}^n$	The sum for i ranging from 1 to n.
V.S	The statistical variable.
$X_{\max}$	Maximum value.
$X_{\min}$	Minimum value.
NC	Number of classes.
a_i	Amplitudes.
h_i	Corrected frequencies.
x_i	Class centers.
Mo	The mode.
Me	The median.
Q_i	Quartiles (Q1, Q2, Q3).
D_i	Deciles (D1 ... D9).
C_i	Centiles (C1 ... C99).
$\bar{X}$	The arithmetic mean of a statistical series X.
H_x	The harmonic mean of a statistical series X.
Q_x	The quadratic mean of a statistical series X.
G_x	The geometric mean of a statistical series X.
σ_x	The standard deviation of X.
Var(X)	The variance of X.
CV	Coefficient of variation.
E	Range.
I(O)	Interquartile range.
I(D)	Interdecile range.
I(C)	Intercentile range.
I(O)R	Relative interquartile range.
I(D)R	Relative interdecile range.
I(C)R	Relative intercentile range.
C_p	Skewness coefficient.
Cov(X,Y)	The covariance between variables X and Y.
$r(X,Y)$	The correlation coefficient between variables X and Y.
$r^2(X,Y)$	The coefficient of determination between variables X and Y.
F_x	The cumulative distribution function of characteristic X.

General Introduction

"Everything is number," declared Pythagoras nearly 25 centuries ago. In recent decades, a wealth of previously latent information has emerged, and businesses, more than any other institutions, have found themselves at the forefront of an abundance of available data. More often than not, it has seemed justifiable—and indeed beneficial—to exploit this information.

Improving the management of organizations, whether businesses or other establishments, requires a shift from a posteriori regulation to a priori regulation. This proactive regulation relies on a precise understanding of the phenomena one seeks to control. It is at this juncture that the statistician plays a crucial role, leveraging their expertise to make the best possible use of available information to extract meaningful insights from facts. For decision-makers, statistics is a tool that aids in enhancing their decision-making processes.

The use of statistics as a decision-support tool demonstrates, in practical terms, how to construct and apply statistical processes to achieve greater efficiency and to improve products, services, and systems of all kinds.

By mastering the use of statistics as a decision-support and quality improvement tool, students, technicians, managers, and professionals across various sectors will acquire practical instruments to tackle the numerous challenges they face effectively.

1

CHAPTER

General overview of statistics

-

In this chapter :

-History and definitions of statistics

-Basic concepts of statistics

- The population (N)
- The sample (n)
- The statistical unit
- The different types of statistical characteristics (Variables)

I.1 Introduction

The primary purpose of statistics is to extract meaningful insights from data, whether numerical or otherwise, collected during the study of a phenomenon. It is essential to distinguish between **statistical data**, which are the results of observations gathered during the study of a phenomenon, and the **statistical method**, which focuses on the rational analysis of this data. The statistical method encompasses several stages.

I.2 Historical Background and Definitions of Statistics

To illustrate the role of statistics in various domains (such as decision-making, among others), we begin with its definitions and scope.

Historically, the term "statistics" (derived from the Latin word **status**, meaning "state") was not commonly used before the 18th century. At that time, it referred to the collection of information, particularly quantitative data, necessary for political authorities to govern the state. However, it can be argued that statistics, at least in the form of enumeration, has been practiced by all civilizations.

The great empires of antiquity provide the earliest evidence of this practice. Similarly, centralized European monarchies of the 17th century, as part of the establishment of the modern state, shared the same interest in enumeration. In contemporary societies, statistical practices have become institutionalized, appearing commonplace today, and many sectors of human activity rely on statistics in one form or another.

Early on, the types of information deemed essential to the state were those required for collecting taxes and recruiting conscripts for war efforts. It is, therefore, unsurprising that the first data collected primarily concerned populations and economies.

Following its application in economics and demography, statistics expanded to encompass the entirety of the sciences. It has since evolved into a scientific discipline, drawing upon mathematics (algebraic calculations, analysis, probability theory) and computer science for practical applications.

I.2.1 Definitions of Statistics (What is Statistics?)

Statistics is simultaneously a formal science, a method, and a technique. It encompasses the processes of collecting, analyzing, interpreting, and presenting data to make it comprehensible and accessible.

As a mathematical discipline, statistics should not be confused with "a statistic," which is a numerical value derived from observations. In various fields, the term "statistics" (plural) commonly refers to collections of numerical data presented in tables or graphs; these figures represent observations related to numerous phenomena.

In its singular form, "statistics" refers to the set of methods used for the numerical study of large-scale phenomena (statistics as a science).

The term "statistics" also refers to the outcomes of analyses based on statistical methods. This activity is organized into three primary branches:

- **Data Collection:** Gathering raw information.
- **Data Processing:** Also known as **descriptive statistics**, which involves summarizing and organizing collected data.
- **Data Interpretation:** Also referred to as **statistical inference**, which relies on sampling theory and mathematical statistics.

The primary goals of these activities are:

- **To inform;**
- **To support decision-making.**

It is important to note that these branches are not distinct or independent domains. Data processing and interpretation can only occur after data collection. Conversely, mathematical statistics provides the rules and methods for data collection to ensure the data can be properly analyzed and interpreted.

Therefore, statistics as an analytical method comprises two levels:

1. **Descriptive statistics;**
2. **Theoretical or mathematical statistics.**

I.2.1.1 Descriptive Statistics

Descriptive statistics encompass the various techniques and methods used to describe a large dataset with appropriate tools, focusing on extracting the most relevant information. It employs graphical representations and statistical measures, such as central tendency indicators (mean, median, mode) and dispersion indicators (variance, standard deviation).

Descriptive statistics can be summarized as follows:

Key Concepts in Descriptive Statistics

- **Variables:** The basic elements of statistical analysis.
 - **Types of Variables:**
 - **Discrete Variables:** Variables that take distinct, separate values.
 - **Continuous Variables:** Variables that can take any value within a range.
- **Summarizing a Distribution:**
 - Statistics employs two main approaches for summarization: graphical representation and numerical methods.
- **Presenting Statistical Series:**
 - Use of tables or graphs.
 - Frequency and count distributions, including cumulative frequencies and counts.
 - Grouping data into categories or intervals for simplification.

Key Measures in Descriptive Statistics

1. **Central Tendency Indicators:**
 - Examples: Mode, median, quartiles, mean, etc.
2. **Dispersion Indicators:**
 - Examples: Variance, standard deviation, coefficient of variation, range, interquartile range.
 - **Shape Measures:** Include skewness (symmetry of distribution) and kurtosis (peakedness of distribution).

Analyzing Relationships Between Variables

- **Continuous Variables:**
 - Use of correlation coefficients and scatterplots.
 - Regression functions to make predictions and estimations.
- **Discrete Variables:**
 - Use of contingency tables to study associations.

Forecasting and Time Series Analysis

- Statistical forecasting often relies on analyzing chronological data (time series) to identify trends and make predictions.

I.1.1.2 Theoretical or Mathematical Statistics

Theoretical statistics extends descriptive statistics by enabling the formulation of laws based on the observation of samples, i.e., limited subsets drawn from a population. It is particularly relevant in surveys and sampling studies. This branch relies on both descriptive statistics and probability theory to derive its conclusions.

From our perspective, statistics can be conceptualized based on the following definition:

Statistics is the science that systematically studies, through mathematical modeling, the methods of utilizing and processing data (or information), with the goal of guiding and supporting reasoning or decision-making in practical situations subject to uncertainty.

At the first level, **descriptive statistics** examines these methods to produce primarily descriptions of information. At the second level, **inferential statistics** (or differential statistics) seeks to extend these described data to unobserved domains of validity, ideally with risk control over the inductive reasoning involved.

This technique finds applications across a wide variety of fields.

I.2.2 Fields of Application

In 1963, W. Weaver stated, "**Statistics and probability theory are two important domains integrated into our daily activities. The industrial world and insurance companies heavily rely on probabilistic laws. Even physics is fundamentally probabilistic in nature.**"

Statistics is applied in numerous fields, including:

- **Geophysics:** For weather forecasting, climatology, pollution studies, and analyses of rivers and oceans.
- **Demography:** Population censuses provide a snapshot of a population at a given time, enabling representative sampling for further studies.
- **Economics and Social Sciences (Econometrics):** Statistical methods are used to study the behavior of population groups or economic sectors. Environmental issues also rely heavily on statistical data.
- **Sociology:** Statistical data serves as primary material for surveys, while statistical methods are applied as tools for data processing.
- **Marketing:** Opinion surveys are crucial tools for decision-making and investment strategies.
- **Physics:** Statistical mechanics and thermodynamics (statistical physics) are used to deduce global behavior (macroscopic) from the behavior of individual particles (microscopic).
- **Metrology:** Applied in the study of measurement devices and the accuracy of measurements.
- **Medicine and Psychology:** For understanding disease behaviors, their frequency, and the validity of treatments or diagnostic methods.
- **Insurance and Finance:** Risk assessment and probabilistic modeling.
- **Others:** Various other domains incorporate statistical methods for analysis and decision-making.

I.2.3 The Statistical Process

I.2.3.1 Data Collection

Statistical surveys always begin with a preliminary phase where the variables to be studied are identified.

The next step is selecting the population to study. This involves addressing the issue of sampling, which includes:

- Choosing the population to be surveyed.

- Determining the sample size.
- Ensuring the sample's representativeness.

Preprocessing the data is a crucial step. Transformations of the initial data (e.g., taking logarithms) can significantly enhance the efficiency and accuracy of subsequent statistical analyses.

I.2.3.2 Data Processing

The results of a statistical survey consist of numerical data (e.g., heights, salaries) or qualitative data (e.g., spoken languages, preferred brands). To make such data exploitable, it is essential to classify and summarize it visually or numerically. In some cases, data compression may also be required. This is the domain of **descriptive statistics**.

The approach to data processing differs depending on whether the study focuses on a single variable or multiple variables.

Study of a Single Variable

- **Grouping and Summarizing Data:** Grouping data, calculating frequencies, and constructing graphs provide an initial visual summary of the studied statistical characteristic. For quantitative continuous variables, the histogram is the most common graphical representation.

- **Statistical Reductions:**
The numerical values of a statistical characteristic are distributed across many observations, making it challenging to analyze or interpret several hundred or thousands of values. To address this, statistical reductions are used to calculate key values that summarize the data. These include central tendency measures (e.g., mean, median) and dispersion measures (e.g., variance, standard deviation). Some measures, such as variance, are specifically designed for theoretical exploitation of the data.

Study of Multiple Variables

- **Two Variables:** The analysis of two variables often involves creating a scatterplot to visualize their relationship. Mathematical tools, such as correlation analysis or linear regression, may be applied to study the potential link between the two phenomena.

- **More than Two Variables :**

For studies involving more than two variables, advanced techniques such as **multidimensional analysis** are employed. These include:

- Principal Component Analysis (PCA);
- Independent Component Analysis (ICA);
- Multiple Linear Regression.

These methods leverage statistics to uncover relationships within large datasets. For forecasting and predictions, **time series analysis** is commonly used.

Summary of Statistical Objectives

The primary aim of statistics is to:

- Make **estimations**;
- Enable **comparisons**;
- Provide **forecasts**.

In a general sense, statistics is a scientific discipline that encompasses methods for collecting, presenting, summarizing, and analyzing data. However, it is not merely the systematic application of mathematical formulas. A high level of mathematical expertise is not required to understand statistics; what is essential is logical reasoning and the ability to use and interpret statistical tools effectively.

Key Principles of Statistics

1. **Learning the Statistical Tool:**

This is the phase of acquiring and assimilating fundamental statistical concepts.

2. **Applying the Statistical Tool:**

This phase involves intelligently applying statistical concepts, such as collecting information, constructing tables, creating graphical representations, and calculating parameters.

3. **Interpreting Results:**

This is the reflection phase, where the degree of mastery of statistical tools is measured by the ability to explain, comment on, and analyze the measured phenomena to draw conclusions and make informed decisions.

I.3 Basic Concepts of Statistics

To understand statistics, it is essential to begin by defining the terms used to describe numerical observations.

I.3.1 Populations and Statistical Units

In statistics, we work with **populations**. The term "population" originates from demography, the study of human populations, which played a central role in the early development of statistics, particularly through population censuses. However, in statistics, the term applies to any group being studied, such as university students, households, computers in a company, or any other set of entities under observation.

A population consists of **individuals**, also referred to as **statistical units**. For example:

- If the statistical study is conducted on students in a lecture hall, the **population** is the group of students, and each student is a **statistical unit**.

Statistics studies the characteristics of these individuals. However, it does not focus on individuals themselves but rather on how they contribute to understanding the population as a whole. **Descriptive statistics**, as the name suggests, aims to describe a given population.

I.3.2 Characteristics and Variables

I.3.2.1 Definition

To study a population, the statistician focuses only on the characteristics of interest, where a characteristic is a variable that describes the individuals within the population.

A characteristic is a property possessed by statistical units that allows them to be described and distinguished from one another. Each statistical unit can be analyzed based on one or more characteristics.

For example, if one is studying the population of students in a lecture hall, the analysis can be conducted from various perspectives:

- **Demographic:** By examining the age of the students.
- **Economic:** By focusing on the income of the students.
- **Sociological:** By investigating their leisure activities.
- **Anthropometric:** By analyzing their height.
- Or from any other relevant perspective.

In each example, a different characteristic is being studied: age, income, leisure activities, or height.

Within a given population, a characteristic may vary between individuals. This variability is referred to as the **modalities** of the characteristic.

Example :

- For the population of students in a lecture hall, if the characteristic studied is **age**, the modalities of the characteristic would be 18 years, 19 years, 20 years, etc.
- For a population of cars, if the characteristic studied is **color**, the modalities would be blue, green, white, etc.

The term **statistical variable** is often used interchangeably with **characteristic**, and the modalities of the characteristic represent the values taken by the variable.

I.3.2.2 Modalities

Modalities are the different categories or possible states of a characteristic. They represent the various situations, cases, or values a characteristic can take. A characteristic may have one or several modalities.

- For qualitative characteristics, modalities represent distinct categories (e.g., male/female for gender, or artisan/worker for profession).
- For quantitative variables, modalities are numerical values (e.g., 18, 19, 20 for age, or 50kg, 60kg, 70kg for weight).

I.3.2.3 Qualitative and Quantitative Characteristics

Characteristics can be broadly divided into **qualitative** and **quantitative** categories:

1. Quantitative Characteristics:

- These are measurable characteristics where numerical calculations are possible.
- Examples: Age, weight, height.
- Calculations such as average weight, maximum height, or minimum age can be performed.
- Modalities of quantitative characteristics are inherently measurable.

2. **Qualitative Characteristics:**

- These are non-measurable characteristics where calculations are not meaningful.
- Example: Gender (male/female).
- While one can count the number of males and females in a population, the concept of an "average gender" is nonsensical.

To differentiate the two types:

- **Qualitative characteristics** are referred to as **characteristics**.
- **Quantitative characteristics** are referred to as **variables**.

Types of Qualitative Characteristics

1. **Nominal Characteristics:**

- These are qualitative characteristics whose modalities cannot be ordered.
- Example: Socioprofessional categories (artisan, worker, etc.) or marital status (single, widowed, etc.).

2. **Ordinal Characteristics:**

- These are qualitative characteristics whose modalities can be ordered or ranked in some way.
- Example: Educational level (primary, secondary, higher education).

Example 1 :

A satisfaction survey asks consumers to evaluate a service by selecting one of the following six categories:

(a) Poor, (b) Mediocre, (c) Average, (d) Fairly Good, (e) Very Good, (f) Excellent.

These are **ordinal modalities** because they can be ranked: an "excellent" service is better than a "good" service, and so on. However, unlike quantitative values, it is not possible to say, for instance, whether a service rated as "excellent" is twice or four times better than a service described as "average." While a ranking can be established, quantification is not feasible.

Note: Some ordinal modalities can be converted into quantitative values. These are essentially quantitative values that appear as ordinal qualitative modalities.

Example 2 :

Shirts are classified by size: XS, S, M, L, XL, XXL, XXXL. These are **falsely ordinal modalities** because a correspondence table exists to define the actual size in centimeters for each category.

Nominal Modalities

Qualitative modalities that cannot be ranked or ordered are called **nominal modalities**.

Example 3 :

A group of individuals is asked what a specific perfume evokes for them. They must choose one of the following options:

(a) Adventure, (b) Sensuality, (c) Comfort, (d) Nostalgia.

It is clear that no comparison or ranking can be established among these modalities. They are **nominal modalities**.

Note: Some purely nominal modalities are occasionally encoded using numbers. For example, the gender of individuals in a population may be coded as "1" for men and "2" for women. This is an attempt to quantify a purely nominal variable. Such variables are referred to as **pseudo-numeric variables**. Using this encoding, it is possible to calculate a mean, which would represent the proportion of men in the population or sample.

I.3.2.4 Discrete and Continuous Characteristics

Based on the form of the variable's values, quantitative characteristics can be categorized as either **discrete** or **continuous**.

a) Discrete Characteristics

Discrete (or discontinuous) characteristics take only isolated values and have a finite or countable number of modalities. These values may or may not be integers.

Examples :

- The number of pages in a book.
- The number of people in a family.
- The number of children per household (e.g., 0, 1, 2, 3, ...).

It is impossible for these variables to take values between integers (e.g., between 0 and 1, or 1 and 2). Therefore, these are discrete characteristics.

b) Continuous Characteristics

Continuous characteristics have an infinite number of modalities and can take any value within a given interval.

Examples:

- Height.
- Weight.
- Household income or revenue (e.g., 2,900,000.1 DA, 2,900,000.12 DA, ...).

Continuous characteristics allow for infinitely precise values, making them distinct from discrete variables.

2

CHAPTER

Representation of data : Statistical tables

Objec

.....

In this chapter :

- Frequencies and counts

- The sum operator
- The product operator

- Presentation of statistical data in table form

- Distribution tables of counts and different types of frequencies
- Distribution table of relative frequencies and percentages
- Distribution table of cumulative frequencies

II.1 Introduction

Organizing data into a table often helps in understanding the phenomenon under study. However, in some cases, tables alone are insufficient to provide a clear picture. For this reason, another method of presentation is used: graphs. Graphs offer a quick overview of the phenomenon and allow for the visualization of key characteristics at a glance, though some information may be lost in the process.

The choice of representation method for a statistical variable depends on the nature of the variable (qualitative or quantitative). The most commonly used and recommended representations are tables and diagrams (graphs). The type of graph used varies according to the nature of the characteristic being studied.

II.2 Frequencies and Counts

The **total count** refers to the number of individuals in the statistical population under study. The total count is denoted by N .

Example

Consider a group of thirty students and observe their ages. The total count of the statistical population under study is thirty ($N=30$).

II.2.1 Definition and Notation

The count of a modality x_{ix} of a characteristic x is the number of individuals exhibiting this modality. The count corresponding to the i -th modality of the characteristic x is denoted as n_{ix} .

Example:

Let us consider a group of thirty-five ($N=35$) students. A table can be constructed to organize the information about their ages, with the first column representing the observed ages and the second column showing the count for each age.

Age (students)	Count (mi)
18	5
19	4
20	10
21	11
22	3
23	1
24	1
Total	35

In general, for a variable with k modalities, the total count N is equal to the sum of the counts for each modality. This can be written as $n_1, n_2, \dots, n_k$, where:

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

II.2.2 The Summation Operator

To express the sum of elements compactly, the **summation operator**, represented by the uppercase Greek letter "Sigma" « Σ », is used.

For simplicity, the sum can be written as:

$$n_1 ; n_2 ; \dots ; n_k \text{ per } \sum_{i=1}^k n_i$$

This notation is read as "the sum of n_i for i varying from 1 to k ."

General form:

$$\sum_{i=1}^n a_i$$

This is read as "the sum of a_i for i varying from 1 to n " and indicates that the values of a_i are added while incrementing i by 1, starting from the lower bound ($i=1$) and ending at the upper bound ($i=n$).

For example:

$$\sum_{i=1}^n a_i = a_1, a_2, \dots, a_n$$

Remark:

The summation operator (Σ Sigma) is a convenient notation used to simplify expressions involving sums. For instance, instead of writing:

$$N = n_1 + n_2 + \dots + n_k$$

it is more concise to write:

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

When dealing with counts, the total count for a variable with k modalities can be expressed as:

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

This notation is especially useful in statistical analysis to represent and compute sums efficiently.

II.2.3 The Product Operator

To express a product of elements compactly, the **product operator**, symbolized by the uppercase Greek letter Pi « $\prod$ », is used.

Example:

Consider four values of a variable xxx, indexed by iii: x_1, x_2, x_3, x_4 . The product of these four values is expressed as:

$$\prod_{i=1}^4 x_i = x_1 \times x_2 \times x_3 \times x_4.$$

The expression on the left is read as "the product of x_i for i ranging from 1 to 4.

II.3 Presentation of Statistical Data in Tabular Form

Raw statistical data, in their unprocessed form, are often incomprehensible and practically unusable. To make them meaningful and useful, they must be organized, classified, and analyzed using tables and graphs as supports.

II.3.1 Frequency Distribution Tables and Types of Frequencies

Statistical tables serve as tools for classifying and presenting the units of a statistical population.

For a statistical population N , described by a characteristic with k modalities, presenting the data in a table involves counting the **frequency** (number of statistical units) corresponding to each modality.

A frequency table can be constructed regardless of the nature of the characteristic (qualitative or quantitative).

Structure of a Frequency Distribution Table

Data arranged and classified in a table form a **distribution**. A **frequency distribution table** consists of two columns:

Modalities / Character Values		Frequencies (n_i)
m_1	x_1	N_1
m_2	x_2	N_2
m_3	x_3	N_3
.	.	.
.	.	.
.	.	.
M_k	K_k	N_k
Total		$N = \sum_{i=1}^p n_i$

Example 1: Frequency Distribution of Exam Grades

The following statistical series represents the grades obtained by 20 students in an exam:

D-B-E-C-D-B-D-C-E-A-B-E-C-D-B-D-D-A-E-C

- **Population:** The population consists of 20 students ($N=20$).
- **Nature of the Characteristic:** Qualitative ordinal, as the grades can be ranked.
- **Number of Modalities:** 5 modalities (A,B,C,D,E).
- **Frequencies (n_{in_ini}):** The number of students who obtained each grade.

Modalities (C_i)	Frequencies (n_i)
$C_1=A$	$n_1=2$
$C_2=B$	$n_2=4$
$C_3=C$	$n_3=4$
$C_4=D$	$n_4=6$
$C_5=E$	$n_5=4$
Total	$N = \sum_{i=1}^P n_i = 15$

Example 2: Frequency Distribution of Students per Class

The following statistical series represents the number of students per class in a primary school :

30, 30, 25, 32, 26, 30, 26, 26, 28, 30, 30, 27, 26, 27, 25

- **Population:** The population consists of 15 classes ($N=15$).
- **Nature of the Characteristic:** Quantitative discrete, as the number of students is countable.
- **Number of Values:** 6 values (25,26,27,28,30,32).
- **Frequencies (n_{in_ini}):** The number of classes corresponding to each value.

Values (x_i)	Frequencies (n_i)
$x_1=25$	$n_1=2$
$x_2=26$	$n_2=4$
$x_3=27$	$n_3=2$
$x_4=28$	$n_4=1$
$x_5=30$	$n_5=5$
$x_6=32$	$n_6=1$
Total	$N = \sum_{i=1}^P n_i = 15$

Remark :

Continuous variables are represented by a very large number of values, which necessitates grouping these values into **classes** to avoid overly lengthy and cumbersome tables. To determine these classes and construct a frequency distribution table, the following steps must be followed:

Steps to Create Frequency Distribution Classes**1. Calculate the Range (EEE) of the dataset:**

$$E = \text{Max}(x_i) - \text{Min}(x_i)$$

2. Determine the Number of Classes (NC):

- For $N \leq 100$: Use **Yule's theorem**: $NC = 2.5 \times N^{1/4}$
- For $N > 100$: Use **STURGE's theorem**: $NC = 1 + 3.32 \log N$

1. Calculate the Class Width (aia_iai):

$$ai = \frac{E+1}{NC}$$

This is the number of values included in each class.

2. Define the First Class Limits ([L0–L1[:

- $L0 = \text{Min}(x_i)$: The lower limit.
- $L1 = L0 + ai$: The upper limit.

3. Write the Remaining Classes:

- Continue adding the class width (aia_iai) to the upper limit of the previous class to determine the limits of subsequent classes.

Note:

- For continuous variables, classes have overlapping boundaries ([10–14[, [14–18[, [18–22[).
- For discrete variables with many values, classes can be non-overlapping ([10–14], [15–19], [20–24]).

When a frequency distribution table includes classes, it is necessary to calculate the **class midpoint** (x_i) using the formula :

$$x_i = \text{Lower Limit} + \text{Upper Limit} / 2.$$

Example 3: Distribution of Hourly Wages

The following dataset represents the hourly wages (in DA) of 50 employees from a company:

62,63,82,63,64,65,66,68,65,81,62,67,70,68,79,62,72,75,69,63,63,71,70,68,77,68,62,73,76,70,68,80,65,62,65,64,85,65,64,67,65,70,67,66,63,88,70,73,62,68,62, 63, 82, 63, 64, 65, 66, 68, 65, 81, 62, 67, 70, 68, 79, 62, 72, 75, 69, 63, 63, 71, 70, 68, 77, 68, 62, 73, 76, 70, 68, 80, 65,

62, 65, 64, 85, 65, 64, 67, 65, 70, 67, 66, 63, 88, 70, 73, 62, 68, 62, 63, 82, 63, 64, 65, 66, 68, 65, 81, 62, 67, 70, 68, 79, 62, 72, 75, 69, 63, 63, 71, 70, 68, 77, 68, 62, 73, 76, 68, 80, 65, 62, 65, 64, 85, 65, 64, 67, 65, 70, 67, 66, 63, 88, 70, 73, 62, 68

Population: The dataset includes 50 employees (N=50).

Characteristic Studied: Hourly wages (quantitative continuous variable).

Since the number of values is large, grouping them into classes is recommended.

Steps to Create Classes

1. Calculate the Range (R):

$$E = \text{Max}(xi) - \text{Min}(xi) = 88 - 62 = 26$$

2. Determine the Number of Classes (NC):

Since $N = 50$ $N \leq 100$, use Yule's theorem:

$$NC = 2,5 \sqrt[4]{50} \Rightarrow NC = 2,5 \times 2,659 \Rightarrow NC = 6,64 \Rightarrow NC \approx 7$$

3. Calculate the Class Width (ai):

$$1. ai = \frac{E+1}{NC} \Rightarrow ai = \frac{26+1}{7} \Rightarrow ai = 3,857 \Rightarrow ai \approx 4$$

4. Define the first class limits [L0-L1 [:

- $L0 = \text{Min}(xi) = 62$
- $L1 = L0 + ai = 62 + 4 = 66$
- First class: $[62-66[$

2. Create the Remaining Classes:

Class interval	Frequency (ni)	Midpoint (xi)
$[62-66[$	15	64
$[66-70[$	20	68
$[70-74[$	7	72
$[74-78[$	3	76
$[78-82[$	3	80
$[82-86[$	1	84
$[86-90[$	1	88
Total	$\sum ni = 50$	/

Classes	Effectifs ni	Centre de classe xi
[62-66[	20	$x_1 = \frac{62+66}{2} = 64$
[66-70[	12	$x_2 = \frac{66+70}{2} = 68$
[70-74[	9	$x_3 = \frac{70+74}{2} = 72$
[74-78[	3	$x_4 = \frac{74+78}{2} = 76$
[78-82[	3	$x_5 = \frac{78+82}{2} = 80$
[82-86[	2	$x_5 = \frac{82+86}{2} = 84$
[86-90[	1	$x_6 = \frac{86+90}{2} = 88$
Total	$\sum ni = 50$	/

II.3.2 Frequency Distribution Table (Relative Frequencies and Percentages)

It is common to transform a frequency distribution table into a table of **relative frequencies** or **percentages** for easier interpretation and comparison.

Definition of Relative Frequency

The relative frequency (fif_ifi) of a modality is the proportion of individuals in the total population who exhibit this modality. It is calculated by dividing the frequency (nin_ini) of the modality by the total population size (N):

$$: fi = \frac{ni}{N} = \frac{ni}{\sum ni} \text{ (relative frequency)}$$

Key Differences

- **Frequency (ni)** is an absolute measure and does not directly allow comparisons between distributions.
- **Relative Frequency (fi)** is a relative measure and facilitates direct comparisons.

Example: Relative Frequency Table for Student Ages

Consider a group of 35 students, and the absolute frequencies (n_i) for their ages are grouped as follows:

Age (xi)	Absolute frequency (ni)	Relative frequency (fi)
18	5	$05/35 = 0,143$
19	4	$04/35 = 0,114$
20	10	$10/35 = 0,285$
21	11	$11/35 = 0,314$
22	3	$03/35 = 0,086$
23	1	$01/35 = 0,028$
24	1	$01/35 = 0,028$
Total	$\sum_{i=1}^P n_i = 35$	$\sum_{i=1}^P f_i = 01$

Converting Relative Frequencies to Percentages

To improve readability, relative frequencies are often converted into percentages. This is calculated as :

$$fi\% = fi \times 100 = \frac{n_i}{N} \times 100$$

Example: Percentage Table for Student Ages

Age (xi)	Absolute frequency (ni)	Relative frequency (fi)	Percentage (fi%)
18	5	$05/35 = 0,143$	$0.143 \times 100 = 14.3\%$
19	4	$04/35 = 0,114$	$0.114 \times 100 = 11.4\%$
20	10	$10/35 = 0,285$	$0.285 \times 100 = 28.5\%$
21	11	$11/35 = 0,314$	$0.314 \times 100 = 31.4\%$
22	3	$03/35 = 0,086$	$0.086 \times 100 = 8.6\%$
23	1	$01/35 = 0,028$	$0.028 \times 100 = 2.8\%$
24	1	$01/35 = 0,028$	$0.028 \times 100 = 2.8\%$
Total	$\sum_{i=1}^P n_i = 35$	$\sum_{i=1}^P f_i = 01$	$\sum fi\% = 100$

Conclusion

Using relative frequencies and percentages simplifies the interpretation of statistical distributions. While absolute frequencies measure the raw count, relative frequencies and percentages allow for direct comparisons between distributions and make the data more accessible for analysis and decision-making.

