PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY CENTER SI-ELHOUES–BARIKA

INSTITUTE OF SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

Course & Exercises:

1st Year of the Academic Degree in Mathematics and Computer Science

---

Descriptive Statistics & Introduction to Probability

Prepared by
Dr. Rania KHALLOUT

2023/2024

Preface

This course is designed for first-year undergraduate students in Mathematics and Computer Science. It covers the key topics of the subject, supported by relevant examples. Application exercises are included at the end of each chapter to help students assess their understanding and prepare for quizzes, tests, and final exams.

Based on my experience teaching this module for several years, I have created this booklet, which encompasses all the essential concepts related to the module, in accordance with the program outlined on Canvas.

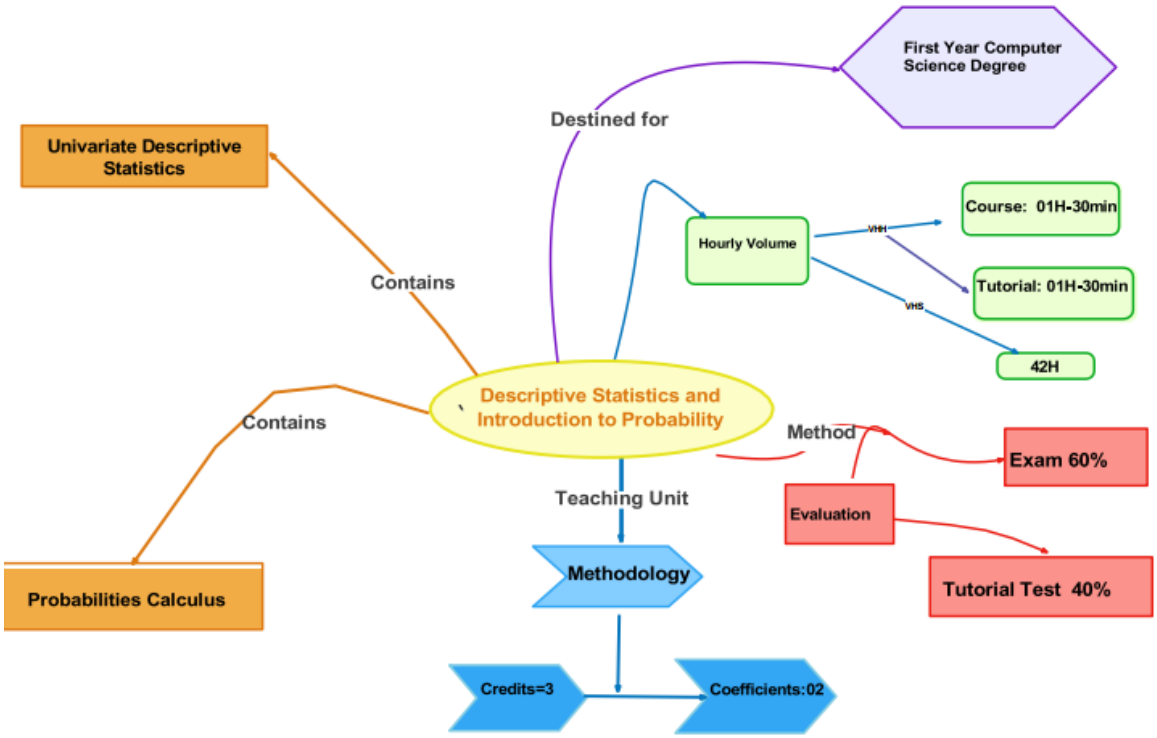

FIGURE 1: The mental map of the module.

Contents

CHAPTER 1

Basic concepts and statistical vocabulary

Statistics is the science of collecting, organizing, analyzing, and interpreting data. Its goal is to extract meaningful insights from data and support decision-making processes. **Statistical software**: Excel, SPSS, R, Matlab, Python, ...

This chapter introduces the fundamental terminology and concepts of statistics. It also discusses how to represent statistical data through tables.

1.1 Statistical vocabulary: key terms

1.1.1 Population

Refers to the entire group of items or individuals that share a common characteristic and are the subject of a statistical study.

Definition 1.1.1 *A **population** is the set on which a statistical study is based, denoted by* Ω.

Example 1.1.2 *Consider a group of students in Section A. If we examine the number of siblings, each student has:*

$$\Omega = \textit{the set of students in section A.}$$

Example 1.1.3 *For studying traffic in a city, the population consists of all vehicles that might circulate in the city on a specific date:*

$$\Omega = \textit{The set of vehicles.}$$

1.1.2 Sample

A sample refers to a subset of individuals, items, or data points selected from a larger population. It is used to make inferences or generalizations about the entire population without studying every member of it.

Definition 1.1.4 *A sample is the subset of the population. The sample size is noted by n.*