Example 4: Frequency Distribution Table

The following table represents the heights (in cm) of 50 children:

Height (cm)	Frequencies (ni)	Relative Frequencies (fi)	Percentages (fi%)
[110–113[	10	$f_1 = \frac{10}{50} = 0,2$	$0.2 \times 100 = 20\%$
[113–116[	15	$f_2 = \frac{15}{50} = 0,3$	$0.3 \times 100 = 30\%$
[116–119[	12	$f_3 = \frac{12}{50} = 0,24$	$0.24 \times 100 = 24\%$
[119–122[	7	$f_4 = \frac{7}{50} = 0,14$	$0.14 \times 100 = 14\%$
[122–125[	3	$f_5 = \frac{3}{50} = 0,06$	$0.06 \times 100 = 6\%$
[125–128[	2	$f_6 = \frac{2}{50} = 0,04$	$0.04 \times 100 = 4\%$
[128–131[	1	$f_7 = \frac{1}{50} = 0,02$	$0.02 \times 100 = 2\%$
Total	$\sum_{i=1}^p ni = 50$	$\sum fi = 1$	$\sum fi\% = 100$

II.3.3 Cumulative Frequency Distribution Table

Cumulative frequencies allow determining the number of observations above or below a particular value or class within a dataset.

Types of Cumulative Frequencies

1. **Cumulative Ascending Frequencies (NCi↑) :**

The number of individuals with values less than or equal to a specific modality.

2. **Cumulative Descending Frequencies (NCi↓):**

The number of individuals with values greater than or equal to a specific modality.

Example 1: Number of Rooms per Household

Number of Rooms (xi)	Number of Households (ni)	Cumulative Ascending Frequencies (NCi↑)
1	5	5
2	10	5+10=15
3	20	15+20=35
4	30	35+30=65
5	25	65+25=90
6	10	90+10=100
Total	100	/

Cumulative Descending Frequencies

Number of Rooms (xi)	Number of Households (ni)	Cumulative Descending Frequencies (NCi↓)
1	5	100-5=95
2	10	95-10=85
3	20	85-20=65
4	30	65-30=35
5	25	35-25=10
6	10	10-10=0
Total	100	/

Example 2 : Employee delays

The following table represents employee delays (in minutes) in arriving at work:

Delay (minutes)	Number of Employees (ni)	Cumulative Descending Frequencies (NCi↓)	Cumulative Ascending Frequencies (NCi↑)	Relative Frequency (fi)	Cumulative Relative Frequency (FCi↑)	Cumulative Percentage (FCi↑%)
[00–05[	4	72	4	0.050	0.050	5%
[05–10[	7	68	11	0.10	0.150	15%
[10–15[	10	61	21	0.140	0.290	29%
[15–20[	15	51	36	0.210	0.500	50%
[20–25[	18	36	54	0.250	0.750	75%
[25–30[	12	18	66	0.170	0.920	92%
[30–35[	6	6	72	0.080	1.001	100%
Total	72	/	/	1.001	1.001	100%

Insights from Cumulative Frequencies

1. **Number of employees with a delay greater than 10 minutes** : 61.
2. **Number of employees with a delay less than 20 minutes**: 36.
3. **Proportion of employees with a delay less than 20 minutes**: 0,500.
4. **Percentage of employees with a delay less than 25 minutes**: 75%.
5. **Number of employees delayed between 10 and 25 minutes**:
 - From $NCi↑$: $54-11=43$.
 - From $NCi↓$: $61-18=43$.

Cumulative frequencies enable detailed insights into data distributions, facilitating comparisons, proportions, and percentile analyses.

3

CHAPTER

Representation of data : Statistical graphs

In this chapter :

- Graphical representations of a qualitative characteristic
- Graphical representations of a quantitative characteristic
 - Case of a discrete (Discontinuous) variable
 - Case of a continuous variable

III.1 Introduction

Although a statistical table contains all the information collected from the sample, it is very useful to translate this information into a graph to obtain a visual summary.

A graph is a means of synthetic and rapid visual communication. It should not consist of simply copying an existing figure, which is often unnecessary, but rather of constructing a new, effective image that is quickly understandable, memorable, and useful.

A graph is beneficial to the author as it allows them to discover what the information contains, first in its global form, by identifying the relationships between the various elements that define it. It serves as a tool for reflection, enabling the formulation of questions and hypotheses to develop and support the commentary that will be made.

It is evident that a graph should never be used as a mere aesthetic accessory. Depending on the nature of the variable being studied, different types of representations are used.

III.2 Presentation of Statistical Data in Graphical Form

III.2.1 Graphical Representations of a Qualitative Characteristic

In this case, the main graphs that can be used are:

III.2.1.1 Bar Chart (Organ Pipes Chart)

The graph consists of rectangles with constant widths but lengths proportional to the frequencies (absolute, relative, or percentages). On the x-axis, the modalities are represented, while on the y-axis, the frequencies or counts for each modality are displayed.

Example 1:

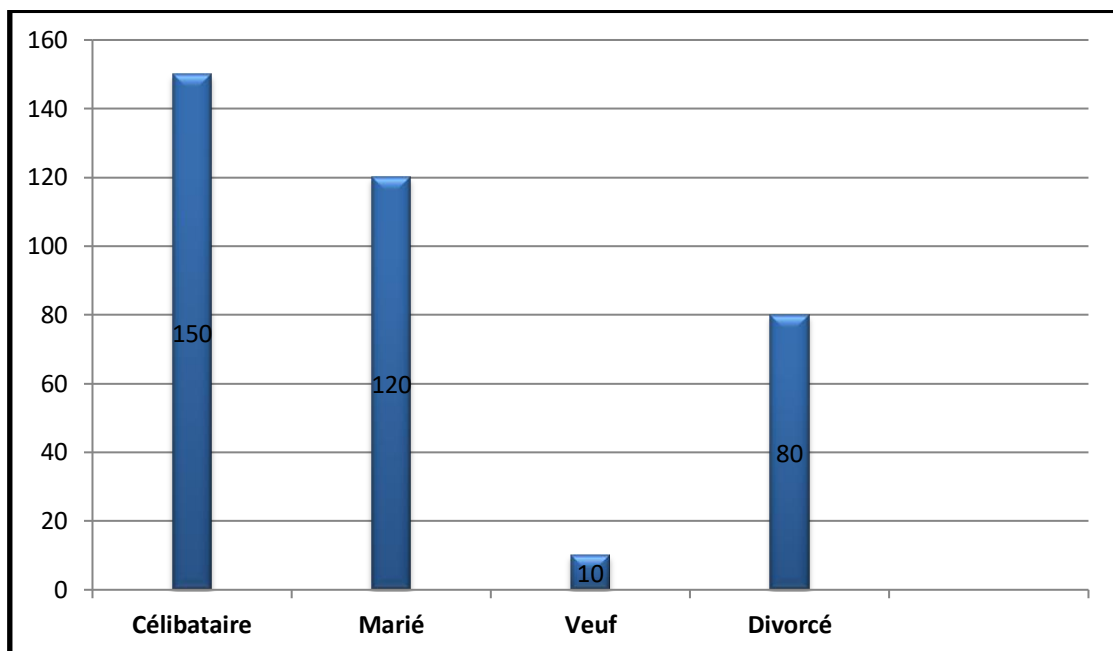
The following table represents the distribution of a sample of individuals from a city in Algeria according to their marital status.

Task:

Represent this data using a bar chart.

Modalities	Counts (ni)	Relative Frequency (fi)	Percentage (fi%)
Single	150	0.42	42%
Married	120	0.33	33%
Widowed	10	0.03	3%
Divorced	80	0.22	22%
Total	360	≈1	100%

Bar Chart (Organ Pipes Chart)



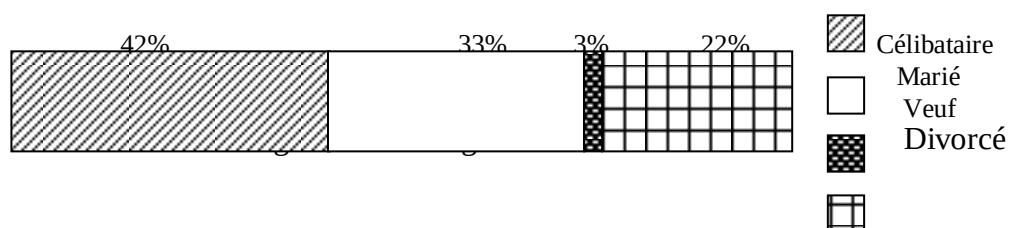
III.2.1.2 The Rectilinear Diagram

It is a rectangle divided into sections, each section representing a modality of the characteristic expressed as a percentage.

Example 2:

Let us take the previous example and draw the rectilinear diagram.

Diagramme rectiligne



III.2.1.3 The Pie Chart (Sector Diagram)

A pie chart is a circle divided into sectors, with each sector representing a modality of the characteristic. These diagrams are particularly suitable for political or socio-economic data. The angle corresponding to each sector is calculated using the following formula:

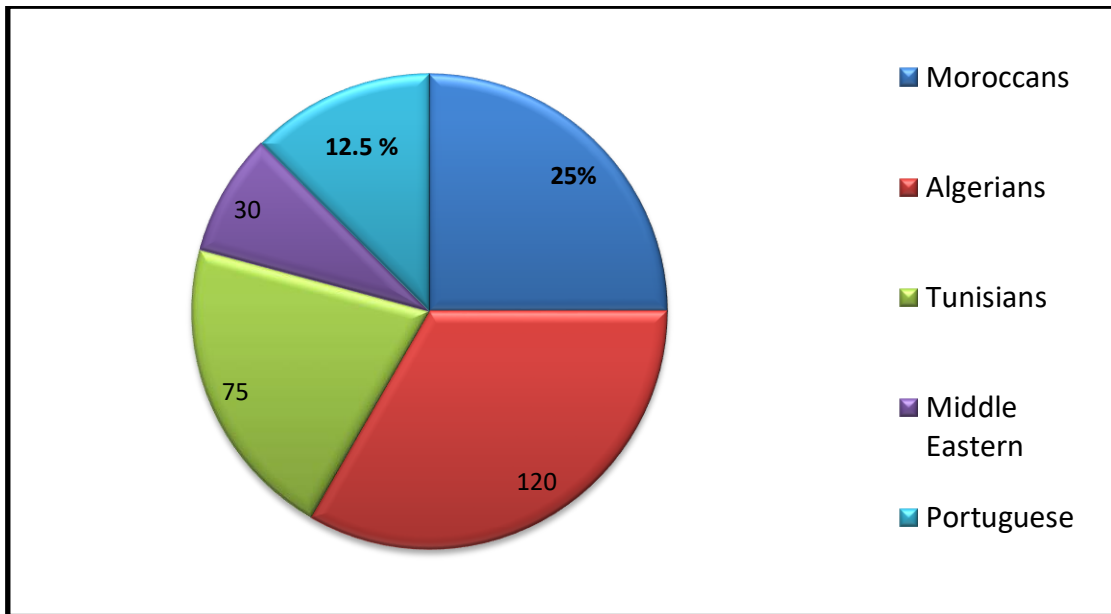
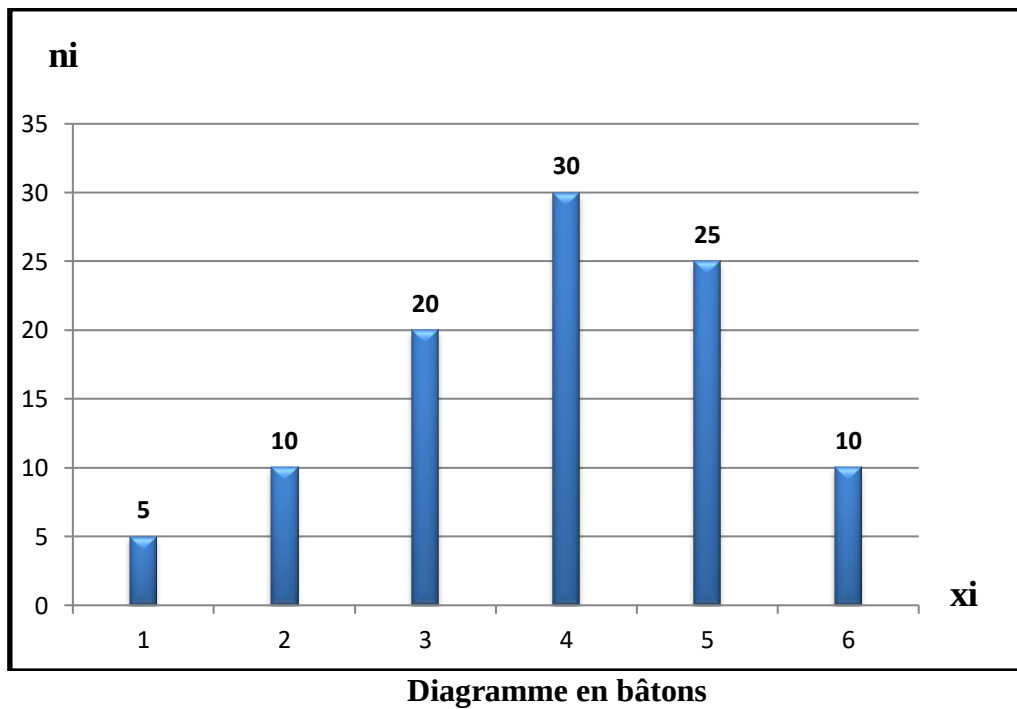
$$\alpha = f_i \times 360$$

Example 3:

The table below represents the distribution of foreigners in France according to their nationality. Represent this data using a pie chart.

Nationality	Count (n_i)	Relative Frequency (f_i)	Angle ($\alpha = f_i \times 360^\circ$)
Moroccans	150	0.25	90°
Algerians	200	0.333	120°
Tunisians	125	0.208	75°
Middle Eastern	50	0.083	30°
Portuguese	75	0.125	45°
Total	600	≈ 1.00	360°

The Pie Chart (Sector Diagram)



III.2.2 Graphical Representations of a Quantitative Variable

In this case, it is essential to distinguish between a discrete variable and a continuous variable.

III.2.2.1 Case of a Discrete (Discontinuous) Variable

The different graphical representations include:

a) The Stick Diagram (Differential Chart)

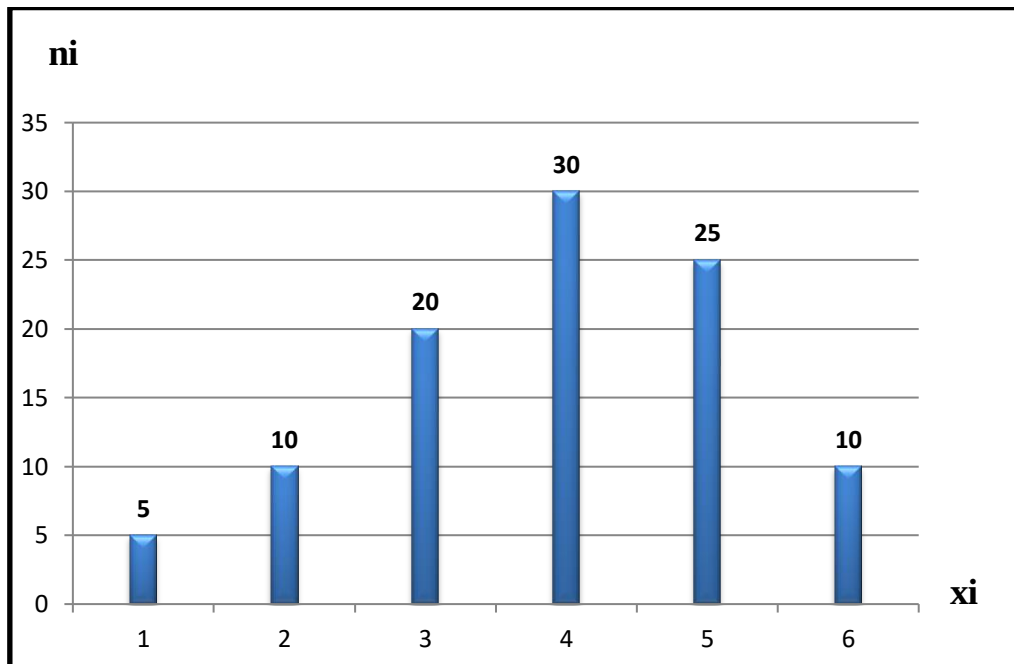
This type of chart consists of simple vertical bars, where the height of each bar is proportional to the frequency (absolute, relative, or percentage) of each value of the variable.

Example 4:

The following table presents the distribution of a sample of housing units in a neighborhood according to the number of rooms. Represent these data graphically.

Number of Rooms (x_i)	Count (n_i)	Cumulative Frequency (Ascending, $N_{ci}\uparrow$)	Cumulative Frequency (Descending, $N_{ci}\downarrow$)
1	5	5	100
2	10	15	95
3	20	35	85
4	30	65	65
5	25	90	35
6	10	100	10
Total	100	/	/

Stick Diagram (Differential Chart)



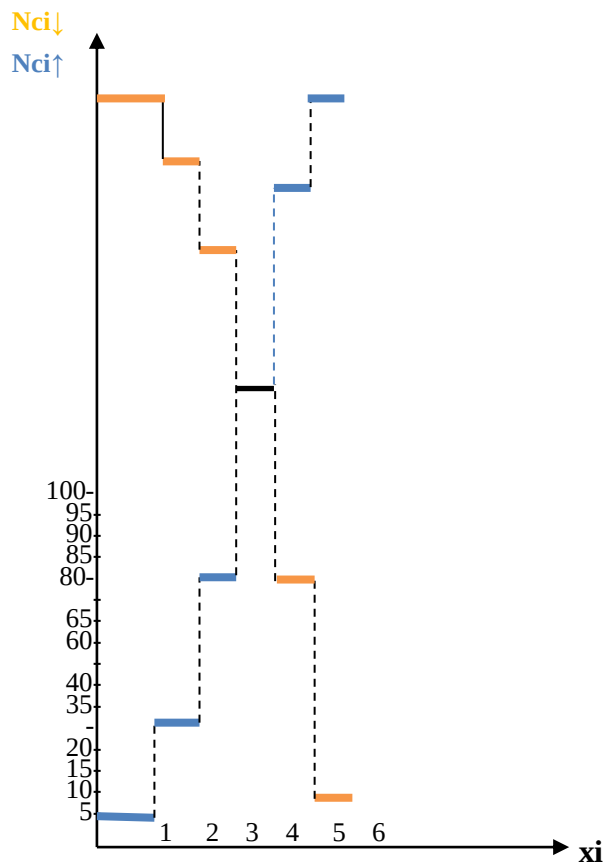
b) Cumulative Diagram (Integral Chart)

This representation consists of line segments that depict the cumulative frequencies—either ascending (increasing) or descending (decreasing)—of the variable's values.

Cumulative Step Diagram (Ogive Chart)

This type of graph visually represents the cumulative distribution of the data, showing the progressive accumulation of frequencies as the variable increases.

Cumulative Diagram (Integral Chart)



III.2.2.2 Case of a Continuous Variable

Among the different types of graphs used for continuous variables, we distinguish:

a) The Histogram

A histogram is the graphical representation of data when grouped into classes; in other words, it represents a statistical series where the variable is quantitative and continuous.

The histogram consists of adjacent rectangles with varying heights corresponding to the frequency (absolute or relative) and widths corresponding to the class intervals (amplitude).

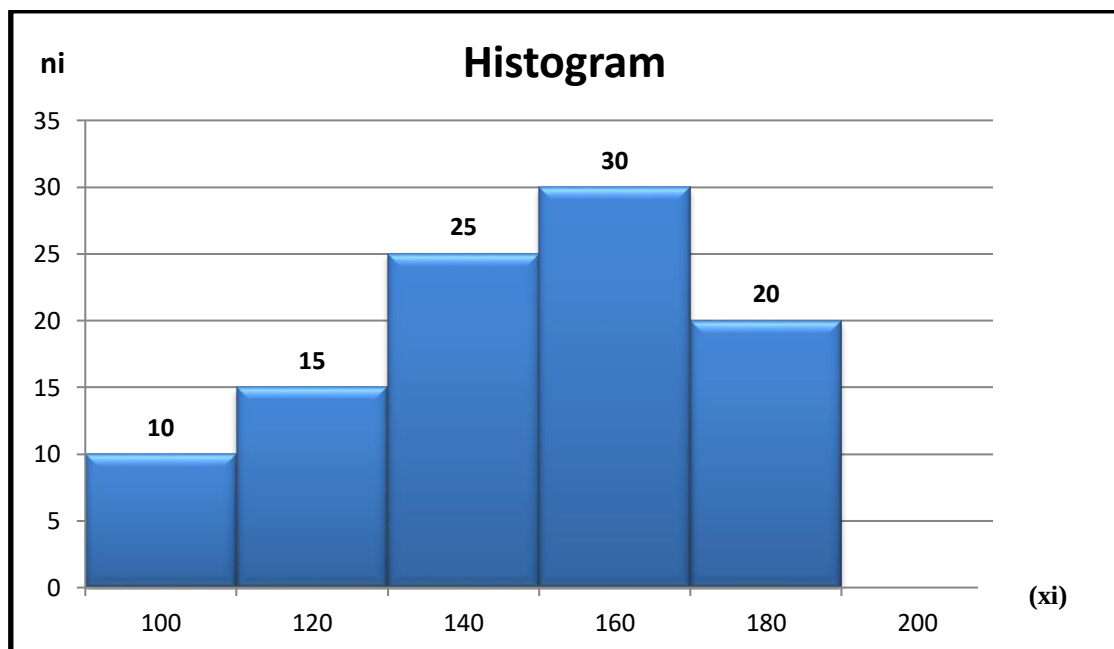
Note: When constructing a histogram, two scenarios may arise:

1. **Equal class widths:** The histogram can be constructed directly.

Example 5:

The following table represents the hourly income (in Algerian Dinars) of a sample of employees. Represent these data using a histogram.

Class Intervals	Frequency (n_i)
[100 - 120[	10
[120 - 140[	15
[140 - 160[	25
[160 - 180[	30
[180 - 200[	20
Total	100

Histogram Representation

Second Situation: Unequal Class Widths

When class widths (amplitudes) are unequal, the frequencies (n_i) must be adjusted before constructing the histogram. The adjustment follows these steps:

1. **Calculate the class widths (a_i)** for each class interval.
2. **Find the greatest common divisor (GCD) of the amplitudes** to determine the unit amplitude (a_u).
3. **Determine the unit amplitude (a_u)** using the most frequent or minimum class width.
4. **Compute the correction coefficients (C_i)** using : $C_i = \frac{a_i}{a_u}$
5. **Calculate the adjusted frequencies (h_i)** using : $h_i = \frac{n_i}{C_i}$

Example 6 :

The table below represents the distribution of monthly income (in thousands of Algerian Dinars) for a sample of workers in a private company. Construct the histogram based on the corrected frequencies.

Class Intervals	Frequency (n_i)	Class Width (a_i)	Correction Coefficient ($C_i = a_i/a_u$)	Corrected Frequency ($h_i = n_i/C_i$)
[10 - 15[	30	5	1	30
[15 - 25[	26	10	2	13
[25 - 30[	10	5	1	10
[30 - 45[	21	15	3	7
[45 - 65[	8	20	4	2
[65 - 70[	5	5	1	5
Total	100	/	/	/

Note:

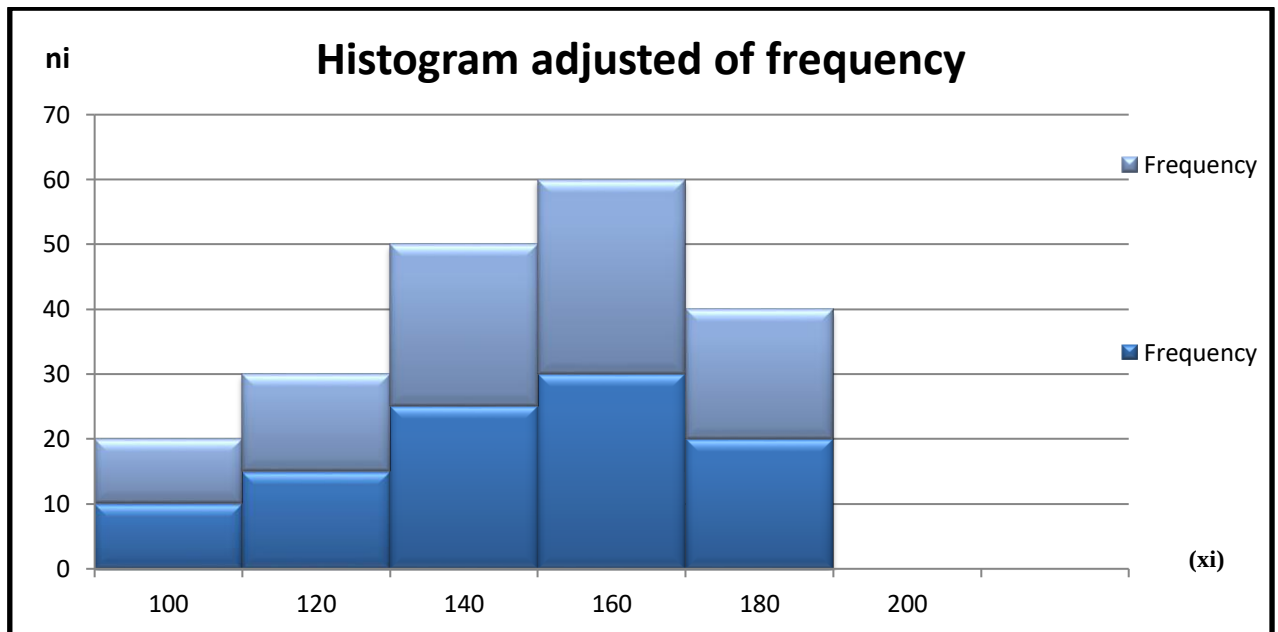
In this case, where class widths are unequal, the frequencies must be adjusted before plotting the histogram.

The **greatest common divisor (GCD)** of the class widths (a_i) is 5, which means that the **unit amplitude (a_u)** is : $a_u=5$

Using this value, the correction coefficients and adjusted frequencies were calculated in the previous table. Now, the histogram should be plotted based on the corrected frequencies (h_i).

Histogram representation

The histogram should be constructed using the corrected frequencies (h_i), ensuring that the class widths are properly accounted for in the graphical representation.



b) The Frequency Polygon

The **frequency polygon** is constructed by connecting the midpoints of the tops of the histogram's rectangles with straight-line segments. To ensure closure, the polygon is extended at both ends along the horizontal axis.

Example 7:

The following table represents the distribution of a sample of students according to their weight (in kg). Construct the frequency polygon for this distribution.

Class Intervals (Weight in Kg)	Frequency (n_i)	Class Midpoint (m_i)
[40 - 50[	X_1	$m_1 = (40+50)/2$
[50 - 60[	X_2	$m_2 = (50+60)/2$
[60 - 70[	X_3	$m_3 = (60+70)/2$
[70 - 80[	X_4	$m_4 = (70+80)/2$
[80 - 90[	X_5	$m_5 = (80+90)/2$
Total	N	/

Steps to draw the frequency polygon:

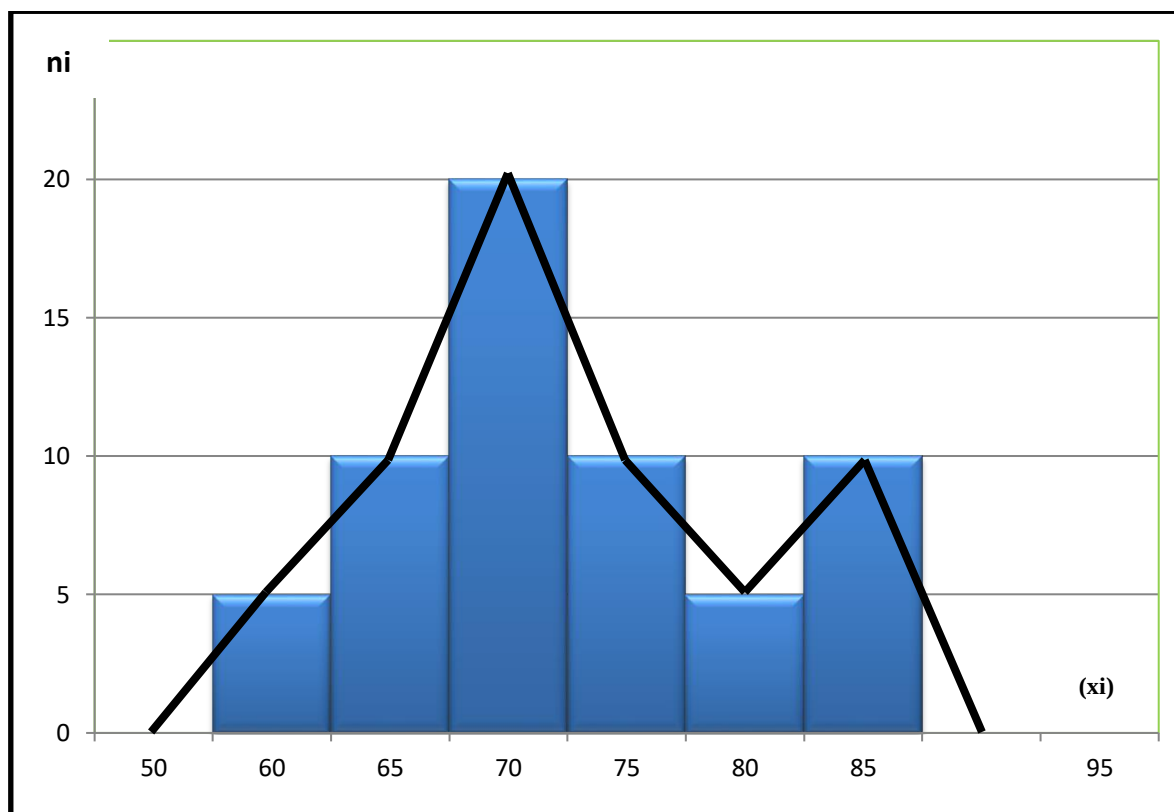
1. Calculate the **midpoint** (m_i) for each class using: $m_i = \text{Lower Bound} + \text{Upper Bound} / 2$.
2. Plot the **midpoints** on the x-axis and their corresponding **frequencies** on the y-axis.
3. Connect these points using straight-line segments.
4. Extend the polygon at both ends by adding extra points at zero frequency, ensuring closure on the horizontal axis.

This graphical representation helps visualize the distribution of the data more smoothly compared to a histogram.

The Frequency Polygon

The **frequency polygon** provides a more realistic view of the distribution by eliminating the breaks between classes, offering a smoother representation of the data.

The frequency polygon



c) Cumulative Curves (Increasing and Decreasing)

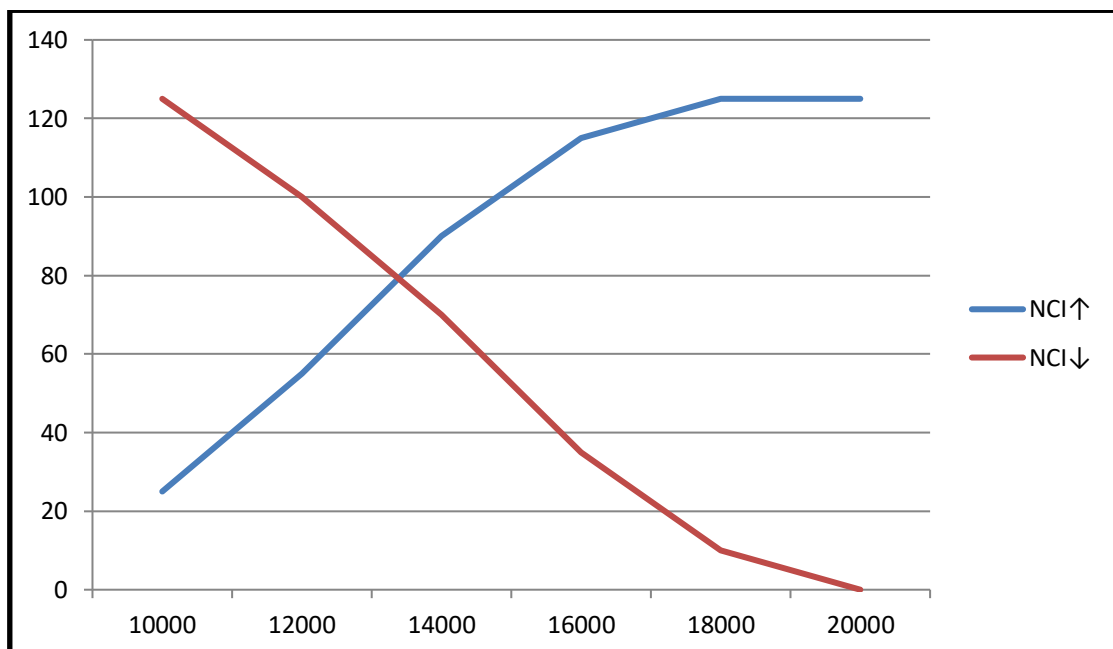
Cumulative curves are obtained by connecting the points derived from class boundaries on the x-axis and cumulative frequencies on the y-axis.

Example 8:

The table below represents the distribution of employees in a company based on their monthly salary (in Algerian Dinars).

Class Intervals (Salary in DA)	Frequency (n_i)	Upper Bound (Less than)	Cumulative Frequency (Ascending, $N_{ci}\uparrow$)	Lower Bound (More than)	Cumulative Frequency (Descending, $N_{ci}\downarrow$)
[10,000 - 12,000[	25	Less than 12,000	25	More than 10,000	125
[12,000 - 14,000[	30	Less than 14,000	55	More than 12,000	100
[14,000 - 16,000[	35	Less than 16,000	90	More than 14,000	70
[16,000 - 20,000[	25	Less than 20,000	115	More than 16,000	35
[20,000 - 25,000[	10	Less than 25,000	125	More than 20,000	10
Total	125				

Cumulative Curves (Increasing and Decreasing)



Steps to Construct Cumulative Curves:

1. Ascending Cumulative Curve (Ogive):

- Plot the **upper class limits** on the x-axis.
- Plot the **cumulative ascending frequencies** ($N_{ci}\uparrow$) on the y-axis.
- Connect the points with a smooth curve.

2. Descending Cumulative Curve:

- Plot the **lower class limits** on the x-axis.

- Plot the **cumulative descending frequencies** ($N_{ci\downarrow}N_{ci\downarrow}N_{ci\downarrow}$) on the y-axis.
- Connect the points with a smooth curve.

These cumulative curves provide a clear overview of the data distribution, helping to identify trends and patterns in salary distribution.

4

CHAPTER

Measures of central tendency

In this chapter :

-The Mode

- Case of a discrete variable
- Case of a continuous variable

-The Median

- Case of a discrete variable
- Case of a continuous variable

-The Arithmetic Mean

- Case of a discrete variable
- Case of a continuous variable

-Generalization of the Mean

- Geometric mean
- Harmonic mean
- Quadratic mean

IV.1 Introduction

When analyzing the statistical distribution of a quantitative variable, statistical processing of information is not limited to graphical representations. The primary goal of a statistician is to characterize this distribution numerically and summarize the information using well-chosen numerical indicators, known as **measures of central tendency**. These summarized values are called **central tendency measures** because they highlight the main values of the distribution.

The challenge lies in determining the most relevant numerical measure. Several indicators can be constructed to characterize a statistical series. This chapter explores these indicators before studying how to compare two situations using indices.

The **role of position and dispersion parameters** is to convey, through calculations, statistical information reflecting a real-world phenomenon.

This chapter introduces some of the most commonly used indicators, including **mode, median, and arithmetic mean**, known as **central tendency measures** or **position indicators**. We will explain their calculation methods and significance, distinguishing between **ungrouped data** and **grouped data** (either by individual values or by class intervals).

IV.2 The Mode

The **mode** of a statistical series, denoted as "**Mo**", is the **value (modality) with the highest frequency** (or the largest count).

- The mode is **well-defined for both qualitative variables and discrete quantitative variables**.
- It is a **position indicator that is not affected by extreme values**.

IV.2.1 Case of a Discrete Variable

The **mode is directly identified** from the frequency table as the **value corresponding to the highest count**.

Example 1:

In a small business, the number of sick leave days taken by each of 12 employees over a year was recorded as follows:

Data: 11, 5, 3, 5, 7, 3, 11, 13, 8, 3, 0, 3

The most frequently occurring value is **3 days**, which appears **four times**.

Thus, the **mode** of the series is:

Mo=3 days.

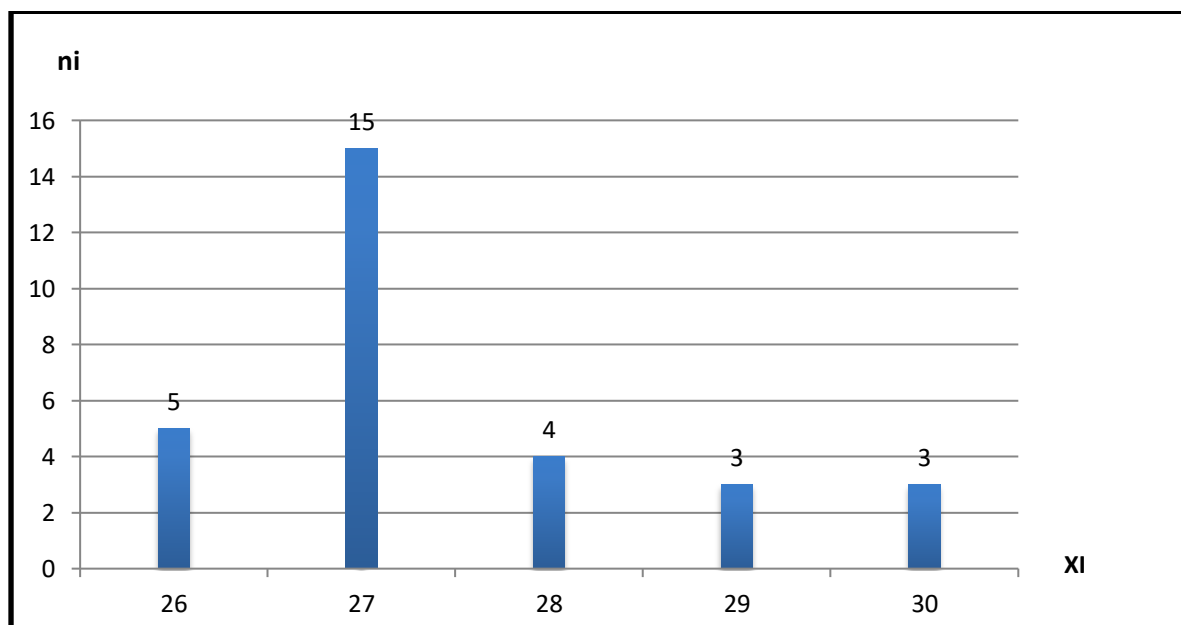
Example 2:

The following table represents the distribution of marriage age among 30 individuals:

Marriage Age (xix_xi)	26	27	28	29	30
Frequency (ni)	5	15	4	3	3

- The highest frequency (15) corresponds to the age 27.
- Thus, the **mode** is: Mo=27Mo = 27Mo=27

Graphically, the mode corresponds to the **x-axis value of the tallest bar in a bar chart**.



IV.2.2 Case of a Continuous Variable

The **mode** of a continuous statistical variable, denoted as **M₀**, is the value corresponding to the **highest frequency per unit of class width (amplitude)**.

For a **continuous quantitative variable**, we refer to the **modal class**, which is the **class with the highest frequency density**.

The determination process follows two steps:

1. **Identifying the modal class** (the class with the highest frequency).
2. **Calculating the mode using an interpolation formula.**

However, two cases must be distinguished depending on whether the **class widths (amplitudes) are equal or unequal**.

IV.2.2.1 Case of Equal Class Widths

In this case, the **modal class** is the class with the highest frequency (n_i), denoted as **[L₀,L₁[**.

The mode is calculated using the following formula:

$$M_0 = L_0 + a_i \left[\frac{m_1}{m_1 + m_2} \right]$$

$$M_0 = 15 + \left[\frac{(20-10)}{(20-10)+(20-15)} \right] * 5 = \mathbf{18,33}$$

Where:

- **L₀L₀** = **Lower boundary of the modal class**
- **L₁** = **Upper boundary of the modal class**
- **a_i** = **Width (amplitude) of the modal class**
- $m_1 = n_i - n_{i-1}$
- $m_2 = n_i - n_{i+1}$
(Difference between the modal frequency and the previous class frequency)

Alternatively, the mode can also be calculated using **relative frequencies** :

$$M_0 = L_0 + a_i \left[\frac{f_i - f_{i-1}}{f_i - f_{i-1} + f_i - f_{i+1}} \right]$$

Example 3: Distribution of a Population of 40 Newborns by Birth Length

The modal class	Class (Length in cm)	Amplitude	Frequency (n _i)
	[48-50[	2	3
	→ [50-52[	2	10
	[52-54[	2	9
	[54-56[	2	7
	[56-58[	2	7
	[58-60[	2	3
	[60-62[	2	1
	Total	/	40

The mode is calculated as follows:

$$Mo = 50 + 2 \left[\frac{(10-3)}{(10-3)+(10-9)} \right] = 50 + 2 \left[\frac{7}{7+1} \right] = 51,75$$

IV.2.2.2 Case of Unequal Class Widths

When class widths (amplitudes) are **unequal**, the **modal class** is the class with the **highest corrected frequency**.

Before calculating the mode, the frequencies (n_i) must be adjusted:

- **Corrected frequency (h_i):** $h_i = \frac{n_i}{c_i}$

-

- **Coefficient (C_i):** $c_i = \frac{a_i}{au}$

- **Unit amplitude (a_{ua}):**

- The smallest class width in the dataset.

- The mode is then calculated using :

$$M_0 = L_0 + a_i \left[\frac{m_1}{m_1 + m_2} \right]$$

Where:

- L₀ = lower bound of the modal class
- a_i = class width of the modal class

- $M1 = h_i - h_{i-1}$ (difference between the corrected frequency of the modal class and the preceding class)
- $M2 = h_i - h_{i+1}$ (difference between the corrected frequency of the modal class and the following class)

This approach ensures that the mode is accurately determined even when the class intervals are not uniform.

Example 4: Distribution of 100 People by Age Group

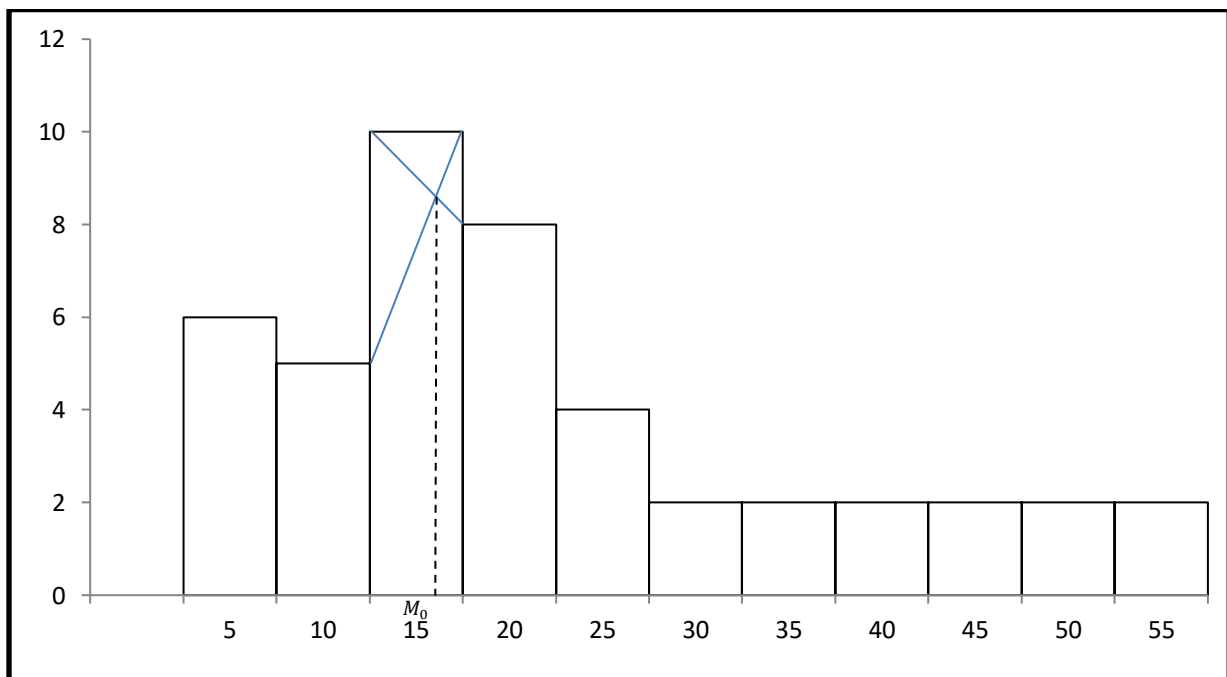
Age Classes	Frequency (ni)	Amplitude (ai)	Coefficient (ci)	Corrected Frequency (hi)
[5 - 10[	11	5	1	11
[10 - 15[	10	5	1	10
[15 - 20[	20	5	1	20
[20 - 30[	30	10	2	15
[30 - 40[	18	10	2	9
[40 - 60[	11	20	4	2.75
Total	100	/	/	/

Given that the **unit amplitude (aua_uau)** is 5, the mode is calculated as:

$$Mo = 15 + 5 \left[\frac{20 - 10}{[20 - 10] + [20 - 15]} \right] = 15 + \frac{50}{10 + 5} = 15 + \frac{50}{15}$$

$$Mo = 18,33$$

Graphical determination of the Mode (Mo)



The **mode (Mo)** can be estimated **graphically** by considering the **intersection of the diagonals of trapezoid ABCD**. This trapezoid is constructed using:

- The **top vertices of the rectangle** associated with the **modal class**
- The **top vertices of the rectangles of the adjacent classes**

Remarks:

- If all categories have the **same frequency**, the **distribution is uniform**, meaning it **has no mode**.
- If the dataset has **only one mode**, the distribution is **unimodal**.
- If the dataset has **two or more modes**, the distribution is called **bimodal** or **multimodal**, respectively.

IV.3 The Median

The **median** of a distribution, denoted as **Me**, is the **value that divides the statistical population into two equal parts**, with half of the individuals having values below and the other half above it.

Thus, the **median is the value such that 50% of the population is below it and 50% is above it** when the data is ordered in ascending order.

IV.3.1 Case of a Discrete Variable

To determine the median of an **ungrouped dataset**, the observed values must first be arranged in **ascending order**. The median is the "**middle**" **value** in this ordered series.

- If the number of observations (**nnn**) is **odd**, the median is the value at the $\frac{n+1}{2}$ **position** in the ordered series.
- If the number of observations (**nnn**) is **even**, the median falls **within the median interval**, which is determined by the two central values located at positions $\frac{n}{2}$ and $\frac{n}{2}+1$. In this case, the median is the midpoint of these two values.

Example 5 :

The following data represent the distribution of a group of **nine households** according to the **number of people per household**:

Raw data:

3-1-4-6-2-4-3-5-7.

Step 1: Arrange the data in ascending order:

1-2-3-3-4-4-5-6-7

Step 2: Identify the median position

- The number of observations **$n=9$**
- The median is at position: $(9+1)/2=10/2=5$
- The **5th value** in the ordered dataset is **4**.
- **Thus, the median is:** $Me=4$ people.

Alternative Method to Determine the Median:

- If the number of observations **n is odd**, it can be written as **$n=2p+1$** $n = 2p + 1$, and the median is found at position **$p+1$** .
- If the number of observations **n is even**, it takes the form **$n=2p$** $n = 2p$, and the median lies within the **median interval** between the two central values at positions **p and $p+1$** .

This alternative method provides a systematic approach to locating the median, ensuring consistency in statistical analysis.

Example 6 :

Consider the distribution of **12 households** according to the number of people per household:

Raw data:

5-3-2-3-6-3-5-4-7-2-1-4

Step 1: Arrange the data in ascending order

1-2-2-3-3-3-4-4-5-5-6-7

Step 2: Identify the median position

- The number of observations **$n=12$** **$n=12$ (even)**
- The median is found between the **6th** and **7th** values:
- et $\frac{n}{2} + 1 = \frac{12}{2} + 1 = 7$ donc $Me = \frac{3+4}{2} = 2,5$
- The **6th value** is **3** and the **7th value** is **4**
- The median is calculated as:

$$Me = 3 + 4 / 2 = 7 / 2 = 3.5.$$

Thus, the **median number of people per household is 3.5**.

Determining the Median for a Discrete Variable Using Cumulative Frequencies

For a **discrete variable**, the **median can also be calculated using cumulative frequencies**.

The median is the **value associated with a cumulative frequency of $N/2$** (where **N is the total frequency**) or a cumulative percentage of **50%**.

Consider the **age distribution of a group of 30 students** in a course:

Age	Frequency (n_i)	Relative Frequency (f_i)	Cumulative Frequency	Cumulative Percentage
18	2	$2/30 = 06,7\%$	2	$2/30 = 06,7\%$
19	4	$4/30 = 13,3\%$	6	$6/30 = 20,0\%$
20	9	$9/30 = 30,0\%$	15	$15/30 = 50,0\%$
21	11	$11/30 = 36,0\%$	26	$26/30 = 86,7\%$
22	4	$3/30 = 10,0\%$	30	$30/30 = 100\%$
Total	30	1=100%	/	/

Finding the Median (Me)

- The **total number of students is 30**.
- The **median is the value below which $30/2 = 15$ students fall**.
- From the **cumulative frequency column**, we see that the 15th student falls in the **age group of 20 years**.
- **Thus, the median age is:**

Me=20 years.

This method allows for a systematic determination of the median in datasets where values are grouped into categories.

Handling the Median When It Does Not Fall on an Exact Value

When the **median does not correspond exactly** to a value in the dataset, the **conventional approach is to take the next higher value**.

For a **continuous and grouped variable**, the median can **only be estimated approximately**. In such cases, **linear interpolation** is used, assuming that frequencies are **uniformly distributed within a class**.

Properties of the Median

- The median provides **valuable insights into the central tendency** of a statistical distribution.
- **It is not influenced by extreme values**, making it a **robust measure of central tendency** compared to the mean.

IV.3.2 Case of a Continuous Variable

Regardless of whether the class widths are **equal or unequal**, the **calculation method for the median remains the same**.

Steps to Calculate the Median

1. **Identify the Median Class**
 - This is the class where the cumulative frequency **first exceeds $N/2$** ($N/2$ is half of the total frequency).

2. Use the Median Formula

- Let $[L_0, L_1[$ be the **median class**, where:
 - L_0 = **lower boundary of the median class**
 - a_i = **class width**
 - N_i = **cumulative frequency of the median class**
 - N_{i-1} = **cumulative frequency before the median class**
 - N = **total frequency**

Mathematical Formula for the Median

$$M_e = L_{min} + \left[\frac{RME - S}{\dot{S} - S} \right] * a_i$$

Where:

$$RME = \frac{\sum ni}{2} \quad (\text{half of the total frequency}).$$

- S = cumulative frequency before the median class.
- $\dot{S}$ = cumulative frequency of the median class.

Example Calculation

$$M_e = 20 + \left[\frac{(50 - 41)}{71 - 41} \right] * 10$$

Alternative Formulas

1. Using absolute frequencies:

$$2. \quad M_e = L_0 + a_i \left| \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \right|$$

2. Using relative frequencies:

$$3. \quad M_e = L_0 + a_i \left| \frac{0,5 - f_{i-1}}{f_i - f_{i-1}} \right|$$

These formulas allow for an **accurate estimation of the median** when working with **continuous grouped data**.

Example 7: Determining the Median for a Grouped Continuous Variable

The following table represents the **distribution of 100 individuals by age group**:

Age Class	Frequency (ni)	Cumulative Frequency (Ni↑)	Relative Frequency (fi)	Cumulative Relative Frequency (fi↑)
[5 - 10[	11	11	0.11	0.11
[10 - 15[	10	21	0.10	0.21
[15 - 20[	20	41	0.20	0.41
[20 - 30[	30	71	0.30	0.71
[30 - 40[	18	89	0.18	0.89
[40 - 60[	11	100	0.11	1.00
Total	100	/	1.00	/

Calculating the Median (Me)

1. Identify the Median Class

- The total number of individuals: **$N=100$**
- The median position: $N/2=100/2=50$
- The cumulative frequency **just before 50** is **41**.
- The cumulative frequency that **exceeds 50** is **71**.
- So, the **median class** is **[20 - 30[**.

2. Apply the Median Formula

- $L_0=20$ (Lower bound of the median class)
- $ai=10$ (Class width)
- $N/2=50$
- $N_{i-1}= 41$ (Cumulative frequency before the median class)
- $N_i= 71$ (Cumulative frequency of the median class)

$$Me = 20 + 10 \left| \frac{50 - 41}{71 - 41} \right| = 20 + 10 \left[\frac{9}{30} \right]$$

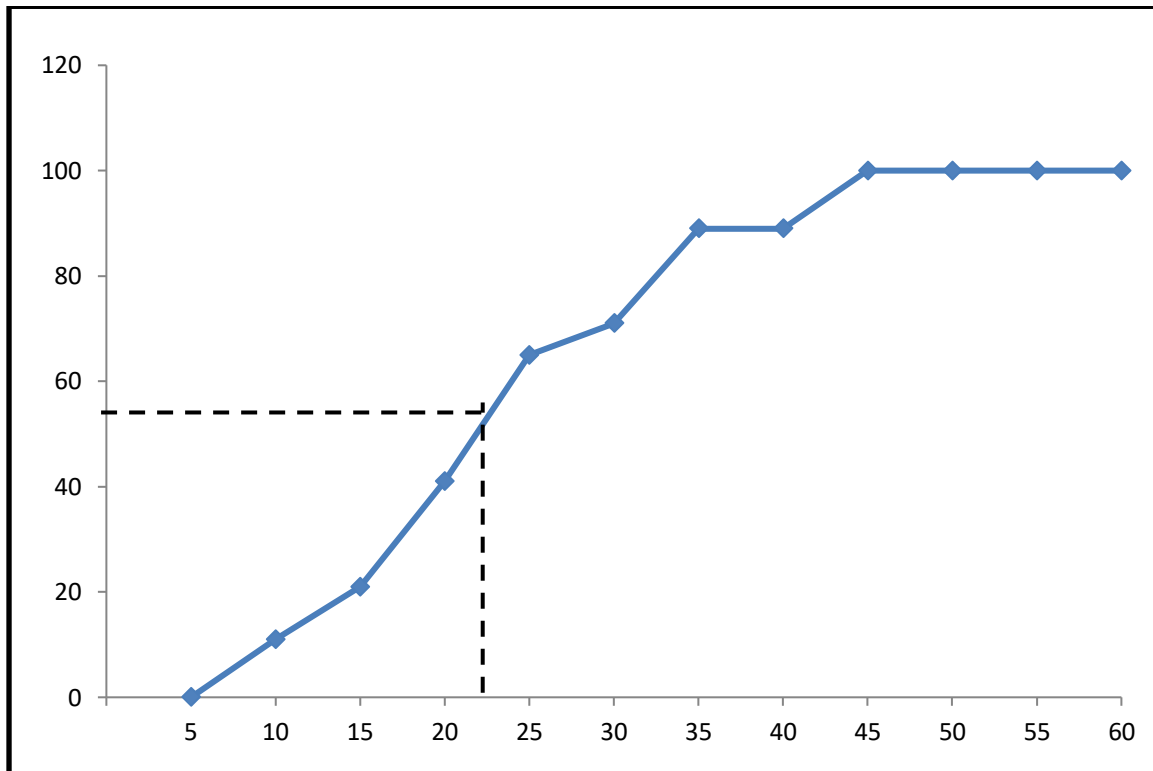
$$Me = 23 \text{ years}$$

Alternative Method: Using Relative Frequencies

$$Me = 20 + 10 \left| \frac{0,5 - 0,41}{0,71 - 0,41} \right| = 20 + 3 = 23 \text{ years}$$

Graphical Determination of the Median (Me)

The **median** can also be determined **graphically** using a **cumulative frequency curve (ogive)**. The **median** is found at the **intersection of the horizontal line at N/2 (or 50% for relative frequency) with the cumulative curve**, and the corresponding **x-axis value gives the median**.



IV.4 Means (Averages)

We assume that **X** is a **quantitative variable** defined over a population of **N** individuals.

IV.4.1 Arithmetic Mean

The **arithmetic mean** is the **most commonly used measure of central tendency**.

However, it has the **drawback of being sensitive to outliers**, making it **less reliable in cases where extreme values are present**.

IV.4.1.1 Ungrouped Data (Simple Arithmetic Mean)

The **arithmetic mean** is defined as the **sum of all observed values divided by the total number of observations**.

One key property of the arithmetic mean is that **it preserves the total sum of the observed values**: if two observations in a statistical series are modified while keeping their sum constant, the mean of the series remains unchanged.

Given a set of N observations of the variable X :

$x_1, x_2, \dots, x_N$

The **simple arithmetic mean** of these N values, also called the **mean of variable X** is denoted as $\bar{X}$ and is calculated as:

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{N}$$

Using summation notation:

$$\sum : \bar{X} = X = \frac{1}{N} \sum_{i=1}^N X_i$$

Example 8: Evolution of a Company's Production (in thousands) from 2010 to 2017

Year	2010	2011	2012	2013	2014	2015	2016	2017
Production	5.45	5.65	5.29	5.67	5.55	4.94	5.12	5.84

$$\bar{X} = \frac{5,45+5,65+5,29+5,67+5,55+4,99+5,12+5,84}{8} = 5,445 \approx 5,45$$

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_n}{N}$$

Thus, the **arithmetic mean production over the period 2010-2017 is approximately 5.45 thousand units**.

This method provides a simple and effective way to summarize the central tendency of a dataset.

IV.4.1.2 Grouped Data

a) Case of a Discrete Variable

When the values of a **quantitative variable X** are organized into a **frequency table**, each observed value **xi** is associated with a weight (or coefficient), which corresponds to its frequency **ni** in the dataset.

The **weighted arithmetic mean** of **K** values (x1,x2,...,xK), each having an associated frequency (n1,n2,...,nK), is given by:

$$\bar{X} = \frac{n_1x_1 + n_2x_2 + \dots + n_Kx_K}{n_1 + n_2 + \dots + n_K} = \bar{X} = \frac{1}{N} \sum_{i=1}^K n_i x_i$$

Using summation notation:

$$\bar{X} = \sum_{i=1}^k f_i x_i$$

Example 9: Distribution of 10 Households by Number of Children

Number of Children (xi)	Frequency (ni)	Relative Frequency (fi)	Product ni*xi	Product fi*xi
0	2	0.2	0	0.0
1	2	0.2	2	0.2
2	1	0.1	2	0.2
3	3	0.3	9	0.9
4	2	0.2	8	0.8
Total	10	1.0	21	2.1

Using the formula:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^K n_i x_i = \frac{21}{10} = 2,1 \text{ children}$$

or using relative frequencies:

$$\bar{X} = \sum_{i=1}^K f_i x_i = 2,1 \text{ children}$$

b) Case of a Continuous Variable

To calculate the **mean of a continuous variable**, where observed values are grouped into **statistical classes**, we assume that the values are **uniformly distributed within each class**.

This assumption allows us to approximate the **mean of a class** by using its **class midpoint** (x_i), also known as the **class center**.

For a class with **midpoint x_i** and **frequency n_i** , we approximate the total sum of values in the class by:

$$\sum = n_i x_i$$

This means we **treat all values within a class as concentrated at the class midpoint**.

Example 10: Distribution of 100 People by Age Group

Age class	Frequency (n_i)	Class Midpoint (x_i)	Product $n_i * x_i$
[10 - 20[	10	15	150
[20 - 30[	30	25	750
[30 - 40[	20	35	700
[40 - 50[	40	45	1800
Total	100	/	3400

The mean is calculated as:

$$\text{Soit : } \bar{X} = \frac{3400}{100} = 34 \text{ years.}$$

Thus, the **average age** in the dataset is **34 years**.

Properties of the Arithmetic Mean

1. Sum of Deviations from the Mean is Zero

$$2. \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Proof:

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

3. The Arithmetic Mean is Easy to Compute and Interpret

- It is the most commonly used statistical measure.
- However, it is **sensitive to outliers**, making it **less reliable for skewed distributions**.

IV.4.2 Generalization of the Mean

Beyond the **arithmetic mean**, other types of means exist, each suited to specific contexts. The choice of which mean to use depends on the nature of the data and the **circumstances in which they are analyzed**.

IV.4.2.1 Geometric Mean

The **geometric mean** is widely used in **analyzing the evolution of a variable over time**, particularly when dealing with **growth rates, percentages, or multiplicative processes**.

- **Ungrouped data → Simple Geometric Mean**
- **Grouped data → Weighted Geometric Mean**

a) Simple Geometric Mean

The **geometric mean** of a dataset is the **n-th root of the product of all NNN observed values**. It is denoted as **G** and calculated using:

$$G = \sqrt[n]{x_1 x_2, \dots, x_n} = (x_1 x_2 \dots x_n)^{1/n}$$

$$\text{Ou : } \log G = \sum_{i=1}^K \frac{\log x_i}{N}$$

Example 11: Calculate the Geometric Mean of the following values: 2, 4, 8, 10

Method 1: Direct Calculation

$$G = \sqrt[4]{2 \cdot 4 \cdot 8 \cdot 10} = \sqrt[4]{640} = 5,029$$

Method 2:

$$\log G = \frac{1}{4}[\log 2 + \log 4 + \log 8 + \log 10]$$

$$\log G = \frac{1}{4}[0,30 + 0,60 + 0,90 + 1] = \frac{2,8}{4} = 0,7$$

$$G = 10^{0,7} = 5,0118$$

Method 3:

$$G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \dots x_K^{n_K}} = (x_1^{n_1} \cdot x_2^{n_2} \dots x_K^{n_K})^{1/n}$$

$$= x_1^{f_1} \cdot x_2^{f_2} \dots x_K^{f_K}$$

$$\text{Ou: } \log G = \sum_{i=1}^K \frac{(n_i \log x_i)}{N}$$

Example 12: Average Growth Rate of National Income

For the annual growth rates **7.2%, 6.3%, 7.0%, and 4.8%** (converted to 107.2, 106.3, 107, and 104.8):

$$G=107.2 \times 106.3 \times 107 \times 104.8.$$

b Application of the Geometric Mean

The geometric mean is useful when the **values of a variable multiply over time** (geometric progression). It helps estimate the **average growth rate of an evolving variable**, such as:

- **Average profit growth rate of a company over a period**
- **Average population growth rate**
- **Average rate of return on investments**

Example 12: Average Growth Rate of National Income

Over the past **four years**, the **annual growth rates** of national income were:

- **Year 1:** +7.2%
- **Year 2:** +6.3%
- **Year 3:** +7.0%
- **Year 4:** +4.8%

Find the average annual growth rate.

Solution

Using the geometric mean formula:

$$G = \sqrt[4]{107,2 \times 106,3 \times 107 \times 104,8}$$

$$\text{Log } G = \frac{1}{4} [\log 107,2 + \log 106,3 + \log 107 + \log 104,8]$$

$$\text{Log } G = 2,027$$

$$\Rightarrow G = 10^{2,027} = 106,3\%$$

During this period, the growth rate of GDP is 6.3%.

Example 13: Average Annual Growth Rate of Profits Over 9 Years

Three different management teams led a company over different periods:

- **First period (4 years):** Annual profit increased by **50%**

- **Second period (3 years):** Annual profit increased by **17%**
- **Third period (2 years):** Annual profit decreased by **30%**

Find the average annual growth rate over these 9 years.

Solution:

Using the **weighted geometric mean formula**:

$$G = \sqrt[9]{150^4 \cdot 117^3 \cdot 70^2}$$

$$\log G = \frac{1}{9} [4 \log 150 + 3 \log 117 + 2 \log 70] = 2,06$$

Or using logarithms:

$$G = 10^{2,06} = 114,8\%$$

Thus, the **average annual growth rate** is **14.8%**.

Key Remarks on the Geometric Mean

- The **geometric mean is undefined if any value is zero.**
- It is primarily used **for calculating average ratios or indices** in economic, financial, and demographic studies.

IV.4.2.2 Harmonic Mean

IV.4.2.2 Harmonic Mean

The **harmonic mean** is used **when working with ratios**, especially **for calculating average rates such as speed or price per unit.**

a) Simple Harmonic Mean

The harmonic mean (H) is defined as:

$$H = \frac{1}{\frac{1}{N} \left(\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_K} \right)} = \frac{N}{\sum_{i=1}^K \frac{1}{X_i}}$$

Example 14: Average Price of Books Over Three Years

A student spent **3,600 DA** per year on books at prices of **400 DA, 600 DA, and 900 DA** per book.

Step 1: Calculate the number of books bought per year

$$\frac{3600}{400} = 9 \text{ Books during the first year}$$

$$\frac{3600}{600} = 6 \text{ Books during the second year}$$

$$\text{Et } \frac{3600}{900} = 4 \text{ books during the third year}$$

Using the **harmonic mean formula**:

Il a donc acheté : $360 \left[\frac{1}{400} + \frac{1}{600} + \frac{1}{900} \right] = 9+6+4 = 19$ livres durant les trois années et le prix moyen d'un livre est donc :

$$H = \frac{3 \times 3600}{\left(\frac{1}{400} + \frac{1}{600} + \frac{1}{900} \right)} = \frac{3}{\left(\frac{1}{400} + \frac{1}{600} + \frac{1}{900} \right)} = 568,42 \cong 568 \text{ DA}$$

$$\frac{1}{H} = \frac{1}{10800} \left(\frac{3600}{400} + \frac{3600}{600} + \frac{3600}{900} \right) \rightarrow \frac{1}{H} = 0,00176 \rightarrow H = 568,42$$

$$\frac{1}{H} = \frac{1}{n} \sum \frac{1}{x_i}$$

Thus, the **average book price over three years is 568 DA**.

b) Weighted Harmonic Mean

For weighted data:

$$H = \frac{1}{\frac{1}{N} \left(\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_K}{x_K} \right)} = \frac{N}{\sum_{i=1}^K \frac{n_i}{x_i}} = \frac{1}{\sum_{i=1}^K \frac{f_i}{x_i}}$$

$$\frac{1}{H} = \frac{1}{n} \sum \frac{n_i}{x_i}$$

Where:

- n_{in_ini} is the frequency of each value x_i .

Key Remarks on the Harmonic Mean

- It **cannot be computed** if any observed value is zero.
- It is **mainly used** when equal investments are made in items of different values, such as determining the **average price of assets or goods**.

IV.4.2.3 Quadratic Mean (Root Mean Square - RMS)

The **quadratic mean** is used for **measuring variations, amplitudes, or deviations**.

a) Simple Quadratic Mean

The **quadratic mean** (QQQ) of a dataset is:

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_K^2}{N}} = \sqrt{\frac{1}{N} \sum_{i=1}^K x_i^2}$$

Example 15: Mean Side Length of Three Square Metal Plates

Given three square plates with side lengths **3 cm, 6 cm, and 9 cm**:

1. Compute the areas:

$$\bar{X} = \frac{9 + 36 + 81}{3} = 42 \text{ cm}^2$$

2. Compute the quadratic mean:

$$Q = \sqrt{\frac{3^2 + 6^2 + 9^2}{3}} = \sqrt{42} = 6.48 \text{ cm}$$

Thus, the **average side length is 6.48 cm**.

b) Weighted Quadratic Mean

For weighted data:

$$Q = \sqrt{\frac{n_1x_1^2 + n_2x_2^2 + \dots + n_Kx_K^2}{N}} = \sqrt{\frac{1}{N} \sum_{i=1}^K n_i x_i^2}$$

$$Q = \sqrt{\sum_{i=1}^K f_i x_i^2}$$

Comparison of the Four Means

For $x_1, x_2, \dots, x_N$ observations, when all four means **can be computed**, they follow this order:

$$X_{\min} \leq H \leq G \leq \bar{X} \leq Q \leq H_{\max}$$

Example 16: Compute Four Means for a Statistical Distribution

Class	Frequency (n_i)	Class Center (x_i)	$n_i x_i$	$\frac{n_i}{x_i}$	$\log n_i x_i$	$x_i^2 n_i$
[10-20[	10	15	150	0.66	11.76	2250
[20-30[	30	25	750	1.20	41.93	18750
[30-40[	20	35	700	0.57	30.88	24500
[40-50[	40	45	1800	0.88	66.12	81000
Total	100		3400	3.31	150.69	126500

Step 1: Arithmetic Mean ($\bar{X}$)

$$\bar{X} = \frac{\sum n_i x_i}{N} = \frac{3400}{100} = 34$$

Step 2: Geometric Mean (G)

$$- \log G = \frac{\sum_{i=1}^4 (ni \log x_i)}{\sum_{i=1}^4 ni} = \frac{150,69}{100} = 1,5 \rightarrow G = 10^{1,5} = 31,62$$

Step 3: Harmonic Mean (H)

$$- H = \frac{\sum_{i=1}^4 ni}{\sum_{i=1}^4 \left(\frac{ni}{x_i}\right)} = \frac{100}{3,31} = 30,21$$

Step 4: Quadratic Mean (Q)

$$- Q = \sqrt{\frac{\sum_{i=1}^4 (ni x_i^2)}{\sum_{i=1}^4 ni}} = \sqrt{\frac{126500}{100}} = \sqrt{1265} = 35,56$$

Final Comparison of Means

$$10 \leq 30,21 \leq 31,62 \leq 34 \leq 35,56 \leq 50$$

$$X_{min} \leq H \leq G \leq \bar{X} \leq Q \leq X_{max}$$

Conclusion:

- The **harmonic mean** is the **smallest** because it **weights smaller values more heavily**.
- The **geometric mean** lies **between the harmonic and arithmetic means**.
- The **arithmetic mean** is **larger than both the harmonic and geometric means**.
- The **quadratic mean** is the **largest**, as it gives **more weight to higher values**.

5

CHAPTER

Measures of dispersion

In this chapter :

- The range
- Quantiles
- Interquartile range
- Inter-decile range
- Inter-centile range
- Mean absolute deviation
- Variance (denoted $V(x)$)
- Standard deviation (σ_x)
- Relative dispersion coefficients
- Coefficient of variation (CV)
- Shape characteristics

V.1 Introduction: Measures of Dispersion

Often, **central tendency indicators** (mode, median, and mean) are **not sufficient** to fully summarize or compare two or more statistical series.

For example, the **median** provides a **central value** summarizing the distribution of a population, but this **summary alone does not give a complete picture of the distribution**.

Two distributions can have **similar central tendency measures** but **be very different in their spread**—one may be more **concentrated** while the other is **more dispersed**.

To refine statistical analysis, we introduce **indicators that measure dispersion**. These include:

- **Range (étendue)**
- **Quantiles**
- **Inter-decile range**
- **Standard deviation (écart-type)**

Illustrative Example

Consider two series (A) and (B), each consisting of **9 electric lamps**, with their **lifespans (in hours) recorded**:

Series (A):

780, 790, 790, 800, 800, 800, 810, 810, 820

Series (B):

400, 600, 600, 800, 800, 800, 1000, 1000, 1200

Step 1: Compute Central Tendency Indicators

For both Series **A** and **B**, we find:

- **Mode (M_o) = 800**
- **Median (M_e) = 800**
- **Arithmetic Mean ($\bar{X}$) = 800**

Step 2: Observe Dispersion

Although both series have **the same mode, median, and mean**, their **spread is different**:

- **Series A:** Lifespans are **closely grouped** around 800 hours.
- **Series B:** Lifespans vary significantly from **400 to 1200 hours**, showing **greater dispersion**.

Thus, **measures of dispersion** are necessary to **capture the variability in the data**.

Definition: Measure of Dispersion

A **measure of dispersion** is an **indicator** that estimates how **far individual observations deviate** from each other or from a **central value** (typically the **mean**).

The **most common dispersion measures** include:

1. **Range (étendue)** – the difference between the maximum and minimum values.
2. **Standard deviation (écart-type)** – measures how much values **deviate from the mean**.

These indicators help **better understand variability** in a dataset.

V.2 Absolute Measures of Dispersion

V.2.1 Dispersion Measures That Do Not Use a Central Value

Dispersion parameters such as **range** and **interquartile range** are used for **quantitative statistical variables** to describe their spread **without relying on a central value**.

V.2.1.1 The Range (Étendue)

The **range** is the **difference between the largest and smallest values** in a dataset. It is given by the formula:

$$E = \text{Max}(x_i) - \text{Min}(x_i)$$

Example: Comparing the Range in Series (A) and (B)

For **Series A**:

$$E_A = 820 - 780 = 40$$

For **Series B**:

$$E_B = 1200 - 400 = 800$$

Although the **range is simple and easy to calculate**, it is **highly sensitive to extreme values** or **outliers**, which may distort its representativeness.

V.2.1.2 Quantiles (Quartiles, Deciles, and Percentiles)

Similar to the **median** (which divides the dataset into two equal parts), **quantiles** divide the dataset into **four, ten, or one hundred equal parts**.

These quantiles are classified as:

- **Quartiles**
- **Deciles**
- **Percentiles or Centiles**

a) Quartiles

Quartiles divide the **ordered dataset** into **four equal parts**, producing **three quartile values**:

- **First quartile Q1Q_1Q1**: The value **below which 25% of observations fall**, meaning **75% of observations are above it**.

- **Rank of Q1** : $RQ1 = \sum n_i \frac{1}{4} = \frac{\sum n_i}{4}$

- **Second quartile Q2** : The **median**, dividing the dataset into **two equal halves**.

- **Rank of Q2** : $RQ1 = \sum n_i \frac{2}{4} = \frac{\sum n_i}{2}$

- **Third quartile Q3** : The value **below which 75% of observations fall**, meaning **25% of observations are above it**.

- **Rank of Q3** : $RQ1 = \sum n_i \frac{3}{4} = \frac{3 \sum n_i}{4}$

Interquartile Range (IQR) - Measuring Variability

The **interquartile range (IQR)** is the **difference between the third quartile Q3 and the first quartile Q1** :

$$IQ = Q3 - Q1$$

This interval **contains the middle 50% of observations**, removing **the influence of extreme values**.

- It provides a **better measure of dispersion** than the range when outliers are present.

Thus, quartiles and the **interquartile range** are useful for **assessing variability while reducing sensitivity to extreme values**.

Example 1: Determining the Interquartile Range (IQR)

The following table represents the **distribution of 100 employees** based on their **hourly salary (in DA)**:

Salary Class (DA)	Frequency (ni)	Upper Bound	Cumulative Frequency (NCi↑)
[140 - 150[	10	Less than 150	10
[150 - 160[	26	Less than 160	36
[160 - 170[	40	Less than 170	76
[170 - 180[	16	Less than 180	92
[180 - 190[	8	Less than 190	100
Total	100	/	/

Step 1: Calculate Q1 (First Quartile)

The **rank of Q1** is calculated as:

- $$Q_1 = \frac{\sum ni}{4} = \frac{100}{4} = 25$$

Q1 is in the class [150 - 160[(since 25 falls between 10 and 36).

The **formula for Q1** :

$$Q_1 = 150 + \left[\frac{25-10}{36-10} \right] x (160 - 150)$$

$$\rightarrow Q_1 = 155,769 \text{ DA}$$

Interpretation:

25% of employees earn **less than 155.77 DA per hour**, while 75% earn **more than 155.77 DA per hour**.

Step 2: Calculate Q3 (Third Quartile)

The **rank of Q3** is:

- $$Q_3 = \frac{\sum ni \cdot 3}{4} = \frac{100}{4} = 75$$

- Q3 is in the class [160 - 170[** (since 75 falls between 36 and 76).

- The **formula for Q3** :

$$Q_3 = 160 + \left[\frac{75-36}{76-36} \right] x (170 - 160)$$

$$\rightarrow Q_3 = 169.75 \text{ DA}$$

Interpretation:

75% of employees earn **less than 169.75 DA per hour**, while 25% earn **more than 169.75 DA per hour**.

Step 3: Compute the Interquartile Range (IQR)

$$IQ = Q_3 - Q_1$$

$$IQ = 169.75 - 155.77 = 13.98 \text{ DA}$$

Interpretation:

The middle **50% of employees** have salaries ranging between **155.77 DA and 169.75 DA per hour**.

Step 4: Compare IQR to Range

$$\frac{Q_3 - Q_1}{E} = \frac{13,981}{189 - 140} = \frac{13,981}{49} = 0,285 \text{ ou } 25,5\%$$

This means that **28% of the total salary range (50 DA) corresponds to the central 50% of employees**.

Step 5: Compute the Semi-Interquartile Range (SIQ)

$$IQ_{/2} = \frac{Q_3 - Q_1}{2}$$

This measure is also known as **quartile deviation** and gives an estimate of the average dispersion around the median.

b) Deciles (D1 D9)

Deciles divide the dataset into **10 equal parts**. The most common ones are:

- **First decile (D1):** Below which **10%** of observations fall.

$$RD1 = \sum n_i \frac{1}{10}$$

- **Ninth decile (D9):** Below which **90%** of observations fall.

$$RD1 = \sum n_i \frac{9}{10}$$

Inter-decile range:

$$ID = D9 - D1$$

This **captures the middle 80% of observations**, offering a more robust measure of dispersion than the range.

c) Percentiles (Centiles, P1 P99)

Percentiles divide the dataset into **100 equal parts**:

$$Ip = P_{99} - P_1$$

Percentiles divide the dataset into **100 equal parts**. The most common ones are:

- **First decile (P1):** Below which **01%** of observations fall.

$$RP1 = \sum n_i \frac{1}{100}$$

- **Ninth Percentile (P99):** Below which **99%** of observations fall.

$$RP99 = \sum n_i \frac{99}{100}$$

Inter- Percentiles range:

$$IP = P9 - P1$$

This **captures the middle 98 % of observations**, offering a more robust measure of dispersion than the range.

This **captures 98% of the observations**, reducing sensitivity to extreme values.

Key Remarks

- **The method for determining deciles and percentiles is the same as for quartiles and the median**, using **cumulative frequencies**.
- **IQR and inter-decile range (ID)** are **better measures of dispersion** than the **range**, as they are **less influenced by extreme values**.
- The **semi-interquartile range (SIQ)** is another robust measure of spread, useful when comparing datasets with outliers.

V.2.2 Dispersion Measures Relative to a Central Value (Mean-Based Measures)

The next section will introduce **variance, standard deviation, and coefficient of variation**, which measure **how far data points deviate from the mean**. These provide deeper insights into the distribution and variability of a dataset.

Example 2:

Let's consider the following statistics grades of a group of 6 students:

2 - 17 - 7 - 18 - 3 - 13

The mean of the grades is:

$$\bar{x} = \sum x_i / 6 = 60 / 6 = 10$$

How can we assess the dispersion of this dataset?

By calculating the differences between the observations and the mean, we obtain :

x_i	2	17	7	18	3	13	Σ
(x_i - $\bar{x}$)	-8	+7	-3	+8	-7	+3	0

Since the sum of these differences is zero:

$$\sum_{i=1}^6 (x_i - \bar{x})$$

This is one of the properties of the arithmetic mean: **the sum of the deviations of the observations from their mean is zero** $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

However, to make this calculation useful, we take the absolute values of these differences.

V.2.2.1 Mean Absolute Deviation (MAD)

An alternative approach is to use the absolute values of these deviations:

$$|x_i - \bar{x}|$$

Thus, we compute the mean of these absolute deviations, which effectively represents the dispersion of the data :

$$e = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

for **ungrouped data**, and

$$e = \frac{1}{n} \sum_{i=1}^n n_i |x_i - \bar{x}|$$

for **grouped data**.

However, since absolute values are not convenient for algebraic calculations, most textbooks avoid using them. Instead, we prefer to use the squares of the differences $(x_i - \bar{x})^2$. The arithmetic mean of these squared differences is called the **variance**.

V.2.2.2 Variance $V(x)$

For **ungrouped data**:

$$V(x) = \frac{\sum (x_i - \bar{x})^2}{N} = \frac{1}{N} \cdot \sum (x_i - \bar{x})^2$$

For **grouped data**:

$$V(x) = \frac{\sum n_i (x_i - \bar{x})^2}{\sum n_i} = \frac{1}{\sum n_i} \cdot \sum n_i (x_i - \bar{x})^2$$

Using **relative frequencies**:

$$V(x) = \sum f_i \cdot x_i^2 - \bar{x}^2$$

Note

Variance is expressed in the square of the unit of the variable. For example, the variance of salary is expressed in **square dinars (DA²)**. For this reason, variance itself is not directly interpretable, but its square root is.

V.2.2.3 Standard Deviation σ_x

The **standard deviation**, denoted σ_x , is the square root of the variance. It serves as an indicator of dispersion in a statistical dataset:

$$\sigma_x = \sqrt{V(x)}$$

Note :

- The standard deviation is expressed in the same unit as the variable.
- A higher standard deviation indicates greater dispersion around the mean.

Example 2 (continued): Calculation of Standard Deviation

We compute the standard deviation for the given dataset:

$$\sigma_x = \sqrt{\frac{1}{6} \sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{6} [(-8)^2 + 7^2 + (-3)^2 + 8^2 + (-7)^2 + 3^2]}$$

$$\sigma_x = \sqrt{40,66} = 6,37$$

Interpretation:

- Good students will have approximately the **mean grade (10) plus 6.37**.
- Weaker students will have approximately the **mean grade (10) minus 6.37**.

Example 3:

Let's consider the following statistics grades of a second group of 4 students:
8 - 12 - 9 - 11

Step 1: Compute the Mean

$$\bar{x} = \frac{1}{4} \sum x_i = \frac{40}{4} = 10$$

Step 2: Compute the Variance

$$V(x)_2 = \frac{1}{4} \sum (x_i - \bar{x})^2$$

$$V(x)_2 = \frac{1}{4} [(8 - 10)^2 + (12 - 10)^2 + (9 - 10)^2 + (11 - 10)^2]$$

$$V(x)_2 = 10/4 = 2,5 \quad \sqrt{2,5} = 1,581$$

$$\rightarrow \sigma_{x_2} = \sqrt{V(x)_2} = \sqrt{10} = 3,16$$

Comparison:

- **Example 2:** $\sigma_x = 6.37$
- **Example 3:** $\sigma_x = 3.16$

The dispersion of grades in **Example 2 is twice as high as in Example 3**. This means that the second group of students is **more homogeneous** than the first group.

Koenig's Formula for Variance and Standard Deviation (Practical Methods)

In **Examples 2 and 3**, calculations were straightforward because the mean $\bar{x} = 10$ was an integer, making it easier to compute squared differences. However, in general cases, variance calculations can be tedious, and we use a more practical formula:

Koenig's Formulas:

- **For ungrouped data:**

$$V(x) = \frac{\sum x_i^2}{n} - \bar{x}^2$$

- **For grouped data:**

$$V(\kappa) = \frac{\sum n_i \kappa_i^2}{\sum n_i} - \bar{\kappa}^2$$

- Using relative frequencies:

$$V(\kappa) = \sum f_i \kappa_i^2 - \bar{\kappa}^2$$

Key Relationship:

$$\text{Since } \sigma_{\kappa} = \sqrt{V(\kappa)} \Rightarrow V(\kappa) = \sigma_{\kappa}^2$$

Example 4:

Given the data: **6 - 9 - 10 - 11 - 15**, we will calculate the **standard deviation** using both the **definition formula** and **Koenig's formula**.

κ_i	$\kappa_i - \bar{\kappa}$	$(\kappa_i - \bar{\kappa})^2$	κ_i^2
6	-4.2	17.64	36
9	-1.2	1.44	81
10	-0.2	0.04	100
11	0.8	0.64	121
15	4.8	23.03	225
Total	51	0.00	42.8

Step 1: Compute the Mean

$$\bar{\kappa} = \frac{\sum \kappa_i}{n} = \frac{51}{5} = 10,2$$

Step 2: Compute the Variance Using the Definition Formula

$$\sigma_{\kappa}^2 = \frac{1}{n} \cdot \sum (\kappa_i - \bar{\kappa})^2 = \frac{42,8}{5} = 8,56$$

$$\Rightarrow \sigma_{\kappa} = \sqrt{8,56} = 2,92$$

Step 3: Compute the Variance Using Koenig's Formula

$$\sigma_{\kappa}^2 = \frac{\sum \kappa_i^2}{n} - \bar{\kappa}^2 = \frac{563}{5} - 10,2^2 = 8,56$$

$$\Rightarrow \sigma_{\kappa} = \sqrt{8,56} = 2,92 \text{ Même résultat}$$

Same result using both methods.

Example 5 :

Given the grouped data:

Class	ni	xi	ni.xi	xi- $\bar{x}$	(xi - $\bar{x}$) ²	ni. (xi - $\bar{x}$) ²	ni.xi ²
[10-20[	3	15	45	-10	100	300	675
[20-30[	4	25	100	0	0	0	2500
[30-40[	3	35	105	+10	100	300	3675
Total	10	/	250	/	/	600	6850

Step 1: Compute the Mean

$$\bar{x} = \frac{\sum ni.xi}{N} = \frac{250}{10} = 25$$

Step 2: Compute the Variance Using the Definition Formula

$$\sigma_x^2 = \frac{\sum ni.(xi - \bar{x})^2}{\sum ni} = \frac{600}{10} = 60$$

$$\Rightarrow \sigma_x = \sqrt{60} = 7,745$$

Step 3: Compute the Variance Using Koenig's Formula

$$\sigma_x^2 = \frac{\sum ni.xi^2}{\sum ni} - \bar{x}^2 = \frac{6850}{10} - (25)^2 = 60$$

$$\Rightarrow \sigma_x = \sqrt{60} = 7,745$$

Same result using both methods.

V.3 Relative Dispersion Coefficients

Since the **standard deviation is expressed in the same unit as the variable**, it can be difficult to compare distributions with different units (e.g., salaries in **Dinars** vs. **Euros**). To **eliminate the influence of the measurement unit**, we calculate a **dimensionless relative dispersion coefficient**.

V.3.1 Interquartile Coefficient of Dispersion

$$CQ = \frac{Q_3 - Q_1}{Q_2} \times 100$$

This measures the **spread of the middle 50% of the data** relative to the median.

V.3.2 Coefficient of Variation (CV)

The **coefficient of variation (CV)** is the most commonly used measure when comparing two or more statistical distributions.

Why Use the Coefficient of Variation ?

- When distributions have **different units** or **different orders of magnitude**, comparing the **standard deviation alone is meaningless**.
- The **CV is a dimensionless ratio** that allows **standardized comparisons** between datasets, even if measured in different units.
- It is particularly useful in **economic comparisons**, such as comparing salaries in **different currencies** (e.g., Dirhams vs. Euros).

Formula:

- $\alpha = CV = \frac{\sigma x}{\bar{x}} \times 100$
- The higher the **CV**, the **greater the dispersion**.
- A **lower CV** indicates **higher homogeneity** in the data.

Example : Comparing Two Companies

- **Company A:** CVA=40%
- **Company B:** CVB=35%

Since $CV_A > CV_B$, we conclude that **Company A has more dispersion in salaries than Company B**.

If within **the same company**, we observe:

- **Workers' CV** = 41%
- **Employees' CV** = 78%

We conclude that **workers' salaries are more homogeneous** than **employees' salaries**, as **employees show greater dispersion**.

Exercise 6: Comparing Salaries in Algeria and France

We are given the **mean salary** ($\bar{x}$) and **standard deviation** (σx) for each country:

Country	Mean Salary ($\bar{x}$)	Standard Deviation (σx)
Algeria	25,000 DA	350 DA
France	7,000 EUR	25 EUR

Since salaries are expressed in **different currencies**, we cannot compare **standard deviations directly**. Instead, we compute the **CV** :

Step 1: Compute CV for Algeria

$$CV_{\text{Algérie}} = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{350}{25000} \times 100 = 1,4\%$$

Step 2: Compute CV for France

$$CV_{\text{France}} = \frac{25}{7000} \times 100 = 0,357\%$$

Step 3: Compare the Results

$$VC_F < CV_{\text{Algérie}}$$

Since the **coefficient of variation is lower in France**, we conclude that **French salaries are more homogeneous**, whereas **Algerian salaries show greater dispersion**.

V.4 Shape Characteristics (Caractéristiques de Forme)

These characteristics **describe the general shape of a distribution** without the need for plotting its frequency curve.

1. Symmetric Distribution

A distribution is **symmetric** if observations are **equally dispersed on both sides** of the central value.

- In a **perfectly symmetric distribution**, the following measures are equal:

$$\bar{x} = M_0 = M_e$$

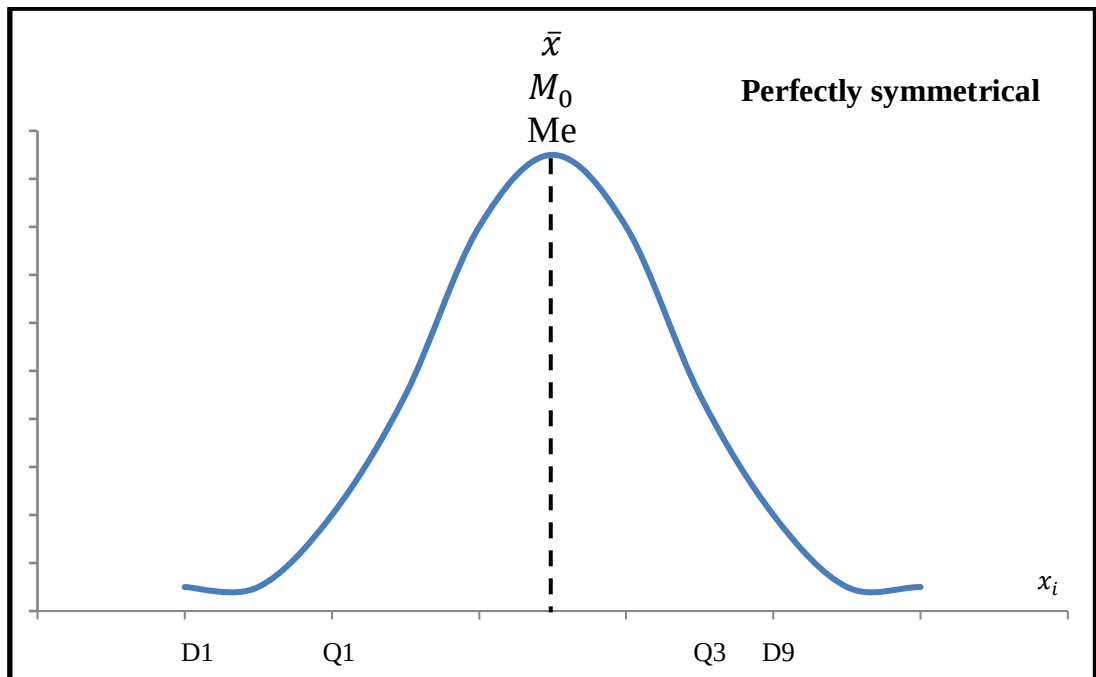
where:

- $\bar{x}$ = **Mean**
- M_0 = **Mode**
- M_e = **Median**
- Additionally, quartiles, deciles, and percentiles are **equidistant from the central value**.

□ Example:

- **Normal distribution (Gaussian curve)** is a common example of a **symmetric** distribution

ni



2. Asymmetric (Skewed) Distribution

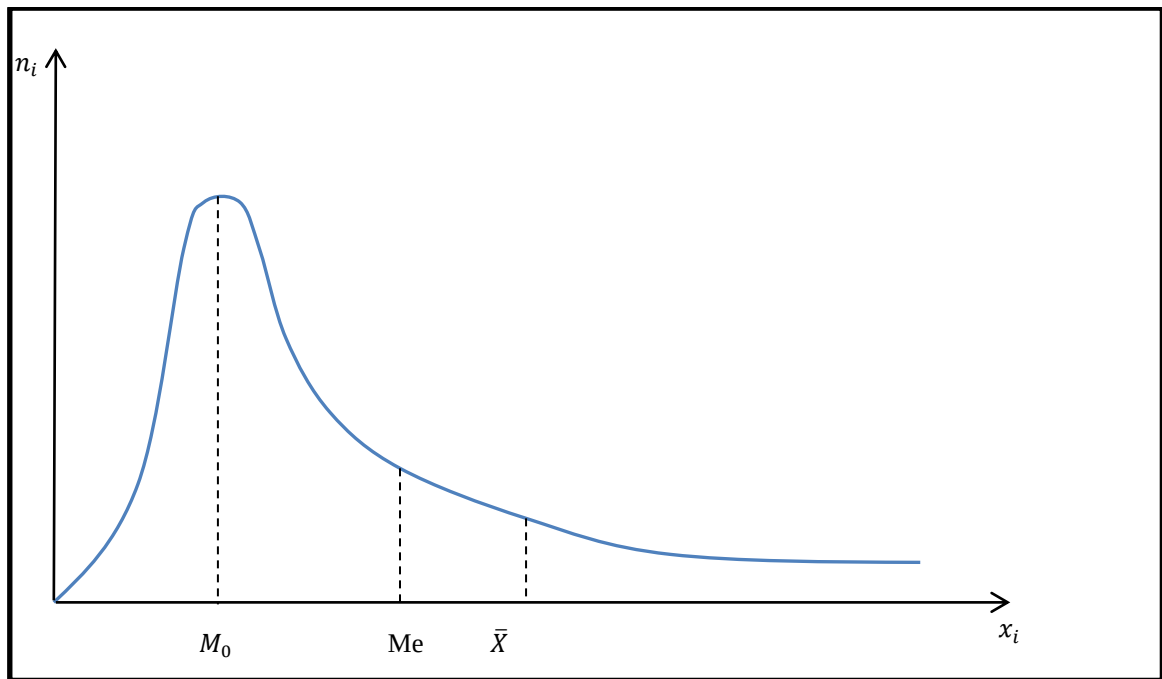
A distribution is **asymmetric** if observations are **not equally dispersed on both sides of the central value**.

This is also called **skewness (dissymétrie)**.

- **Right-skewed (positively skewed) distribution:**
 - **Spread extends more to the right** (higher values).
 - The **mean is greater than the median, which is greater than the mode:**
 $M_0 < Me < \bar{x}$
 - **Example:** Income distribution (a few people earn significantly more than the rest).

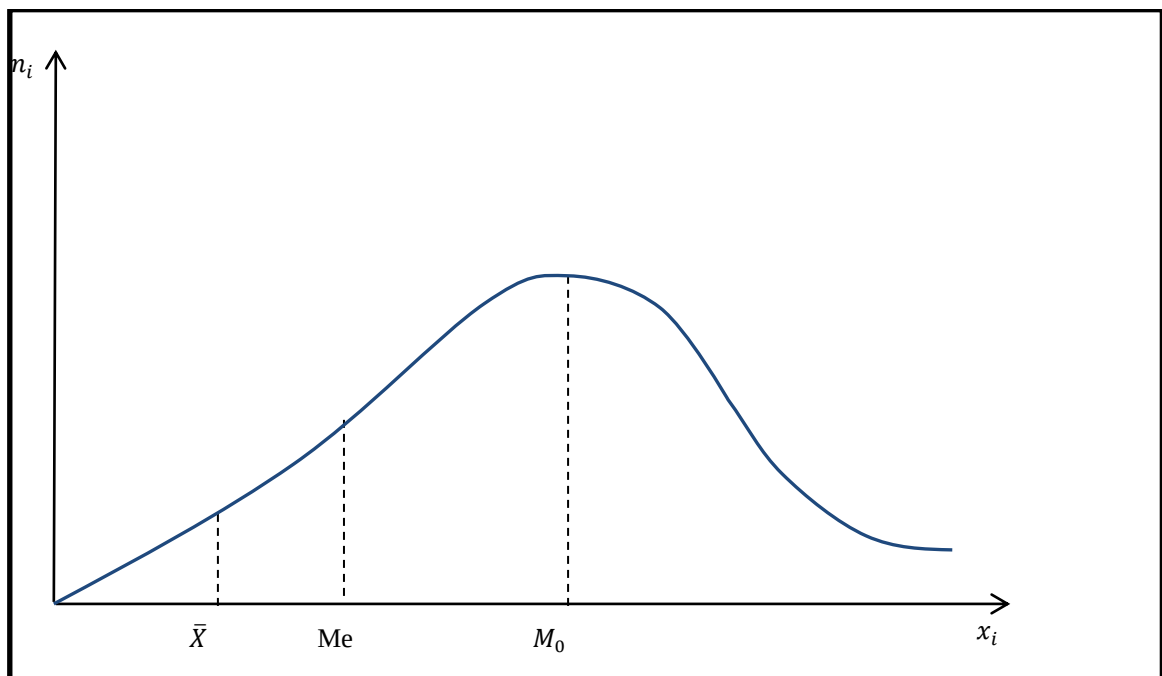
□ Interpretation:

- In **right-skewed** distributions, most values are concentrated on the **left** with a **longer tail on the right**.



3. Asymmetric (Skewed) Distribution

Right -skewed or positively skewed curve (Spread to the Right) $\bar{x} < Me < M_0$



To characterize the asymmetry of the dataset, we use Pearson's coefficient in this course, which is based on the **mean**, **mode**, and **standard deviation**:

- If $cp_1 = \frac{\bar{x} - M_0}{\sigma x}$ avec $-1 < cp_1 < +1$
- If $cp_1=0 \Rightarrow$ **Symmetric curve**
- If $cp_1>0 \Rightarrow$ **Right-skewed curve** ($\bar{x} > M_0$)
- If $cp_1<0 \Rightarrow$ **Left-skewed curve** ($\bar{x} < M_0$)

Note:

For **slightly asymmetric unimodal distributions**, the three central tendency values are related by Pearson's **empirical formula**:

$$\bar{x} - M_0 = 3(\bar{x} - Me)$$

This equation allows the calculation of one of these values if the other two are known.

Using this relationship, we can derive a **second Pearson asymmetry coefficient**:

- $Cp_2 = \frac{3.(\bar{x} - M_0)}{\sigma x}$ $-1 < cp_2 < +1$
- If $cp_2=0 \Rightarrow$ **Symmetric curve**
- If $cp_2>0 \Rightarrow$ **Right-skewed (positive asymmetry)**
- If $cp_2<0 \Rightarrow$ **Left-skewed (negative asymmetry)**

Example 7:

xi	2	3	4	5	6	7	8	Σ
ni	1	6	15	20	15	6	1	64

Step 1: Compute Pearson's Skewness Coefficient Algebraically

$$cp_1 = \frac{\bar{x} - M_0}{\sigma x}$$

$$\bar{x} = \frac{\sum ni \cdot xi}{\sum ni} = \frac{320}{64} = 5$$

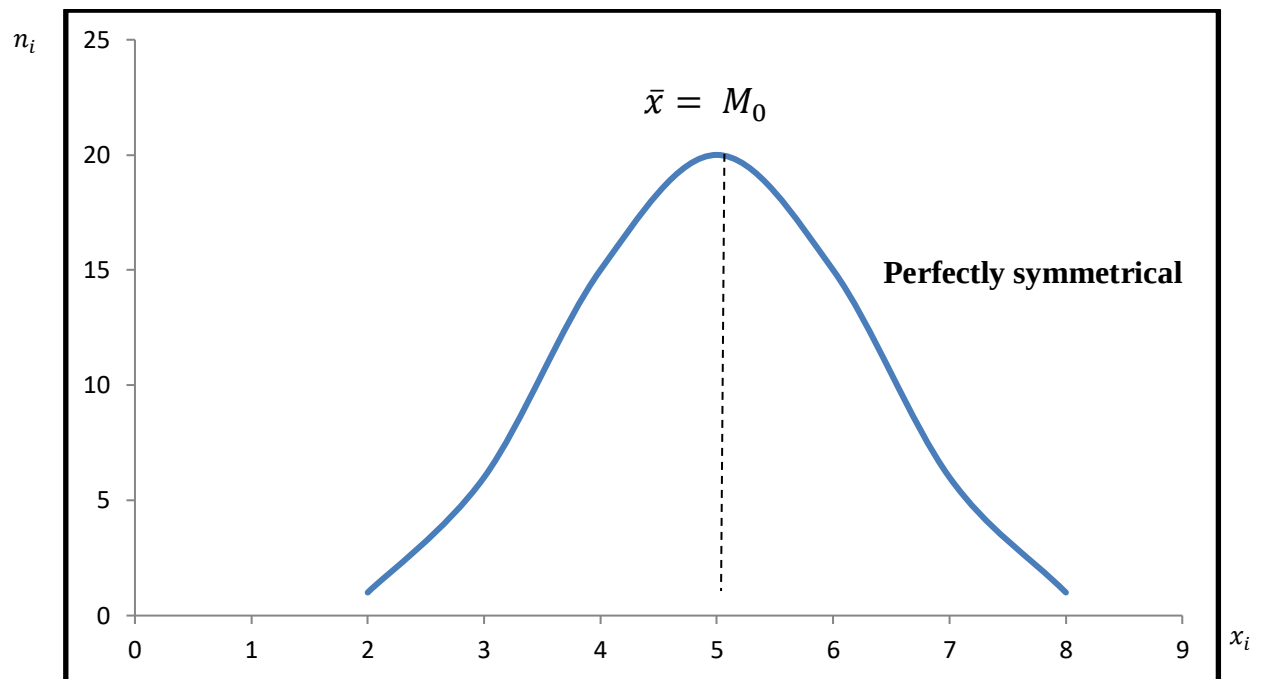
$$M_0 = 5$$

$$\Rightarrow cp_1 = \frac{5-5}{\sigma x} = 0$$

Thus, the **distribution is symmetric**.

Step 2: Graphical Interpretation

Graphically, the curve is **perfectly symmetrical**, confirming the algebraic result.



We join the coordinates of points n_i and x_i . Then, we represent the mode and the mean in the graph.

On joint les points de coordonnées n_i et x_i . Puis on représente dans le graphe le mode et la moyenne.

Example 8 :

Classes	10-20	20-30	30-40	40-50	50-60	60-70	70-80
n_i	11	19	21	30	10	5	4

Task to do : Same question as in Example 7.

$$\sum n_i = 100 \quad \sum n_i x_i = 3900$$

$$\sum n_i x_i^2 = 174700$$

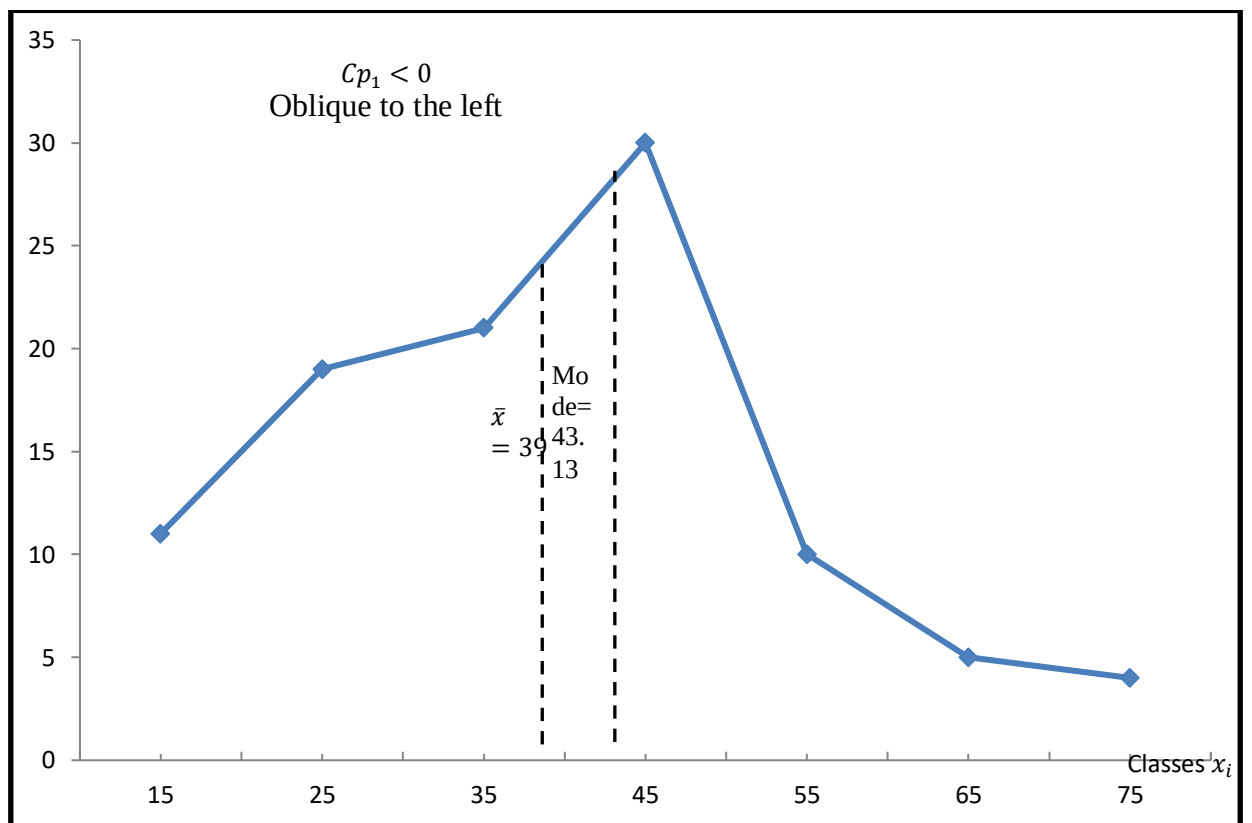
After calculations, we find:

$$\bar{x} = 39 \quad M_o = 43,103 \quad \sigma_x = 15,033$$

$$cp_1 = \frac{39 - 43,103}{15,033} = -0,272 < 0$$

$Cp_1 < 0$ Oblique to the left $\Rightarrow$ Skewed to the left.

Graphically



6

CHAPTER

Measures of shape

In this chapter :

- Skewness

- The principle of skewness
- Skewness indicators

- Kurtosis

- The principle of kurtosis
- Kurtosis indicators

VI.1 Introduction

To **characterize, measure, or summarize** statistical distributions, we rely on:

- **Measures of central tendency**, which provide an overall estimate of the distribution.
- **Measures of dispersion**, which indicate the degree of spread or concentration of observations around a central value (usually the arithmetic mean).

However, to determine whether the **shape of the frequency curve** of two or more distributions is similar, we need to compute **shape parameters**, which measure the **degree of similarity between distributions**.

In reality, this analysis helps determine **whether one distribution is more or less symmetric than another**.

VI.2 Asymmetry

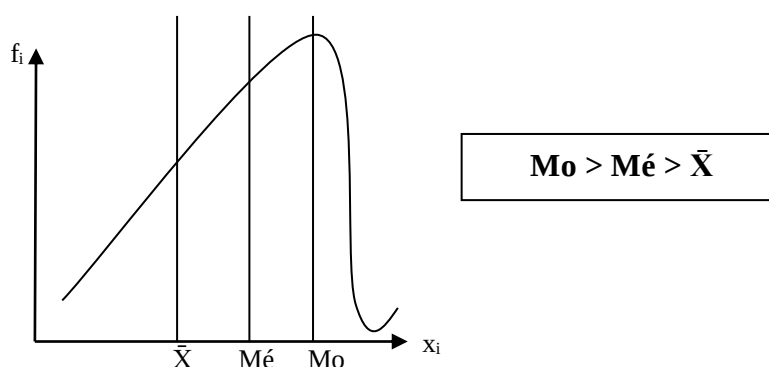
VI.2.1 The Principle of Asymmetry

The asymmetry of a statistical distribution is analyzed **relative to its central value**. It aims to evaluate whether the **distribution is more spread to the left or to the right**, or whether **observations are equally distributed on both sides** of this central value.

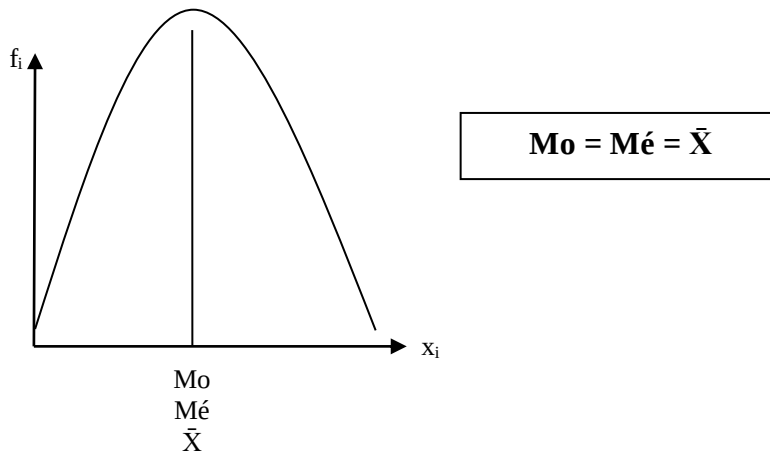
Asymmetry causes the **three central tendency measures** (mode, median, and mean) **to no longer coincide at the center** of the distribution. Instead, they **follow a specific order** depending on whether the distribution is **left-skewed or right-skewed**.

The typical positioning of these measures is as follows:

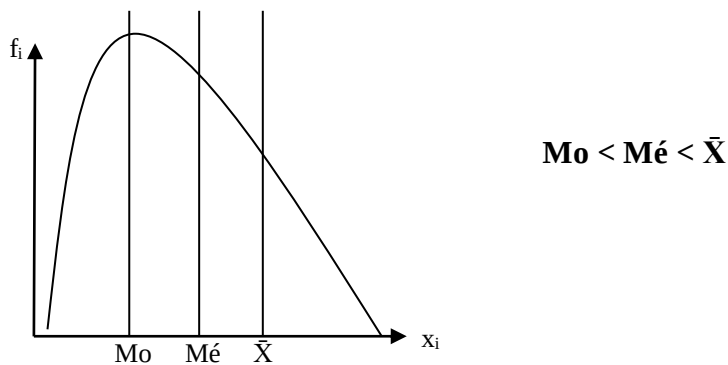
- **Left-skewed distribution (spread to the left)**



- **Symmetric distribution**



- **Right-skewed distribution (spread to the right)**



VI.2.2 Asymmetry Indicators

Different asymmetry indicators aim to **evaluate asymmetry** and are mainly useful for **comparing distributions**. These indicators are based on the effects of **distribution asymmetry** on the **relative positions of central tendency measures and quantiles**.

A- The Quartile Skewness Coefficient (Yule's Coefficient)

This coefficient is given by:

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

It is based on the comparison of the **relative positions of the three quartiles**, particularly the **first and third quartiles with respect to the median**. This comparison helps assess the **spread of the distribution on the left and right**.

Using the same principle, one could calculate a **decile or percentile skewness coefficient** by replacing quartiles with **deciles or percentiles**.

For example, a **decile skewness coefficient** would be:

$$Cd = \frac{(D9-D5) - (D5-D1)}{(D9-D5) + (D5-D1)} = \frac{(D9-D5) - (D5-D1)}{(D9-D1)}$$

- If we focus on **asymmetry near the median (C_q)**, we analyze the **center of the distribution**.
- If we focus on **asymmetry at the extremes (C_d, C_c)**, we analyze values **further from the median**.

Interpretation:

- If **C = 0** → The distribution is **symmetric**.
- If **C < 0** → The distribution is **left-skewed**.
- If **C > 0** → The distribution is **right-skewed**.

B- Pearson's First Skewness Coefficient

This coefficient **compares the relative positions of the mode and the mean**, considering the **dispersion**. It is defined as:

$$Cp1 = \frac{x - Mo}{\sigma(x)}$$

Interpretation:

- If **Cp1=0** → The distribution is **symmetric**.
- If **Cp1<0** → The distribution is **left-skewed**.
- If **Cp1>0** → The distribution is **right-skewed**.

Note: This coefficient is **only valid for weakly asymmetric distributions**.

C- Pearson's Second Skewness Coefficient

Pearson's **second skewness coefficient** is given by:

$$Cp2 = \frac{(M^3_x)^2}{(M^2_x)^3}$$

It is **defined as the ratio of the square of the third central moment to the cube of the second central moment (variance)**:

$$(M^3_x) = m_3 - 3m_1m_2 + 2(m_1)^3$$

$$(M^2_X) = V(X) = m_2 - (m_1)^2$$

where m_i represents the simple moment of order i .

Interpretation:

- A symmetric distribution will have $C_p=0$.
- An asymmetric distribution will have a positive C_p .

Example:

Let's calculate the asymmetry indicators for the height distribution (in meters) of a population of 24 students :

Height (m)	Frequency (n _i)	Cumulative Frequency (N _{ci} ↑)	Class Center (x _i)	n _i * x _i	n _i (x _i - $\bar{X}$) ²	n _i (x _i - $\bar{X}$) ³
< 1.60	6	6	1.55	9.3	0.1134	-0.0156
[1.60 - 1.70]	7	13	1.65	11.55	0.0098	-0.0003
[1.70 - 1.80]	8	21	1.75	14	0.0312	0.0019
[1.80 - 1.90]	2	23	1.85	3.7	0.0528	0.0085
≥ 1.90	1	24	1.95	1.95	0.0689	0.018
Σ	24	/	/	40.5	0.2761	0.0125

Quartiles Calculation:

$$Q_1 = 1,60$$

$$Q_2 = 1,60 + \frac{(1,20-6)(1,70-1,60)}{(13-6)} = 1,685 \text{ m.}$$

$$\bullet \quad Q_3 = 1,70 + \frac{(18-13)(1,80-1,70)}{(21-13)} = 1,7625 \text{ m.}$$

1. Quartile Skewness Coefficient (Yule's Coefficient)

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

$$C_q = \frac{(1,7625 - 1,685) - (1,685 - 1,60)}{(1,7625 - 1,60)} = -0,0461.$$

2. Pearson's First Skewness Coefficient

$$C_{p1} = \frac{x - Mo}{\sigma(x)} = \frac{1,6875 - 1,75}{0,1072} = -0,583$$

3. Pearson's Second Skewness Coefficient

$$C_{p2} = \frac{(M^3_x)^2}{(M^2_x)^3} = \frac{(0,0005)^2}{(0,0115)^2} = 0,1643$$

Conclusion:

All **four indicators agree** and **confirm that the distribution is asymmetric** (since none of them equals zero).

VI.3 Kurtosis (Aplatissement)

VI.3.1 Concept of Kurtosis

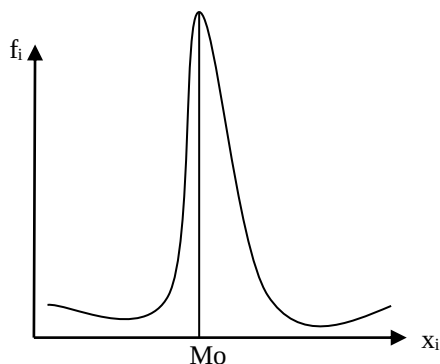
A **statistical distribution** can be **more or less flattened** depending on the proportion of observations that are **concentrated near the mode**.

- The more values are concentrated around the mode, the lower the kurtosis.
- The more spread out the values are, the higher the kurtosis.

Graphical Interpretation:

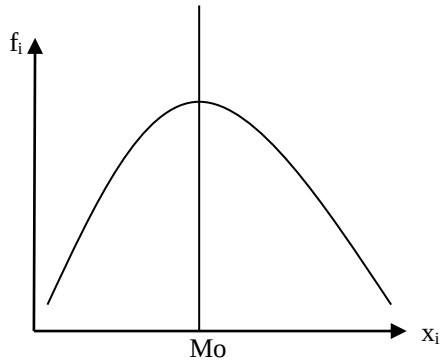
1/ Low Kurtosis (Flat Distribution)

- Observations are widely spread around the mode.
- Example: **Uniform distribution**.



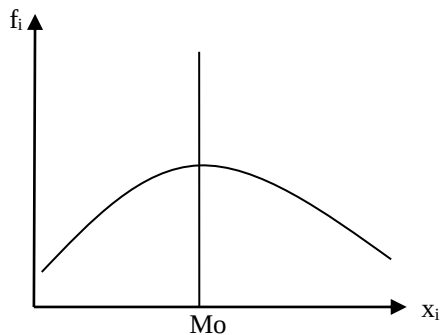
2/ Medium Kurtosis (Normal Distribution - Mesokurtic)

- Example: **Normal (Gaussian) distribution**.



3/ High Kurtosis (Peaked Distribution - Leptokurtic)

- Observations are highly concentrated around the mode.
- Example: **T-distribution with low degrees of freedom**.



VI.3.2 Kurtosis Indicators

A- Kurtosis Coefficient

The kurtosis coefficient **measures the degree of peakness or flatness** of a distribution.

It is calculated using the **fourth central moment ($M_4\bar{X}$)** relative to the mean and the **standard deviation**, and is expressed as:

$$a_1 = \frac{(M_4\bar{X})}{\sigma(x)^4}$$

This coefficient is also sometimes called **Pearson's kurtosis coefficient**.

- For a **normal distribution**, $a_1 = 3$.

- If $a_i < 3$, the distribution is **more flattened (platykurtic)**, and the **lower a_i , the stronger the flattening** (values closer to 1 indicate extreme flattening).
- If $a_i > 3$, the distribution is **more peaked (leptokurtic)**, and the **higher a_i , the less flattening is observed**.

B- Quartile Kurtosis Coefficient

This coefficient is based on **quartiles and deciles** and is calculated as:

$$a_2 = \frac{Q}{(D_9 - D_1)}$$

where:

$$Q = \frac{(Q_3 - Q_1)}{2}$$

- The **higher a_2 , the more flattened** the distribution is.
- This coefficient **always lies between 0 and 1**.

Example:

We calculate the kurtosis indicators for the **height distribution (in meters) of a population of 24 students**:

Height (m)	Frequency (n_i)	Cumulative Frequency (N_i)	Class Center (x_i)	$n_i * x_i$	$n_i (x_i - \bar{X})^2$	$n_i (x_i - \bar{X})^4$
< 1.60	6	6	1.55	9.3	0.1134	0.0021
[1.60 - 1.70]	7	13	1.65	11.55	0.0098	0
[1.70 - 1.80]	8	21	1.75	14	0.0312	0.0001
[1.80 - 1.90]	2	23	1.85	3.7	0.0528	0.0014
≥ 1.90	1	24	1.95	1.95	0.0689	0.0047
Σ	24	/	/	40.5	0.2761	0.0083

Step 1: Compute Pearson's Kurtosis Coefficient

$$\bar{X} = 1,6875 \text{ m}$$

$$\sigma(x) = 0,1072 \text{ m}$$

$$a_1 = \frac{M^4_{\bar{X}}}{\sigma(x)^4}$$

$$M^4_{\bar{X}} = \frac{\sum n_i (x_i - \bar{X})^4}{\sum n_i} = 0,0003 \text{ m}^4$$

$$Q_1 = 1,60 \text{ m}$$

$$Q_3 = 1,7625 \text{ m}$$

$$D_1 = 1,54 \text{ m}$$

$$D_9 = 1,83 \text{ m}$$

$$a_1 = \frac{(M^4_{\bar{x}})}{\sigma(x)^4} = \frac{0,0003}{(0,1072)^4} = 2,2716$$

Step 2: Compute the Quartile Kurtosis Coefficient

$$a_2 = \frac{Q}{(D_9 - D_1)} \quad \text{avec } Q = \frac{(Q_3 - Q_1)}{2}$$

$$a_2 = \frac{\frac{Q_3 - Q_1}{2}}{(D_9 - D_1)} = \frac{\frac{1,7625 - 1,60}{2}}{(1,83 - 1,54)} = 0,2801$$

Conclusion :

- **Pearson's kurtosis coefficient $a_1 = 2.2716$** indicates that the **distribution is flatter than a normal distribution** (since it is lower than 3).
- **The quartile kurtosis coefficient $a_2 = 0.2801$** also confirms that the **distribution is more flattened than a normal distribution**, as $a_2 > 0.263$.

Thus, both indicators **agree** that the distribution is **less peaked than a normal distribution (platykurtic)**.

7

CHAPTER

Measures of concentration

In this chapter :

- Definition of the midrange
- Median-Midrange deviation
- Concentration curve (Visual method)
- GINI Index (Calculation method)

VII.1 Introduction

The concept of **concentration** is generally related to **dispersion**, but instead of measuring deviations between observed variables and a central tendency parameter, it aims to **characterize the difference between an observed distribution and a purely theoretical equal distribution**.

Concentration is used to **describe economic entities** classified based on **quantitative characteristics** whose values can be summed. The corresponding statistical variables are **continuous, positive values**, calculated using the **median, medial, or graphical representation**.

Examples include:

- **Economic entities classified by revenue, number of employees, production, cultivated land area, population, etc.**

The **notion of concentration** originates from the work of **Corrado Gini**, an Italian statistician, who studied the **distribution of wages and incomes**, which have always been at the heart of social struggles.

The **Gini index** is particularly useful because it **measures the degree of inequality** in the distribution of a variable, **independent of the absolute level of values**.

Concentration can be determined **graphically** (using the **Lorenz Curve**) or **mathematically** (using the **Gini Index**).

VII.2 Definition of the Medial

While the concepts of **median and medial** are closely related in both **definition and calculation**, they should not be confused.

- **The median** is the value of the variable such that **half of the individuals** have values below it.
- **The medial** is the value of the variable such that **half of the total mass of the characteristic** lies below it.

By "mass of the characteristic," we mean the products ($n_i * x_i$), which represent the **total contribution of each modality x_i** .

Example:

Let's consider the distribution of **16 women** in a village according to their **number of children** :

Number of children per woman (x_i)	Number of women (n_i)	Total children ($n_i * x_i$)
0	1	0
1	3	3
2	8	16
3	4	12
Total	16	31

If we calculate $n_i x_i$, the products of the frequencies n_i by the variable values x_i , these values represent the **total number of children** in each category.

In total, **16 women have 31 children**.

- The **8 women who have 2 children** together have **16 children**, which represents:

$$\left(\frac{16}{31} = 51,61\%\right)$$

of the **total number of children** in the population.

Thus, the value **$x_i=2$ children per woman** represents **51.61% of the total mass of the characteristic**.

The **medial** is generally denoted as **ML**.

II- Example of Medial Calculation

The **medial** is calculated similarly to the **median**, with one key difference:

- Instead of using the **cumulative frequencies of the number of individuals (n_i)**,
- We use the **cumulative frequencies of the total mass of the characteristic ($n_i * x_i$)**, i.e., the **products**.

Calculation Steps:

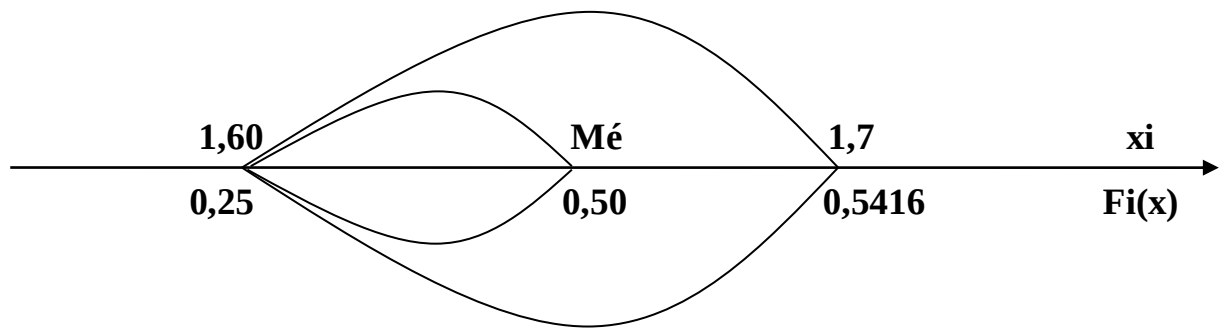
- Compute the series of ($n_i * x_i$)** (i.e., the total contribution of each class to the characteristic).
- Calculate the medial using this series**, as if computing the median from the cumulative series of (n_i).

Example: Calculating the Median and Medial

We analyze the **height distribution (in meters) of 24 students** for comparison.

1. Calculating the Median

Height (m)	Frequency n_i	Cumulative Frequency (N_i)	Cumulative Frequency $F(x)$
< 1.60	6	6	0.25
[1.60 – 1.70[	7	13	0.5416
[1.70 – 1.80[	8	21	0.875
[1.80 – 1.90[	2	23	0.9583
≥ 1.90	1	24	1.00
Σ	24	/	/



The **median class** is **[1.60; 1.70[** m.

Applying **linear interpolation**:

$$\frac{Mé - 1,60}{1,70 - 1,60} = \frac{0,50 - 0,25}{0,5416 - 0,25}$$

$$Mé = 1,60 + \frac{(0,50 - 0,25)(1,70 - 1,60)}{(0,5416 - 0,25)} = \mathbf{1,6857}$$

2. Calculating the Medial

Height (m)	Class Center (xi)	Frequency (ni)	Total Contribution (ni*xi)	Proportion of Total Contribution (ni xi) $= \frac{(ni \ xi)}{\sum_i (ni \ xi)}$	Cumulative Proportion of Total Contribution (ni xi) ↑
< 1.60	1.55	6	9.3	0.2296	0.2296
[1.60 – 1.70[	1.65	7	11.55	0.2852	0.5148
[1.70 – 1.80[	1.75	8	14	0.3456	0.8604
[1.80 – 1.90[	1.85	2	3.7	0.0913	0.9517
≥ 1.90	1.95	1	1.95	0.0481	1.00
Σ	/	24	40.5	1.00	/

Interpretation of Columns:

- The **nixi column** represents the **total height** of students within each height class.
 - Example: The **8 students** in the **[1.70 – 1.80]** m class **collectively measure 14 meters**.
 - The **total height** of the **24 students** is **40.5 meters**.
- The **cumulative proportion column** shows the **percentage of total height** accounted for by each class.
 - Example: Students **under 1.80 meters** contribute **86.04% of the total height**.

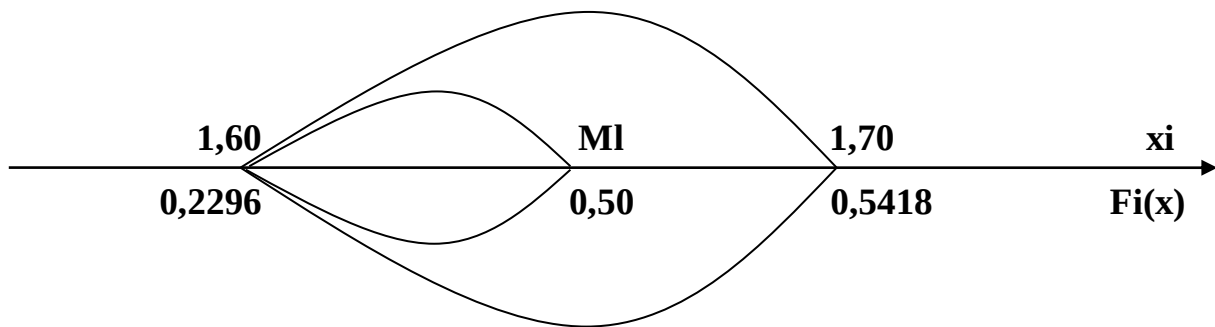
3. Computing the Medial

The **medial class** is [1.60 – 1.70[m.

Applying **linear interpolation**:

$$\frac{ML - 1,60}{1,70 - 1,60} = \frac{0,50 - 0,2269}{0,5148 - 0,2296}$$

$$ML = 1,60 + \frac{(0,50 - 0,2296)(1,70 - 1,60)}{(0,5148 - 0,2296)} = 1,6948 \text{ m}$$



Final Interpretation:

- Since the **median** is 1.6875 m, we conclude that **half of the students are shorter than 1.6875 m.**
- Since the **medial** is 1.6948 m, we conclude that **half of the total height of all students is accounted for by students shorter than 1.6948 m.**

VII.3 Median-Medial Gap

The **medial** is always **greater than the median**. However, rather than the medial itself, the **gap between the median and the medial** is the key **indicator of concentration** to consider. This gap is denoted as ΔM :

$$\Delta M = | ML - Me |$$

- ΔM is always **positive**.
- To **assess the concentration level** of a distribution, ΔM is **compared to its range or variation interval**, i.e., the **difference between the smallest and largest values of the variable in the distribution**.

- The larger ΔM is relative to the range, the higher the concentration of the distribution.

Example (from the previous dataset):

- $\Delta M = |1,6948 - 1,6857| = 0,0091 \text{ m.}$

- The **range** of the distribution is [1.50 ; 2], so :

$$2 - 1.50 = 0.50 \text{ m}$$

$$\frac{0,0091}{0,50} = 0,0182$$

Example: Salary Distribution in a Company

The following table presents the **hourly wage distribution** of 100 employees in company B

Salary Classes (DA)	Frequency (ni)	Class Center (xi)	Relative Frequency (fi)	Cumulative Frequency F(x)	Total Contribution (n*ixi)	Proportion of Total Contribution (ni xi) = $\frac{(ni \cdot xi)}{\sum_i (ni \cdot xi)}$	Cumulative Proportion of Total Contribution (ni xi) ↑
[20, 30[	15	25	0.15	0.15	375	0.082	0.082
[30, 40[	20	35	0.2	0.35	700	0.154	0.236
[40, 50[	30	45	0.3	<u>0.65</u>	1,350	0.296	<u>0.532</u>
[50, 60[	20	55	0.2	0.85	1,100	0.242	0.774
[60, 70[	10	65	0.1	0.95	650	0.143	0.917
[70, 80[	5	75	0.05	1.00	375	0.082	1.00
Σ	100	/	1.00	/	4,550	1.00	/

From this table, we observe that **65% of employees hold 53% of the total salary mass.**

To determine the **concentration indicator**, we must first compute:

1. The median (Me)
2. The medial (ML)
3. The median-medial gap (ΔM)
4. The concentration indicator (IC)

1. Determining the Median (Me)

The **median class** is [40, 50[DA. Using **linear interpolation**:

$$Me = 40 + 10 \cdot \frac{50 - 35}{30} = \mathbf{45 \text{ DA}}$$

2. Determining the Medial (ML)

The **medial class** is [40, 50[DA. Using **linear interpolation**:

$$ML = 40 + 10 \cdot \frac{0,5 - 0,236}{0,296} = \mathbf{48,9189 \text{ DA}}$$

3. Determining the Median-Medial Gap (ΔM)

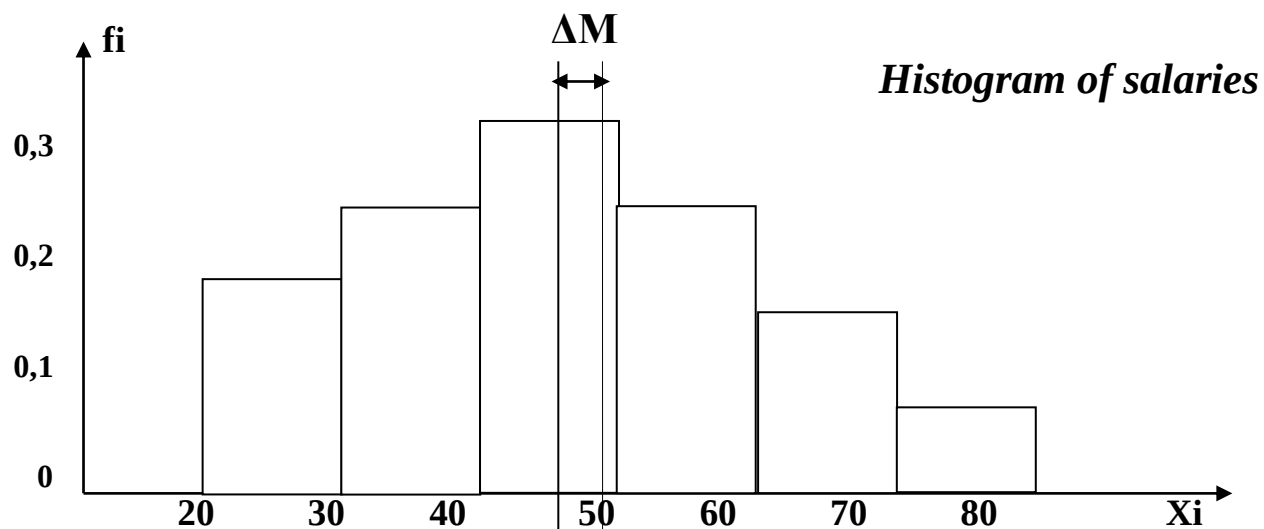
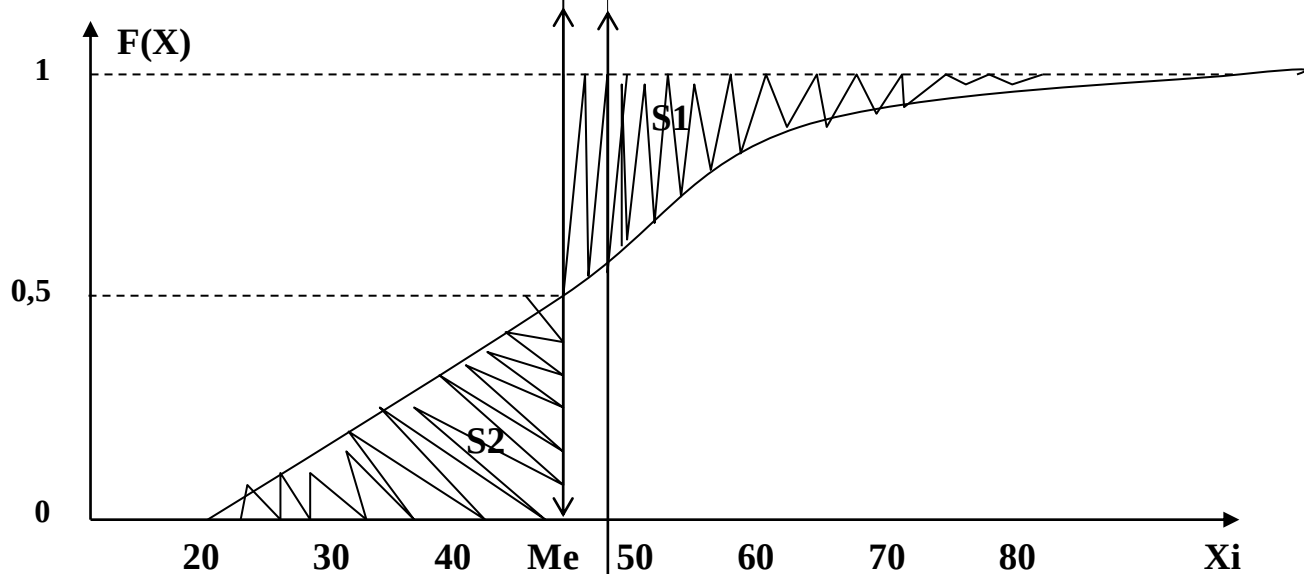
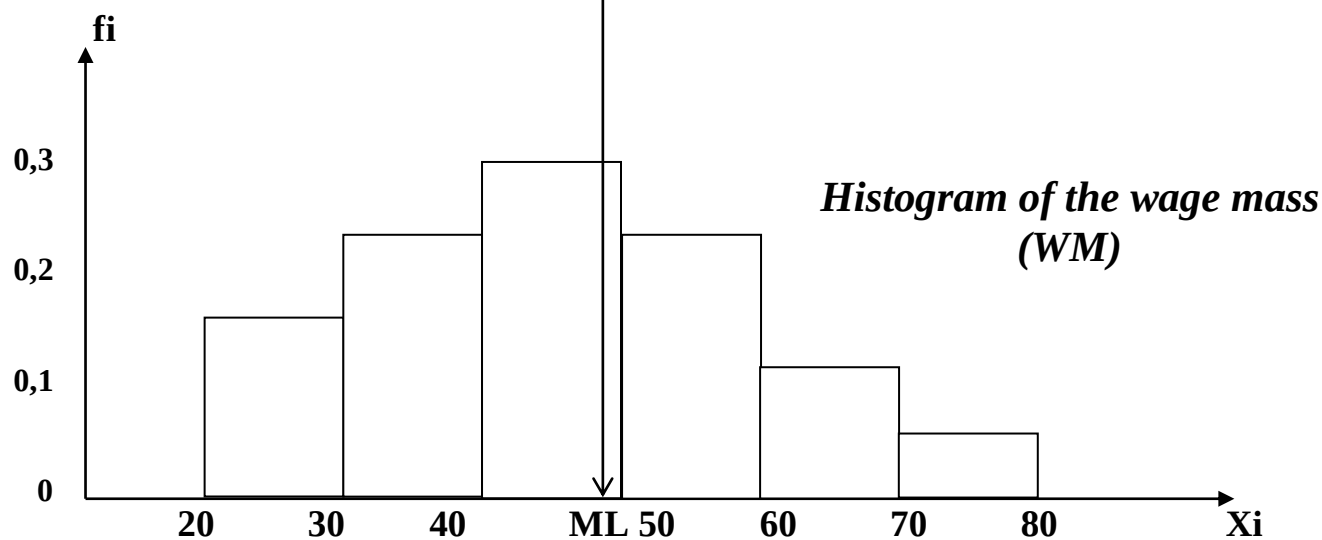
$$\Delta M = |ML - Me| = |48,9189 - 45| = \mathbf{3,92}$$

- If ΔM is large relative to the variable's range, concentration is high.
- If ΔM is small, concentration is low.
- If $\Delta M=0$, there is perfect equality in the distribution.

4. Determining the Concentration Indicator (IC)

$$IC = \frac{\Delta M}{E} = \frac{3,92}{60} = \mathbf{6,5 \text{ \%}}$$

This **indicates particularly low concentration** in the salary distribution.



The Construction of the Two Histograms

Building the **two histograms** allows us to **visualize the importance of concentration**.

It appears that this concentration is **low**, as indicated by the calculation (**small ΔM**).

VII.4 The Concentration Curve (Visual Method)

I- Principle

The general idea behind constructing a **concentration curve** is to **compare the observed distribution of the characteristic's total mass** with a **theoretical "equal" distribution** that represents **zero concentration**.

As in the calculation of the **medial**, the **mass of the characteristic** refers to the **products (n_ix_i)**.

A **completely equal distribution** of the characteristic's mass would be one where **yyy% of the population always represents y % of the total characteristic mass**.

To **assess concentration**, we compare:

1. **Cumulative frequencies of individuals (n_i)**
2. **Cumulative frequencies of the characteristic's mass (n_ix_i)**

The **greater the gap** between the cumulative frequencies of the characteristic's mass and the cumulative frequencies of individuals, **the more concentrated the distribution**.

Cumulative Frequency Formulas

1. **Cumulative Frequency of Individuals:**
2.
$$Fi(x) = \sum fi(x) = \sum_l^i \frac{ni}{N}$$
3. **Cumulative Frequency of the Characteristic's Mass:**
4.
$$Fi(x) = \sum_l^i \frac{(nixi)}{\sum_i(nixi)}$$

Graphical Representation

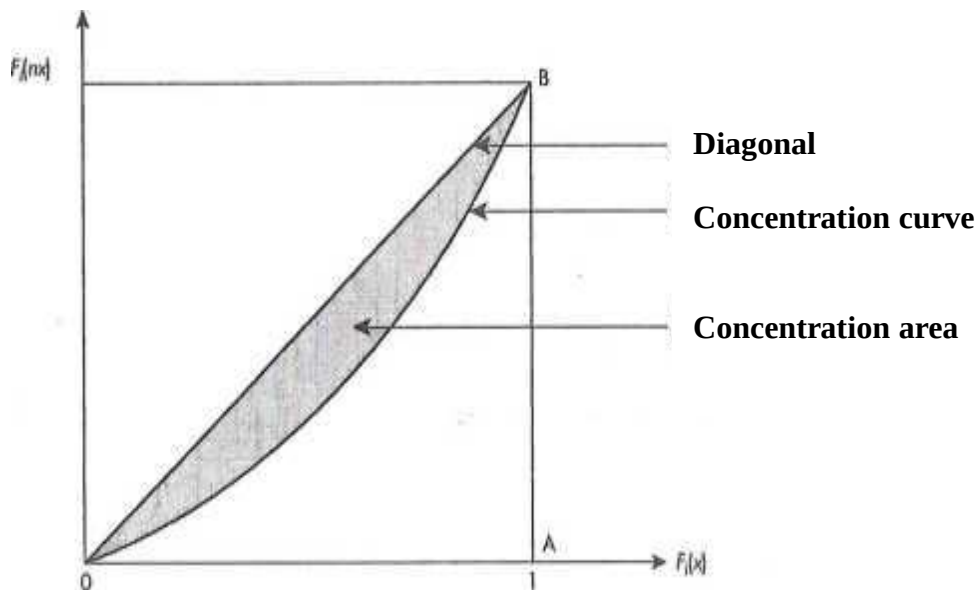
To assess the gap between **the cumulative frequencies of individuals** and **the cumulative frequencies of the characteristic's mass**, we create a **graph** where:

- **X-axis (abscissa)** represents **cumulative frequencies of individuals (F_i(x))**.
- **Y-axis (ordinate)** represents **cumulative frequencies of the characteristic's mass (F_i(nx))**.

By **joining the points (F_i(x);F_i(nx))** corresponding to each characteristic modality, we obtain the **concentration curve** (also known as the **Lorenz Curve**, **Gini Curve**, or **Gini-Lorenz Curve**).

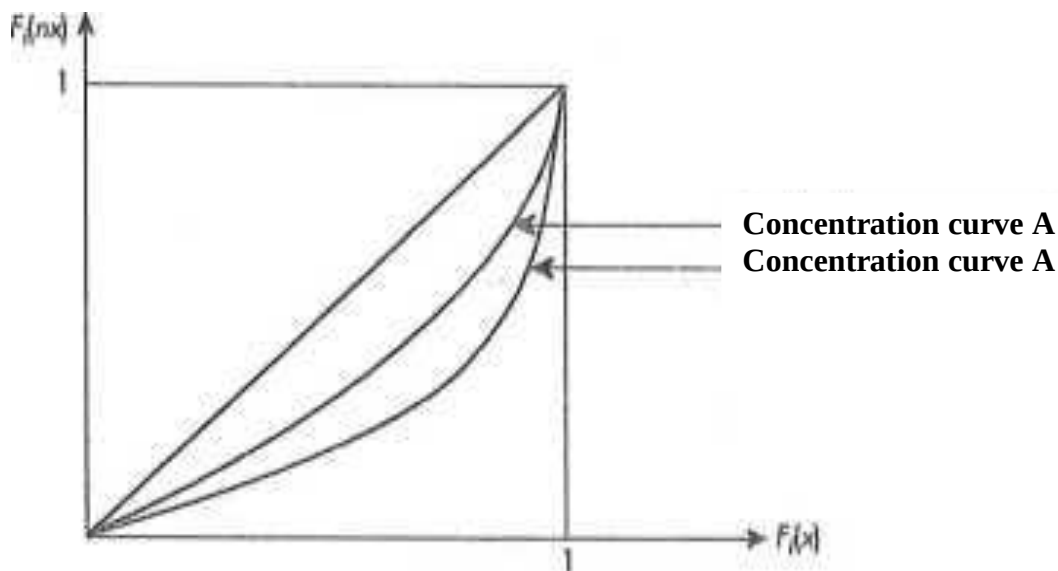
Interpreting the Lorenz Curve

- The **diagonal line** represents the **zero-concentration distribution**, where the **cumulative frequencies of individuals match the cumulative frequencies of the characteristic's mass** (perfect equality).
- By comparing the **Lorenz Curve of the observed distribution** to the **diagonal line**, we can **evaluate the degree of concentration**:
 - The farther the concentration curve deviates from the diagonal, the stronger the concentration.
 - If the curve closely follows the diagonal, the concentration is weak (or zero).



Conclusion :

- The concentration of Distribution B is stronger than that of Distribution A.
- This leads to the idea of **quantifying concentration** by **measuring the area between the diagonal and the concentration curve**.
- This **surface area** is called the **concentration area**, visually represented as the shaded region in the graph below.



II- Example of Constructing a Concentration Curve

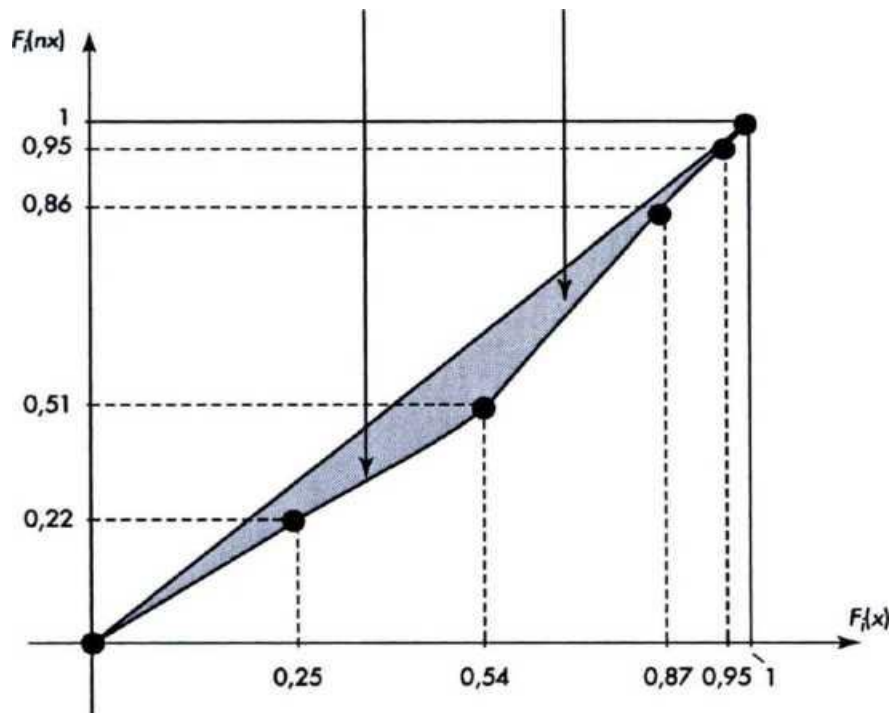
Let's consider the **distribution of 24 students by height** (see Sheet 17, "The Medial and the Medial-Median Interval" for details on intermediate calculations).

The **cumulative frequencies of individuals $F_i(x)$** and **the cumulative frequencies of the characteristic's total mass ($F_i(x)$) and ($F_i(nx)$)** are given as follows:

Height (m) x_i	Cumulative Frequency of Individuals $F_i(x)$	Cumulative Frequency of Characteristic's Mass $F_i(nx)$
< 1.60	0.25	0.2296
[1.60 – 1.70[	0.5416	0.5148
[1.70 – 1.80[	0.876	0.8604
[1.80 – 1.90[	0.9583	0.9517
> 1.90	≈ 1	≈ 1

The **concentration curve** for this distribution is obtained by **joining the points** points ($F_i(x)$; $F_i(nx)$) corresponding to each modality of the characteristic.

Concentration Concentration
curve area

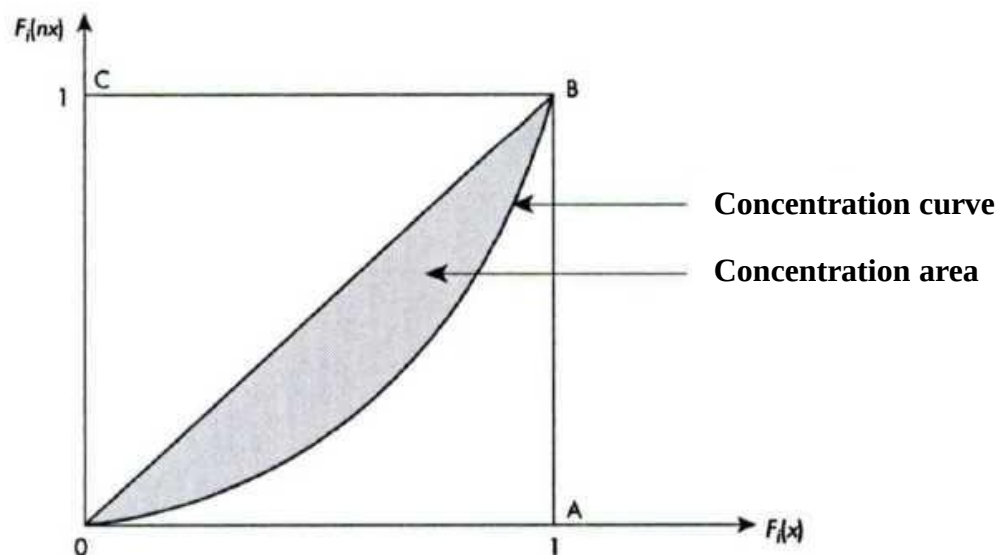


VII.5 The Gini Index (Calculation Method)

VII.5.1 Definition:

The **concentration indicator**, known as the **concentration coefficient** or **Gini index**, is defined as:

$$I = \frac{[\text{aire de concentration}]}{[\text{aire du triangle } OAB]}$$



From the **concentration curve**, this simplifies to :

$I = 2 \times \text{Concentration Area}$.

since the **area of the triangle (OAB)** is equal to **half the total area of the square (OABC)**, which is **equal to 1**.

- The **higher the Gini index**, the **stronger the concentration**.
- The **Gini index ranges from 0 to 1**.

VII.5.2 Methods for Calculating the Gini Index

The main challenge in computing the **Gini index** lies in **measuring the concentration area**. Three main methods exist for estimating this area:

A- The Graphical Method

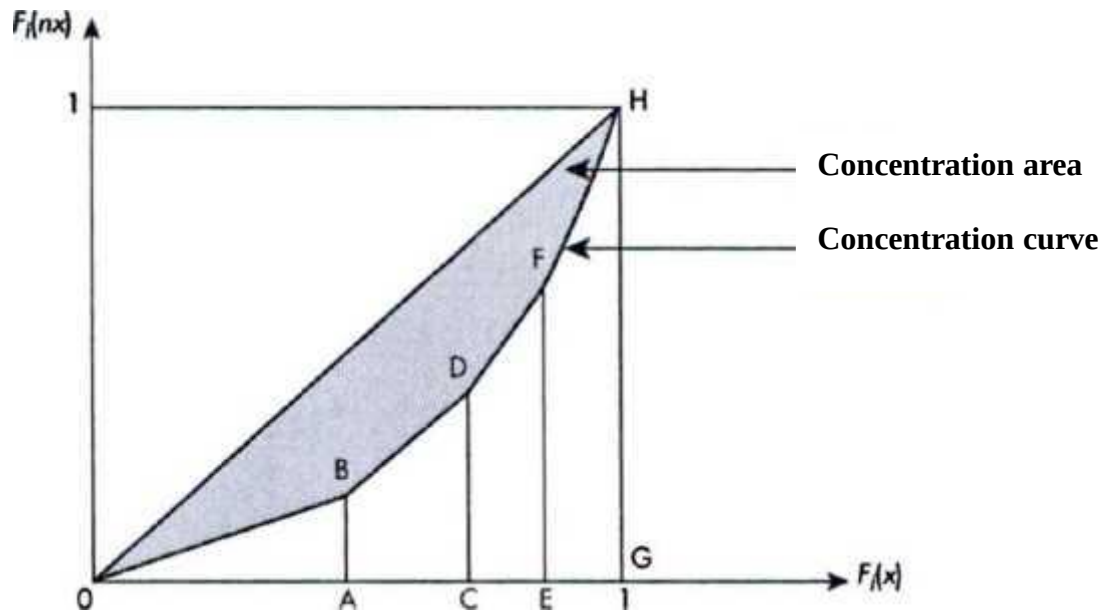
This method involves:

1. **Drawing the concentration curve** of the studied distribution on **graph paper**.
2. **Counting the number of grid squares** approximately enclosed **between the diagonal and the concentration curve**.
3. The **Gini index** is then **twice the measured concentration area**.

B- The Trapezoidal Method

In this approach, the **area between the concentration curve and the x-axis** is divided into **trapezoids**, corresponding to the **number of variable modalities**.

On This Graph, Four Trapezoids Can Be Identified :



Successively, these are: **(OAB)**, **(ACDB)**, **(CEFD)**, and **(EGHF)**.

The **concentration area** is calculated as **the area of triangle (OGH) minus the sum of the trapezoidal areas**.

Since the area of **triangle (OGH)** is equal to $\frac{1}{2}$, we have :

$$\text{aire de concentration} = \left(\frac{1}{2}\right) - \left(\sum \text{des surfaces des trapèzes}\right)$$

Thus, the **Gini Index** is given by:

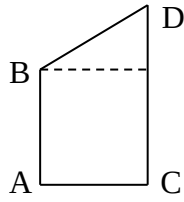
$$\text{Indice de GINI} = 1 - 2 \left(\sum \text{des surfaces des trapèzes}\right)$$

For example, the area of a **trapezoid (ACDB)** is calculated as:

$$[(AB) + (CD)] \cdot \left[\frac{(AC)}{2}\right]$$

Visually:

The area of the **trapezoid for the iii-th class** is given by:



$$S_i = [F_i(n.x) + F_{i-1}(n.x)] \cdot \left[\frac{F_i(x) - F_{i-1}(x)}{2} \right]$$

Summing for all trapezoids:

$$\sum_i S_i = \frac{1}{2} \sum_i [F_i(n.x) + F_{i-1}(n.x)] \cdot [F_i(x) - F_{i-1}(x)]$$

Thus, the **Concentration Area** is:

$$\text{Concentration Area} = \frac{1}{2} \cdot (\sum \text{des surfaces de trapèzes})$$

Thus, the **Gini Index** is:

$$I = 2 \times \text{Concentration Area}$$

$$= 1 - \sum_i [F_i(n.x) + F_{i-1}(n.x)] \cdot [F_i(x) - F_{i-1}(x)]$$

C- The Analytical Method

The **Gini index** can also be directly calculated from the **frequencies and values of the variable** across different distribution classes. It is given by:

$$I = \left(\frac{\sum \sum |x_i - x_j| \cdot n_i \cdot n_j}{2N(N-1) \cdot \bar{x}} \right)$$

where:

$$\bar{x} = \sum \left(\frac{n_i \cdot x_i}{N} \right)$$

Example: Gini Index Calculation for a Student Height Distribution

Height (m)	Frequency (ni)	Class Center (xi)	Cumulative Frequency $F_i(x)$	Cumulative Mass Frequency $F_i(nx)$
< 1.60	6	1.55	0.25	0.2296
[1.60 – 1.70[	7	1.65	0.5416	0.5148
[1.70 – 1.80[	8	1.75	0.876	0.8604
[1.80 – 1.90[	2	1.85	0.9583	0.9517
≥ 1.90	1	1.95	≈ 1	≈ 1

$\Sigma 24$

Now, we compute the **Gini Index** using both the Trapezoidal and Analytical Methods.

C.1 - Gini index calculation using the trapezoidal method

Height (m)	$F_i(nx)$	$F_{i-1}(nx)$	$[F_i(n.x) + F_{i-1}(n.x)]$	$F_i(x)$	$F_{i-1}(x)$	$[F_i(x) - F_{i-1}(x)]$	$[F_i(n.x) + F_{i-1}(n.x)] \cdot [F_i(x) - F_{i-1}(x)]$
< 1.60	0.2296	0	0.2296	0.25	0	0.25	0.0574
[1.60 – 1.70[	0.5148	0.2296	0.7444	0.5416	0.25	0.2916	0.2170
[1.70 – 1.80[	0.8604	0.5148	1.3752	0.876	0.5416	0.3344	0.4598
[1.80 – 1.90[	0.9517	0.8604	1.8121	0.9583	0.876	0.0823	0.1491
≥ 1.90	≈ 1	0.9717	1.9717	≈ 1	0.9583	0.0417	0.0822
Total	/	/	/	/	/	/	0.9655

Thus, the **Gini Index** is:

$$\begin{aligned}
 I &= 1 - \Sigma_i [F_i(n.x) + F_{i-1}(n.x)] \cdot [F_i(x) - F_{i-1}(x)] \\
 &= 1 - 0.9655 \\
 &= \mathbf{0.03}
 \end{aligned}$$

C.2 - Gini Index Calculation Using the Analytical Method

The calculations can be presented in the form of a large two-way table containing the $[|x_i - x_j| \cdot n_i \cdot n_j]$, where the sum across rows and then columns gives the numerator of the Gini index calculation formula :

	x_j	1,55	1,65	1,75	1,85	1,95	Σ
x_i	n_j	6	7	8	2	1	24
n_i		6	7	8	2	1	24
1,55	6	0	4,2	9,6	3,6	2,4	19,8
1,65	7	4,2	0	5,6	2,8	2,1	14,7
1,75	8	9,6	5,6	0	1,6	1,6	18,4
1,85	2	3,6	2,8	1,6	0	0,2	8,2
1,95	1	2,4	2,1	1,6	0,2	0	6,3
Σ	24	19,8	14,7	18,4	8,2	6,3	67,4

The x_i (and x_j) are the class centers of the distribution. The table includes the following:

$[|x_i - x_j| \cdot n_i \cdot n_j]$ such as, for example:

$$5,6 = |1,75 - 1,65| \cdot 8 \cdot 7$$

Thus, after summation across rows and columns (or vice versa) :

$$\Sigma_i \Sigma_j |x_i - x_j| \cdot n_i \cdot n_j = 67,4 \text{ m}$$

On the other hand, we have :

$$N = 24$$

$$X = \left(\frac{\Sigma n_i \cdot x_i}{N} \right) = \frac{40,5}{24} = 1,6875 \text{ m}$$

Hence, the Gini index :

$$I = \left(\frac{\Sigma \Sigma |x_i - x_j| \cdot n_i \cdot n_j}{2N(N-1) \cdot \bar{x}} \right)$$

$$I = \left(\frac{67,4}{2 \cdot 24 \cdot 23 \cdot 1,6875} \right) = \left(\frac{67,4}{1863} \right) = 0,03$$

Both methods give, as expected, the same value for the Gini index, with rounding of intermediate calculations taken into account.

8

CHAPTER

Regression and correlation

In this chapter :

- Simple regression

- The simple linear regression model
- The least squares method

- Linear correlation

- Covariance
- Correlation coefficient
- Coefficient of determination (R^2)

VIII.1 Introduction :

In the previous chapters, we studied statistical distributions with a single characteristic (we presented methods to summarize and represent information related to a single variable). However, a statistical population can be described using one, two, or multiple characteristics (or variables). In this chapter, we will focus on distributions involving two characteristics or two variables (x,y).

For instance, we will study a population of workers based on two characteristics: age (x) and salary (y), or salary (x) and the number of children (y), or age (x) and qualification (y). Similarly, we will analyze salaries in relation to seniority and education level.

This chapter is dedicated to studying the fundamental tools that allow us to examine the relationship between two characteristics and the degree of intensity of this relationship, where one characteristic is independent and the other is dependent.

VIII.2 Simple Regression:

Regression is one of the most well-known and widely used methods in statistics for analyzing quantitative data. It is used to establish a relationship between a quantitative variable and one or more other quantitative variables in the form of a model. If we are interested in the relationship between two variables, we refer to it as **simple regression**, where one variable is expressed as a function of the other. If the relationship involves one variable and multiple others, we refer to it as **multiple regression**. Implementing a regression requires the existence of a cause-and-effect relationship between the variables considered in the model.

It is common to question the relationship between two quantities, particularly in forecasting and estimation problems. Three types of problems may arise:

1. We have a set of experimental points (x_i, y_i) , where x_i and y_i are the values taken by the variables x and y , and we attempt to determine a functional relationship between these two quantities. This relationship, for theoretical or practical reasons, can be written as $y = f(x, a, b, c, \dots)$. The problem consists of optimally adjusting the parameters $a, b, c, \dots$ so that the representative curve of f closely fits the points (x_i, y_i) . This is known as an **analytical fitting problem**.
2. We attempt to determine the **statistical relationship** between the two quantities X and Y . This type of analysis is called **regression analysis**. It assumes that the variation of one of the two variables (X) explains the variation of the other (Y). Different fields of

application have given different names to these two variables. Some of these names include :

X	Y
Explanatory variable	Explained variable
Controlled variable	Response
Independent variable	Dependent variable
To regress	... (Predicted variable / Outcome variable)

In this type of analysis, the values of X are predetermined. Therefore, X is not a random variable. However, the second variable, YYY, is a random variable, and its distribution is influenced by the value of X. From a statistical perspective, this establishes a cause-and-effect relationship. The objective is to identify this relationship.

- Both variables, X and Y, are random, and we seek to determine whether their variations are linked. In this case, there is no explanatory variable or explained variable. Instead, the variables may have common causes of variation, among other factors, that explain their relationship from a statistical standpoint. This constitutes a **correlation problem**. The goal is to measure the degree of association between the variables.

Example: The relationship between an individual's weight and height, or students' scores in two different exams.

Simple Linear Regression:

It involves studying the relationship between two quantitative variables by analyzing how one varies as a function of the other.

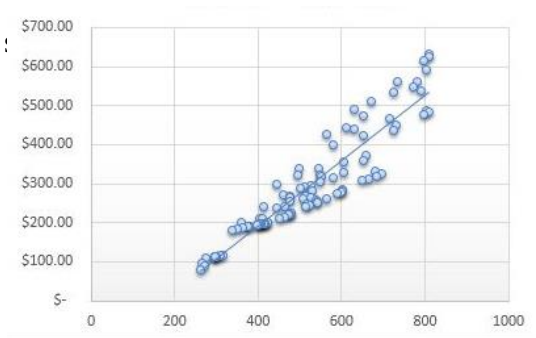
To investigate this relationship, we must define the dependent variable and the independent variable, i.e., the one that has the most influence. To examine the link between two quantitative (discrete) variables, we first create a scatter plot. Usually, we use the Cartesian representation of the pairs (x_i, y_i) with a set of nnn points, known as a **scatter diagram**, where each point corresponds to a row in the dataset, similar to the table shown below. The general shape of this graph indicates whether or not a relationship exists between the two variables.

To determine the type of relationship between the two characteristics, we encounter several **situations**:

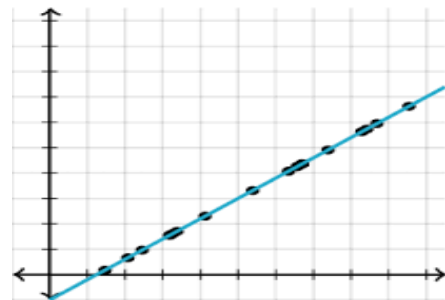
- ☐ **Positive Linear Relationship**
- ☐ **Perfect Linear Relationship**

☐ **No Relationship (or Null Relationship)**

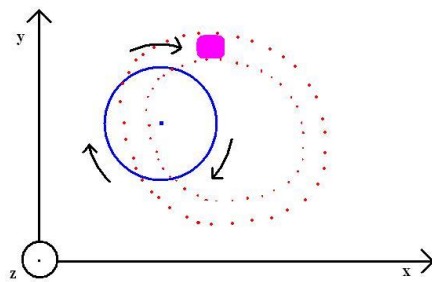
☐ **Negative Linear Relationship**



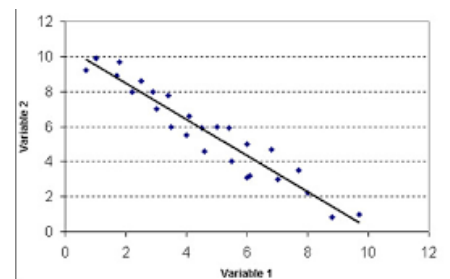
Positive Linear Relationship



Perfect Linear Relationship



No Relationship (or Null Relationship)



Negative Linear Relationship

In these diagrams, each individual in the study is represented by a point whose coordinates (x,y) correspond to the values of the two variables. The line that best fits most of these points is called the **regression line**.

Note: In this chapter, only first-degree (i.e., linear) relationships are considered.

We must formulate the **mathematical equation** that allows us to determine the linear regression line. This line is of the form:

$$y_i = ax_i + b$$

This equation enables us to determine the value of the dependent variable (y) based on the value of the independent variable (x).

The idea is to transform a scatter plot into a straight line that is as close as possible to each point. Therefore, we aim to minimize the deviations between the points and the line.

To achieve this, we use the **least squares method**. This method seeks to explain a scatter plot by a line linking Y to X, that is:

$$Y=aX+b$$

Such that the distance between the scatter plot and the line is minimized. This distance represents the **error**, i.e., the difference between the observed value and the predicted value on the regression line. If the line passes through the center of the points, the error will alternatively be positive and negative, with the sum of errors being zero by definition.

Thus, the **least squares method** consists of finding the values of the parameters *aaa* and *bbb* that **minimize the sum of squared errors**.

$$D : y/\kappa$$

is the regression line, expressed as:

$$\hat{Y}_i = a x_i + b$$

where:

- **a** represents the **slope** of the regression line (or the regression coefficient),
- **b** represents a **constant**.

VIII.2.1 The Simple Linear Regression Model:

Consider a sample of *nnn* individuals. For an individual *iii* ($i=1,...,n_i$), we have observed:

- x_i , the value of the quantitative variable *x* (explanatory or independent variable).
- y_i , the value of the quantitative variable *y* (response or dependent variable).

We want to study the relationship between these two variables, particularly the effect of *X* (independent variable) on *Y* (dependent variable).

First, we can represent this relationship graphically by plotting a **scatter plot** of the *nnn* points with coordinates (x_i, y_i). If the scatter plot has a **linear shape**, we aim to fit a line to these points.

The relationship between y_i and x_i is then written in the form of a **simple linear regression model** (the **least squares regression line**):

$$Y' = ax + b$$

where:

- **a** is called the **slope**. It represents the change in *yyy* when *xxx* changes by one unit.

- **b** is called the **intercept** (constant). It represents the predicted value of y when x=0.

But which method should we use to determine the **best** values for the parameters aaa and bbb in the model?

VIII.2.2 The Least Squares Method:

To determine the values of aaa and bbb, we use the **ordinary least squares (OLS) method**.

- $\hat{Y}_i$ is the predicted value of y on the regression line.
- b is the intercept (the value of $\hat{Y}_i$ when x=0).

The **OLS method** aims to find the regression line that minimizes the sum of the squared deviations between the actual points in the scatter plot and the points on the regression line.

Let l_i represent the deviation:

$$\Rightarrow \sum l_i^2 = \sum (\gamma_i - \hat{\gamma}_i)^2 = \text{Min}$$

The goal is to minimize the sum of squared deviations:

$$\Rightarrow \sum l_i^2 = \sum (\gamma_i - \hat{\gamma}_i)^2 = \text{Min}$$

The least squares method thus consists of minimizing the function U (the sum of the squared errors). The minimization condition is given by :

$$\frac{\partial \sum l_i^2}{\partial a} = 0 \quad \text{et} \quad \frac{\partial \sum l_i^2}{\partial b} = 0$$

For the first derivative, we obtain:

$$\frac{\partial \sum l_i^2}{\partial a} = \sum (2ax_i^2 + 2bx_i - 2x_i\gamma_i) = 0$$

This simplifies to:

$$2a \sum x_i^2 + 2b \sum x_i - 2 \sum x_i \gamma_i = 0$$

Hence :

$$2a \sum x_i^2 + 2b \sum x_i = 2 \sum x_i \gamma_i$$

This can be rewritten as follows:

$$a \sum x_i^2 + b \sum x_i = \sum x_i \gamma_i \quad (\text{équation 1})$$

Now, let's move on to the derivative with respect to b:

$$\frac{\partial \sum l_i^2}{\partial b} = \sum (2b + 2ax_i - 2\gamma_i) = 0$$

This gives:

$$n2b + 2a \sum x_i - 2 \sum \gamma_i = 0$$

Note: The fitting line passes through the mean point $(\bar{\kappa}, \bar{\gamma})$.

In the case where the variable y is independent and the variable κ is dependent, we obtain the regression line of κ on Y :

$$\kappa_i = a' \gamma_i + b'$$

where

$$a' = \frac{\text{COV}(\kappa, \gamma)}{S_\gamma^2} = \frac{(\sum \kappa_i \gamma_i) / n - \bar{\kappa} \bar{\gamma}}{(\sum \gamma_i^2) / n - \bar{\gamma}^2}$$

and

$$b' = \bar{\kappa} - a' \bar{\gamma}$$

Example 1: Given the following table. Find the regression lines Y/κ and κ/Y , estimate the consumption value when the income is 20 UM, and estimate the income when the consumption is 35 UM.

κ_i (income)	γ_i (consumption)	κ_i^2	γ_i^2	$\kappa_i \gamma_i$
5	4	25	16	20
6	5	36	25	30
7	6	49	36	42
9	7	81	49	63
12	8	144	64	96
15	12	225	144	180
54	42	560	334	431

Regression line Y/κ

$$Y_i = a\kappa_i + b$$

With:

$$a = \frac{\text{COV}(\kappa, \gamma)}{S_{\gamma}^2}$$

Parameter calculations:

$$\bar{\kappa} = \frac{\sum \kappa_i}{N} = \frac{54}{6} = 9$$

$$\bar{\gamma} = \frac{\sum \gamma_i}{N} = \frac{42}{6} = 7$$

$$\text{Cov}(\kappa, \gamma) = \frac{431}{6} - 9^2 \cdot 7 = 8,833$$

$$S_{\kappa}^2 = \frac{560}{6} - 9^2 = 12,333$$

Thus :

$$a = \frac{8,833}{12,333} = 0,716$$

and:

$$b = \bar{\gamma} - a\bar{\kappa} = 7 - 0,716 \times 9 = 0,556$$

The regression line is therefore :

$$\gamma_i = 0.716\kappa_i + 0.556$$

Estimation for $\kappa=20$:

$$\gamma = 0.716 \times 20 + 0.556 = 14.876$$

Regression line κ/γ :

$$\kappa_i = a'\gamma_i + b'$$

With:

$$a' = \frac{COV(\kappa, \gamma)}{S_{\gamma^2}}$$

Parameter calculations:

$$Cov(k, y) = \frac{431}{6} - (9 * 7) = 08,833$$

$$S^2_y = \frac{334}{6} - 7^2 = 06,666$$

Thus:

$$a' = \frac{8,833}{06,666} = 1,325$$

and :

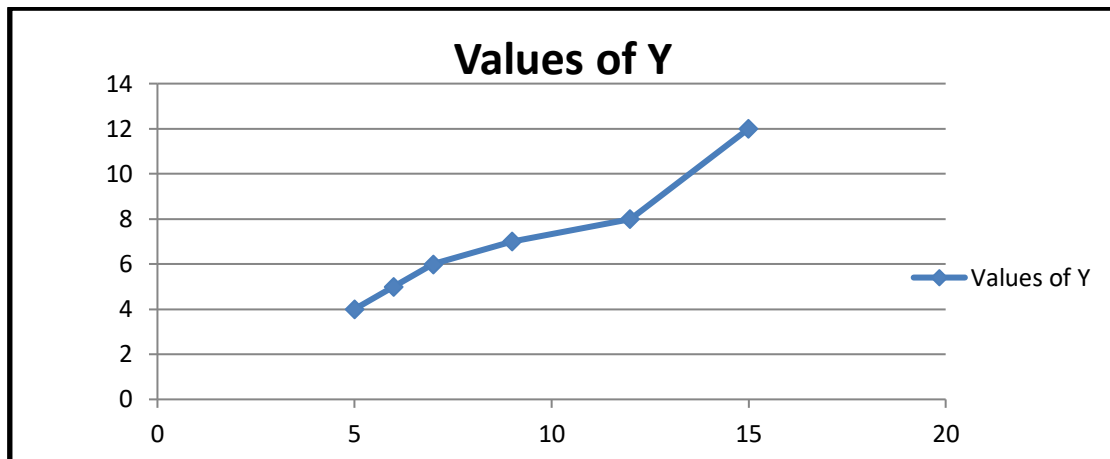
$$b' = \kappa^- - a'\gamma^- = 9 - 1.325 \times 7 = -0.275$$

The regression line is therefore:

$$\kappa_i = 1.325\gamma_i - 0.275$$

Estimation for $\gamma=35$:

$$\kappa = 1.325 \times 35 - 0.275 = 46.1$$



VIII.3 Linear Correlation

In the first section, we used the **least squares method** to determine the constants of the regression equation and to analyze the nature of the relationship between the two variables. In this section, we will **measure the strength of this relationship** using the **correlation coefficient** denoted as **r**, which informs us about the degree of dependence between the variables.

$$\Gamma = \sqrt{a \cdot a'} \Leftrightarrow \begin{cases} \gamma_i = a\kappa_i + b \\ \kappa_i = a'\gamma_i + b' \end{cases}$$

$$\Gamma = \sqrt{\frac{\text{COV}(\kappa, \gamma)}{S_{\kappa}^2} \cdot \frac{\text{COV}(\kappa, \gamma)}{S_{\gamma}^2}} \Rightarrow \Gamma = \frac{\text{COV}(\kappa, \gamma)}{S_{\kappa} \cdot S_{\gamma}}$$

VIII.3.1 What is Covariance?

Correlation quantifies the linear relationship between continuous variables. The **Pearson correlation coefficient** is computed based on the **covariance** between two continuous variables. In fact, the correlation coefficient is simply the **standardized covariance**, which allows us to obtain a value that always ranges between **-1 and +1**, regardless of the scale of measurement of the variables.

Covariance is a measure of the **association** or **relationship** between two variables. To understand covariance, let's revisit the concept of **variance**.

- **Variance** quantifies the **average dispersion** of a variable's values around its mean.

- **Two variables exhibit covariance** when a deviation from the mean of one variable is accompanied by a deviation in the same or opposite direction for the other variable.
- The more this pattern is observed across the dataset, the stronger the **association** between the two variables.

In other words, two variables **covary** when the variation of one variable around its mean **influences** the way the other variable varies around its mean. Covariance, therefore, measures the **shared variance** between two variables.

Just like variance, covariance can be **quantified**:

- The **higher the covariance**, the more the two variables share a significant portion of their variance.

The formula to compute the **covariance** between two continuous variables is:

$$Cov(X,Y) = \frac{\sum[(x_i - \bar{X})(y_i - \bar{Y})]}{N} \quad Cov(X,Y) = \frac{\sum(x_i y_i)}{N} - (\bar{X}\bar{Y})$$

Note:

The **correlation coefficient** is always within the interval [-1, 1].

- **If $r > 0$ (Positive Correlation):**

Any increase in xxx leads to an increase in yyy.

- If $r=1$: The relationship is **perfectly positive**.
- If $r=0.5$: The relationship is **moderately positive**.
- If $r \in]0; +0.5[$: The relationship is **weakly positive**.
- If $r \in]0.5; +1[$: The relationship is **relatively strong and positive**.

- **If $r < 0$ (Negative Correlation):**

Any increase in x leads to a **decrease** in y.

- If $r=-1$: The relationship is **perfectly negative**.
- If $r=-0.5$: The relationship is **moderately negative**.
- If $r \in]-0.5; 0[$: The relationship is **weakly negative**.
- If $r \in]-1; -0.5[$: The relationship is **relatively strong and negative**.

- If $r=0$:

There is **no relationship** between the two variables; they are **completely independent**.

VIII.3.2 The Correlation Coefficient $r(x,y)$:

The **correlation coefficient** measures the strength of the linear relationship between two variables X and Y. This coefficient, denoted as r , is given by :

$$r(x,y) = \frac{COV(x,y)}{\sigma_x \sigma_y}$$

or

$$r^2(x,y) = \frac{COV^2(x,y)}{V(x) V(y)}$$

- $r > 0$ if the correlation is **positive**.
- $r < 0$ if the correlation is **negative**.
- $r^2 = a \cdot a'$, where a and a' are the slopes of the regression equations $Y = aX + b$ and $X = a'Y + b'$.
- r is always **between -1 and 1**.

Example:

A company wants to determine whether there is a **correlation between its advertising expenses** and the **number of units sold** of a product. The data collected over several years are as follows:

Years	01	02	03	04	05	06	07
Units Sold (X)	987	1,578	1,956	2,540	3,091	3,426	3,980
Advertising Expenses (Y)	100	140	205	230	290	354	395

Step 1: Compute Totals and Means

We first calculate $\sum x_i$ and $\sum y_i$, then derive the **mean values** $\bar{X}$ and $\bar{Y}$.

And $\sum (x_i - \bar{X})^2$, $\sum (y_i - \bar{Y})^2$, $\sum (x_i - \bar{X})(y_i - \bar{Y})$

Step 2: Compute variance and covariance

To compute $r(x,y)$, we use a calculation table:

Years	x_i	y_i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$	$(x_i - \bar{X})(y_i - \bar{Y})$
01	987	100	-1,521.3	-144.9	2,314,310	20,984	220,369.1
02	1,578	140	-930.3	-104.9	865,432	10,995	97,547.1
03	1,956	205	-552.3	-39.9	305,020	1,589	22,012.5
04	2,540	230	31.7	-14.9	1,006	221	-471.2
05	3,091	290	582.7	45.1	339,556	2,038	26,305.4
06	3,426	354	917.7	109.1	842,200	11,912	100,162.0
07	3,980	395	1,471.7	150.1	2,165,943	22,543	220,967.4
Total	17,558	1,714	0	0	6,833,465	70,280.86	686,892.3

Step 3: Compute the Correlation Coefficient

$$r(x,y) = \frac{686\,892,3}{\sqrt{6\,833\,465 \cdot 70\,280,86}} = 0,99.$$

Conclusion:

The correlation coefficient $r=0.99$ indicates a **very strong positive correlation** between advertising expenses and the number of units sold.

VIII.3.3 The coefficient R^2 (coefficient of determination) :

The R^2 coefficient is defined as the square of the correlation coefficient between x and y , and it measures the quality of the fit. The R^2 (r^2 %) represents the percentage of the total variation in Y that is explained by the relationship between Y and X . It is the coefficient of determination.

Example 2: The following table represents the annual revenue (CA) and profits of 5 companies, in millions of dinars. Calculate the correlation coefficient and interpret the result. Find the regression lines κ/Y and Y/κ .

κ_i (Revenue)	γ_i (Profit)	$\kappa_i \gamma_i$	γ_i^2	κ_i^2
10	1	10	1	100
20	3	60	9	400
30	2	60	4	900
40	5	200	25	1600
50	4	200	16	2500
Totals		530	55	5500

Calculations for κ and γ :

$$\bar{\kappa} = \frac{\sum \kappa_i}{N} = \frac{150}{5} = 30$$

$$\bar{\gamma} = \frac{\sum \gamma_i}{N} = \frac{15}{3} = 03$$

The correlation coefficient $\Gamma(\kappa, \gamma)$ is given by :

$$r(\kappa, \gamma) = \frac{COV(\kappa, \gamma)}{\sigma_{\kappa} \sigma_{\gamma}}$$

Where the covariance is calculated as follows :

$$Cov(\kappa, \gamma) = \frac{530}{5} - 30 * 3 = 16$$

The standard deviations S_{κ} and S_{γ} are calculated as follows :

$$S_{\kappa} = \sqrt{\frac{\sum \kappa_i^2}{n} - \bar{\kappa}^2} = \sqrt{\frac{5500}{5} - 30^2} = \sqrt{200} \Rightarrow S_{\kappa} = 14.142$$

$$S_{\gamma} = \sqrt{\frac{\sum \gamma_i^2}{n} - \bar{\gamma}^2} = \sqrt{\frac{55}{3} - 3^2} = \sqrt{2} \Rightarrow S_{\gamma} = 1.414$$

The correlation coefficient is then:

$$r(\kappa, \gamma) = \frac{16}{14.142 * 1.414}$$

$$r(\kappa, \gamma) = 0.8$$

This indicates a strong positive relationship between revenue and profit.

Regression line Y/κ :

The regression line is given by:

$$Y_i = a\kappa_i + b$$

With:

$$a = \frac{\text{Cov}(k,y)}{S^2k} \Rightarrow a = \frac{16}{200} = 0,08$$

and:

$$b = \bar{y} - a\bar{k} = 3 - (0,08 \times 30) \Rightarrow b = 0.6$$

The regression line is therefore:

$$\gamma_i = 0.08\kappa_i + 0.6$$

Regression line κ/γ :

The regression line is given by:

$$\kappa_i = a'\gamma_i + b'$$

With:

$$a' = \frac{\text{Cov}(k,y)}{S^2y} \Rightarrow a' = \frac{16}{2} = 8$$

and:

$$b' = \bar{\kappa} - a'\bar{\gamma} = 30 - (8 \times 3) \Rightarrow b' = 6$$

The regression line is therefore:

$$\kappa_i = 8\gamma_i + 6$$

Example 3: The Evolution of Stock and Bond Prices

The table below presents the **evolution of stock and bond prices** over a series of years.

Objectives:

1. Determine if there is a **correlation** between stock price changes and bond price changes.
2. Find the **regression equation** of X on Y.

Year	Stocks (X)	Bonds (Y)
2012	352	1024
2013	360	998
2014	358	980
2015	361	970
2016	366	982
2017	382	972
2018	398	935
2019	406	902
2020	450	895
2021	445	900
Total	3878	9558

Solution :

We compute key statistical measures using the following extended table:

Year	Stocks (X)	Bonds (Y)	$(xi - \bar{X})$	$(xi - \bar{X})^2$	$(yi - \bar{Y})$	$(yi - \bar{Y})^2$	$(xi yi)$
2012	352	1024	-35.8	1281.64	68.2	4651.24	-2441.56
2013	360	998	-27.8	772.84	42.2	1780.84	-1173.16
2014	358	980	-29.8	888.04	24.2	585.64	-721.16
2015	361	970	-26.8	718.24	14.2	201.64	-380.56
2016	366	982	-21.8	475.24	26.2	686.44	-571.16
2017	382	972	-5.8	33.64	16.2	262.44	-93.96
2018	398	935	10.2	104.04	-20.8	432.64	-212.16
2019	406	902	18.2	331.24	-53.8	2894.44	-979.16
2020	450	895	62.2	3868.84	-60.8	3696.64	-3781.76
2021	445	900	57.2	3271.84	-55.8	3113.64	-3191.76
Total	3878	9558	/	11,745.6	/	18,305.6	-13,546.4

The mean:

$$\bar{X} = \frac{\sum xi}{N} = \frac{3878}{10} = 387.8$$

$$\bar{Y} = \frac{\sum yi}{N} = \frac{9558}{10} = 955.8$$

The variance:

$$V(X) = \frac{11745.6}{10} = 1174.56$$

$$V(Y) = \frac{18305.6}{10} = 1830.56$$

The standard deviation:

$$\sigma_X = \sqrt{V(X)} = \sqrt{1174.56} = 34.27$$

$$\sigma_Y = \sqrt{V(Y)} = \sqrt{1830.56} = 42.79$$

The covariance:

$$\text{Cov}(x, y) = \frac{-13546,4}{10} = -1354.64$$

The correlation coefficient:

$$r(x, y) = \frac{1505,16}{34,27 \times 42,79} = -0.92 \Rightarrow \text{Strong negative correlation.}$$

The coefficient R^2 :

$$r^2 = (-0.92)^2 = \mathbf{0.8464}$$

$$R^2 = (-0.92)^2 \times 100 = \mathbf{85.35\%}$$

This means that 85.35% of the variation in bonds (Y) is explained by the relationship between stocks (X) and bonds.

The regression line of Y on X :

It has the form:

$$Y' = aX + b$$

With:

$$a = \frac{\text{Cov}(k, y)}{V(x)} \Rightarrow a = \frac{-1354.64}{1174,56} = \mathbf{-1.15}$$

$$b = Y^- - aX^- \Rightarrow b = 955.8 - (1.15) \times 387.8 = \mathbf{1403.06}$$

Thus, the regression line is :

$$Y' = -1.15X + 1403.06$$

9

CHAPTER

APPLICATION

Chapter 01: General concepts in statistics

Chapter 02: Data representation : Statistical tables

Exercise 01:

The following ordered statistical series represents the **weight in kilograms (kg)** of a group of individuals:

50 - 55 - 61 - 62 - 64 - 65 - 66 - 67 - 68 - 69 - 70 - 72 - 73 - 74 - 75 - 76 - 76 - 77 - 78 - 79 - 80
- 81 - 81 - 82 - 83 - 83 - 84 - 85 - 85 - 85 - 86 - 87 - 87 - 88 - 90 - 92 - 92 - 93 - 93 - 95 - 95 -
96 - 97 - 98 - 100 - 102 - 102 - 104 - 104 - 105 - 107 - 109 - 111 - 118

1. Identify the population and the studied characteristic.
2. Organize these data into a frequency distribution table.
3. Calculate the central tendency parameters. What conclusions can be drawn?

Exercise 02:

The following table shows the distribution of a group of **15 students** based on their **daily expenses**:

Daily Expenses (DA)	0 – 10	10 - 50	50 - 100	100 and more
Number of Students	2	6	6	1

1. Determine the population, statistical unit (individuals), and the nature of the characteristic.
2. Calculate the percentage of students in each daily expense category.
3. Find the percentage of students spending more than 100 monetary units per day.
4. Find the percentage of students spending less than 10 monetary units per day.
5. Find the percentage of students spending less than 100 monetary units per day.
6. Discuss the importance of empirical methods such as Yule's method and Sturges' method, and explain how they are determined.
7. Determine the mode of this distribution. Interpret the results by explaining the advantages and disadvantages of the mode.

Exercise 03:

The following table shows the **distribution of 200 employees** based on their **salaries**:

Salary Range	0-40	40-80	80-120	120-L	L-220	220-300	300-420
Employees	12	N2	N3	34	28	22	6

1. **Identify the population, statistical unit (individuals), and the nature of the characteristic.**
 - Find N2 and N3 given that the **fourth decile is equal to 95**.
2. **Determine the limit LLL given that the arithmetic mean is 130.**
3. **Calculate: $\sum (2i+1)$ and express the following sum using the summation symbol $\sum$: $5+7+9+\dots +17+19$**
4. **Discuss the importance of empirical methods such as Yule's method and Sturges' method, and explain how they are determined.**
5. **The monthly salaries in a small company of 5 employees are as follows: 2500,3200,3800, 4500,8700.**
 - **Calculate the median and arithmetic mean.**
 - **If the last salary is changed to 22,500, what will be the new median and arithmetic mean?**
 - **Compare the median with the mean. What conclusions can be drawn?**
 - **Discuss the advantages and disadvantages of central tendency measures (position parameters).**
 - **When can we consider the geometric mean, arithmetic mean, and quadratic mean as characteristics of central tendency?**

Exercise 04:

In a faculty of **3,000 students**, the choice of **specialization for second-year students** is distributed as follows:

- **Finance** → 40%
- **Management** → 25%

- **Accounting** → 20%
 - **Marketing** → 10%
 - **International Trade** → 5%
1. **Construct a frequency distribution table.**
 2. **What are the possible graphical representations?**
 3. **Draw a graphical representation of your choice.**

Chapter 03: Data representation - Statistical graphs

Exercise 01:

The following table represents the **distribution of workers in Algeria** across different sectors during two periods **(1970 - 1995)**:

Sector	1995 (%)	1970 (%)
Agriculture	14	6
Industry	28	17
Construction	16	11
Administration	25	62
Services	17	4
TOTAL	100	100

Questions:

1. What are the different possible graphical representations for illustrating this table?
2. Graphically represent the 1960 data in two different ways.
3. Represent the entire dataset in a single graph. What conclusions can be drawn?

Chapter 04: Measures of central tendency

Exercise 01:

In a company, employees are classified according to their **years of seniority in the organization**:

Seniority (years)	< 1	1 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
Number of Employees	22	24	52	30	40	20	12

Questions:

1. Identify the population, statistical unit (individuals), and the nature of the characteristic.
2. Calculate the percentage of employees with seniority:
 - Less than 5 years.
 - More than 20 years.
3. Determine the median using the graphical method.
4. Calculate the first quartile Q1 and third quartile Q3, then deduce the interquartile range.

Exercise 02:

To better understand its customer base, the **director of a company conducted a survey**, obtaining the following data on the **age distribution of customers**:

Age Group (years)	45+	40 - 45	35 - 40	30 - 35	25 - 30	20 - 25	15 - 20
Percentage (%)	3	5	10	15	28	26	13

Questions:

1. Draw the histogram and deduce the mode.
2. Plot the cumulative frequency curves (ascending and descending) and deduce the median.
3. Find the mode, median, and arithmetic mean by calculation.
4. Determine the percentage of customers aged ≥ 40 years.

Exercise 03:

The following table represents the distribution of **80 employees** based on their **monthly salary (in 10^3 DA)**:

Salary Range (10^3 DA)	[5-10[	[10-15[	[15-20[	[20-25[	[25-30[
Number of Employees	8	10	12	30	20

Questions:

1. Draw the histogram and determine the mode graphically.
2. Calculate the first quartile Q_1 and the third quartile Q_3 , then interpret the results.
3. What percentage of workers earn a salary between 15,000 and 25,000 DA?

Exercise 04:

The following table represents the **distribution of salaries (in 10^3 DA)** of **50 workers** in a company:

Salary Range (10^3 DA)	[100-120[	[80-100[	[60-80[	[40-60[	[20-40[
Relative Frequency	0.12	0.20	0.30	0.20	0.18

Questions:

1. Calculate the absolute frequency of each class, the cumulative increasing frequencies, and the cumulative decreasing frequencies.
2. Calculate the arithmetic mean using the definition formula and the variable transformation method.
3. Find the median and the mode, then compare the three central values and deduce the shape of the distribution.

Exercise 05:

To better understand its customer base, the **director of a company** conducted a survey and obtained the following data on **the age distribution of customers**:

Age Group (years)	45+	45 - 40	40 - 35	35 - 30	30 - 25	25 - 20	20 - 15
Percentage (%)	3	5	10	15	28	26	13

Questions:

1. Draw the histogram and deduce the mode.
2. Plot the cumulative frequency curves (ascending and descending) and deduce the median.
3. Find the mode by calculation.
4. Determine the percentage of clients aged ≥ 40 years.

Exercise 06:

The following table represents the **distribution of N employees** based on **age in a company A**:

Age Group (years)	[20-25[	[25-30[	[30-35[	[35-40[
Number of Employees	20	?	50	10

Questions:

1. Given that the median is 30.5, complete the table.
2. Calculate the quartiles Q_1, Q_2, Q_3 , the fifth decile D_5 , and the twenty-fifth percentile P_{25} .
3. Calculate the harmonic mean H , geometric mean G , arithmetic mean $\bar{X}$, and quadratic mean Q . Compare the results.
4. In another company B, the average employee age is 30 years, with 90 employees. Calculate the combined average age of employees from both companies.

Exercise 07:

A driver travels a total **distance of 900 km** in three stages at different speeds:

- **162 km at 60 km/h** (18% of the total distance)
- **528 km at 110 km/h** (58.67% of the total distance)
- **210 km at 140 km/h** (23.33% of the total distance)

Question:

- **Calculate the average speed of the trip.**

Exercise 08:

An investor purchases stocks as follows:

- **For 26,000 euros at a stock price of 520 euros per share**
- **For 40,000 euros at a stock price of 500 euros per share**
- **For 37,100 euros at a stock price of 530 euros per share**

Question:

- **What is the average purchase price per share?**

Chapter 05: Measures of dispersion

Exercise 01:

A statistical population is given as follows:

Class Intervals	[0 - 4[	[4 - 10[	[10 - 20[	[20 - 40[
Frequencies	4	20	14	2

Questions:

1. Calculate the arithmetic mean and variance.
2. Determine the coefficient of skewness and interpret the result.
3. Divide each class into two equal-width sub-classes, assigning each a frequency equal to half of the original class frequency.
 - Construct a new table.
 - Recalculate the arithmetic mean and variance.
 - Compare and interpret the results.

Exercise 02:

Calculate the **average annual growth rate** of a product's production over the period **2013-2018**, given the following **successive growth rates (%)**:

Year	2013	2014	2015	2016	2017	2018
Growth Rate (%)	9.6	7.4	5.8	2.7	3.1	2.7

Exercise 03:

The following series represents the **number of absences** of **6 students** in tutorial sessions for a subject:

4,5,6,2,1,0, 5, 6, 2, 1, 0,4,5,6,2,1,0

Questions:

- Calculate the variance (and standard deviation) using:
 - The **definition formula**.
 - The **expanded formula (Koenig's theorem)**.

Exercise 04:

The following table represents the **distribution of 50 employees** in a company according to their **monthly salary (in 10² DA)**:

Salary Range (10² DA)	[120-130[	[130-140[	[140-150[	[150-160[	[160-170[	[170-180[	[180-190[
Frequencies	4	7	10	15	7	4	3

Questions:

1. Calculate the following dispersion measures:
 - First quartile Q1 and third quartile Q3.
 - Range, interquartile range.
 - Fifth decile D5, ninth decile D9, and 90th percentile P90.
 - **Interpret the results.**
2. Calculate the variance and standard deviation using:
 - The **definition formula**.
 - The **expanded formula (Koenig's theorem)**.
3. Calculate the coefficient of skewness algebraically and graphically. Interpret the result.

Exercise 05 :

A survey recorded the **average hourly wage of 100 male employees** in a company :

- **Mean hourly wage: 24 DA**
- **Standard deviation: 1.56 DA**

A sample of **female employees** in the same company had the following **hourly wages (in DA)**:

Salary Range (DA)	20-22	22-24	24-26	26-28	28-30	30-32
Frequencies	10	20	30	25	8	7

Question:

- Compare the two distributions using the coefficient of variation. What conclusions can be drawn?

Exercise 06:

The following dataset provides information on **hourly wages** received by **50 employees** in a company during **March 2018**:

340	360	450	620	370	430	420	1020	310	420
510	300	610	630	470	1050	520	430	810	950
920	770	600	360	480	490	650	710	780	810
430	520	630	710	430	420	510	550	610	410
930	820	830	470	540	610	1020	330	480	550

Questions:

1. Identify the studied population, characteristic, its nature, and modalities.
2. Organize the observations and present the results in a statistical table with class intervals.
3. Draw the histogram and plot cumulative frequency curves (ascending and descending).
4. Determine the mode, median, and arithmetic mean. Compare them.
5. Calculate the variance and standard deviation.

Exercise 07:

The **distribution of 40 employees** in a company based on the **number of dependent children** is given below:

Number of Dependent Children (xi)	0	1	2	3	4	5	6
Number of Employees (ni)	4	7	8	8	6	4	3

Questions:

1. Graphically represent this series.
2. Determine the mode. How does it appear on the graph?
3. Determine the cumulative distribution function and represent it graphically.
4. Calculate the variance and standard deviation of this distribution.

Exercise 08:

The following data are given:

$$\sum_{i=1}^{10} x_i^2 = 1060 \qquad \sum_{i=1}^{10} (x_i - \bar{x})^2 = 250$$

Questions:

1. Calculate the arithmetic mean.
2. Determine the coefficient of variation.

Exercise 09:

The following data relate to a **symmetric salary distribution** of **50 employees**:

$$\text{Variance} = 18,6 \qquad \sum n_i \cdot x_i^2 = 1820$$

Questions:

- Calculate the median and mode of the distribution.

Exercise 10 :

A study on the **revenue** of a population of **small and medium-sized enterprises (SMEs)** provided the following results (in thousands of Algerian Dinars - DA):

Statistic	Value (in DA)
Minimum	3500
Arithmetic Mean	4900
Standard Deviation	650
Mode	4550
Interquartile Range	1100
Median	4600
First Quartile (Q1)	4100
First Decile (D1)	3700
Interdecile Range	2800
Range	5000

Questions:

1. **Classify these parameters into two categories:**
 - **Central tendency and position.**
 - **Dispersion.**
2. **What is the highest revenue in this SME population?**
3. **Calculate the third quartile Q3 and the ninth decile D9.**
4. **Place the parameters on a number line to visualize the distribution.**

Exercise 11:

The following table represents the **distribution of 600 employees** in a company based on the duration of their **home-to-work commute (in minutes)**:

Commute Duration (minutes)	[60- 70[	[50- 60[	[40- 50[	[30- 40[	[20- 30[	[10- 20[	[0- 10[
Number of Employees	40	54	83	151	127	95	50

Questions:

1. Among these three durations: 15 min, 35 min, and 60 min, which seems to be the closest to the average commute duration?
2. Calculate the mean commute duration.
3. Determine the mean absolute deviation and compare it with the standard deviation.
4. Calculate the first decile (D1) and the ninth decile (D9), then deduce an indicator of dispersion around the median (Me).

Exercise 12:

A survey recorded the **average hourly wage of 100 male employees** in a company:

- **Mean hourly wage: 24 DA**
- **Standard deviation: 1.56 DA**

A sample of **female employees** in the same company had the following **hourly wages (in DA)**:

Salary Range (DA)	[20-22[	[22-24[	[24-26[	[26-28[	[28-30[	[30-32[
Number of Employees	10	20	30	25	8	7

Question:

- **Compare the two distributions using the coefficient of variation. What conclusions can be drawn?**

Exercise 13:

A statistical population is given as follows:

Class Intervals	[2 - 6[	[6 - 12[	[12 - 22[	[22 - 42[
Frequencies	4	20	14	2

Questions :

1. Calculate the arithmetic mean and variance.
2. Determine the coefficient of skewness and interpret the result.
3. Divide each class into two equal-width sub-classes, assigning each a frequency equal to half of the original class frequency.
 - o Construct a new table.
 - o Recalculate the arithmetic mean and variance.
 - o Compare and interpret the results.

Exercise 14:

A survey recorded the **average hourly wage of 100 male employees** in a company:

- Mean hourly wage: 24 DA
- Standard deviation: 1.56 DA

A sample of **female employees** in the same company had the following **hourly wages (in DA)**:

Salary Range (DA)	[20-22[	[22-24[	[24-26[	[26-28[	[28-30[	[30-32[
Number of Employees	10	20	30	25	8	7

Question:

- Compare the two distributions using the coefficient of variation. What conclusions can be drawn?

Chapter 06: Shape parameters

Exercise 01:

We calculate the skewness indicators for the **distribution of a population of 24 students** according to their **height (in meters)** :

Height (m)	< 1.60	[1.60 - 1.70[	[1.70 - 1.80[	[1.80 - 1.90[	≥ 1.90
Frequencies (ni)	6	7	8	2	1

Questions:

1. Calculate the required quartiles.
2. Determine the quartile coefficient of skewness (Yule's coefficient).
3. Calculate Pearson's first skewness coefficient.
4. Calculate Pearson's second skewness coefficient.

Exercise 02:

We calculate the kurtosis indicators for the **distribution of a population of 24 students** according to their **height (in meters)**:

Height (m)	< 1.60	[1.60 - 1.70[	[1.70 - 1.80[	[1.80 - 1.90[	≥ 1.90
Frequencies (ni)	6	7	8	2	1

Questions:

1. Calculate Pearson's kurtosis coefficient.
2. Determine the quantile-based kurtosis coefficient.

Chapter 07: Measures of concentration

Exercise 01:

The **income distribution** (in thousands of DA) for **500 workers** in a company is presented in the following table:

Income (10 ³ DA)	[12 - 16[	[16 - 20[	[20 - 24[	[24 - 28[	[28 - 32[
Number of Workers	75	100	150	100	75

Questions:

1. Plot the Lorenz Curve.
2. Calculate the Gini Index (GI).

Exercise 02:

The **income distribution** (in thousands of DA) for employees in a company is presented in the following table:

Income (10 ³ DA)	[15 - 25[	[25 - 35[	[35 - 45[	[45 - 55[	[55 - 65[	[65 - 75[	[75 - 85[	[85 - 105[
Number of Employees	300	170	130	110	60	40	20	10

Questions:

1. Plot the Lorenz Curve.
2. Calculate the Gini Index (GI).

Exercise 03:

The following table presents the **distribution of 24 students** based on their **height (in meters)**:

Height (m)	[1.50 - 1.60[	[1.60 - 1.70[	[1.70 - 1.80[	[1.80 - 1.90[	[1.90 - 2.00[
Number of Students	6	7	8	2	1

Questions:

1. Plot the Lorenz Curve.
2. Calculate the Gini Index (GI).

Chapter 08: Regression and Correlation

Exercise 01:

The following table presents the **milk production over 10 years** in a given company:

Year (t)	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Production (10 ² L)	10	12	12	13	14	16	17	18	20	22

Questions:

1. Graphically represent the data. What do you observe?
2. Determine the equation that fits the data using the least squares method.
3. If nothing changes, what will the milk production level be in 2023?

Exercise 02:

For **11 months**, the **consumption** of two goods XXX and YYY has been observed:

Consumption of X	70	50	51	40	41	55	56	52	58	60	65
Consumption of Y	50	69	70	80	81	65	66	67	64	65	67

Questions:

1. Calculate the covariance between X and Y.
2. Determine the Pearson correlation coefficient $r(x,y)$.
3. Calculate the coefficient of determination $r^2(x,y)$.

Exercise 03:

The **linear regression equations** are given as follows:

- Regression of Y on X: $Y=X+30$
- Regression of X on Y: $X=\frac{1}{4}Y+60$

Questions:

1. Calculate the Pearson correlation coefficient.
2. Calculate the arithmetic means of variables X and Y.
3. Calculate the covariance between X and Y, and the variance of X, given that the variance of Y is 40.
4. If the correlation coefficient is zero, how can X be estimated from Y?

Bibliography :

- ALBARELLO, BOURGEOIS et GUYOT., (2007), **"Statistique descriptive : Un outil pour les praticiens-chercheurs"**, De Boeck, 174 p.
- ANDERSON, SWEENEY, WILLIAMS, CAMM et COCHRAN., (2015), **"Statistiques pour l'économie et la gestion"**, 5^{ème} édition, De Boeck, 921 p.
- BAILLARGEON, G., (2015), **"Statistique appliquée pour les sciences de la gestion et les sciences économiques"**, 3^{ème} édition, Les Éditions SMG, 772 p.
- Bijma, F., Jonker, M., & van der Vaart, A., (2017), **"Descriptive Statistics. In An Introduction to Mathematical Statistics"**, (pp. 21–44). Amsterdam University Press.
<http://www.jstor.org/stable/j.ctt1v2xsxr.6>.
- Calot. G., (1984), **"Cours de statistique descriptive"**, Dunod, Paris.
- CHESNEAU et KACHOUR., (2015), **"Intervalles de confiance et tests statistiques : moyennes, variances et proportions"**, Ellipses, 229 p.
- D'ASTOUS, A., (2019), **"Le projet de recherche en marketing"**, Chenelière Éducation, 460 p.
- EMC., (2015), **"Data science and big data analytics : Discovering, analyzing, visualizing and presenting data"**, Wiley, 410 p.
- FIELD, A., (2018), **"Discovering statistics using IBM SPSS statistics"**, 5th edition, SAGE Publication, 1070 p.
- GONICK et SMITH., (2005), **"The cartoon guide to statistics"**, Herper, 230 p.
- HAHN et MACÉ (2012), **"Méthodes statistiques appliquées au management"**, Pearson, 342 p.
- Kaur P, Stoltzfus J, Yellapu V., (2018), **"Descriptive statistics"**. Int J Acad Med 2018;4:60-3.
- Kimberly Leung, Chris Rasmussen, Samuel S. P. Shen, and Dov Zazkis., (2014), **"Calculus from a Statistics Perspective"**. The College Mathematics Journal, 45(5), 377–386.
- Klarmann, M., & Feurer, S., (2018), **"Control Variables in Marketing Research"**. Marketing : ZFP – Journal of Research and Management, 40(2), 26–40.
- Liorzou, A., (1980), **"Initiation pratique à la statistique"**, Eyrlles, Paris.

- MARR, B., (2015), "**Big Data : Using SMART big data analytics and metrics to make better decisions and improve performance**", Wiley, 246 p.
- MARR, B., (2016), "**Big Data in practice : How 45 successful companies used big data analytics to deliver extraordinary results**", Wiley, 308 p.
- NEWBOLD, Carlson et THORNE., (2013), "**Statistics for Business and Economics**", Pearson Education, 772 p.
- PLAISENT, ZHENG, KHADHRAOUI et BERNARD., (2018), "**Concepts et outils des sondages web : Introduction à LimeSurvey et SurveyMonkey**", Presses de l'Université du Québec, 149 p.
- PROVOST et FAWCETT., (2013), "**Data sciences for Business : What you need to know about data mining and data-analytic thinking**", O'Reilly Media, 386 p.
- PUPION, P.-C., (2012), "**Statistiques pour la gestion : Applications avec Excel, SPSS, AMOS et Smart PLS**", Dunod, 445 p.
- PY, B., (2010), "**La statistique sans formule mathématique : comprendre la logique et maîtriser les outils**", Pearson Education, 306 p.
- Porter, T. M., (2022-09), "**Probability and statistics**". Encyclopedia Britannica. <https://www.britannica.com/science/probability>.
- SHARDA et DELEN et Turban., (2018), "**Business Intelligence, Analytics, and Data Science : A Managerial Perspective**", Pearson, 486 p.
- Stigler, S. M., (1986), "**The history of Statistics : The measurement of uncertainty before 1900**". Harvard Univ. Press.
- Taylor, Courtney., (2020), "**Definition of a Percentile in Statistics and How to Calculate It**". Thought Co, Aug. 26, 2020, [thoughtco.com/what-is-a-percentile-3126238](https://www.thoughtco.com/what-is-a-percentile-3126238).
- Witte, R. S., & Witte, J. S., (2017), "**Statistics**". Wiley.
- WONG., (2013), "**The Wall Street Journal Guide to information graphics : The dos & don'ts of presenting data, facts, and figures**", Norton, 157 p.
- Zhang, Lei, and Yimin Yi., (2018), "**What Contributes to the Rising House Prices in Beijing ? A Decomposition Approach**". Journal of Housing Economics, vol. 41, 2018, pp. 72–84.

- جلاطو جيلالي، (2007)، "الإحصاء مع تمارين و مسائل محلولة"، ديوان المطبوعات الجامعية OPU. الجزائر.
- د. سليم لعور، (2016)، "مبادئ في الإحصاء الوصفي"، ديوان المطبوعات الجامعية OPU. الجزائر.
- فؤاد البهي السيد، (2000)، "مبادئ الإحصاء"، دار المعارف. مصر.
- عبد العزيز السنبل، (2002)، "الإحصاء الوصفي"، مكتبة الخريجي. المملكة العربية السعودية.
- زكريا الشريبي، (2010)، "الإحصاء الوصفي و التطبيقي بإستخدام الحاسوب"، مكتبة الرشد. السعودية.
- عبد الرحيم محمد، (2015)، "مبادئ الإحصاء الوصفي و الإستدلالي"، دار التعليم الجامعي. مصر.
- محمد فوزي عبد الرحمن، (2008)، "الإحصاء الوصفي في العلوم الإجتماعية و السلوكية"، مكتبة الأنجلو المصرية. مصر.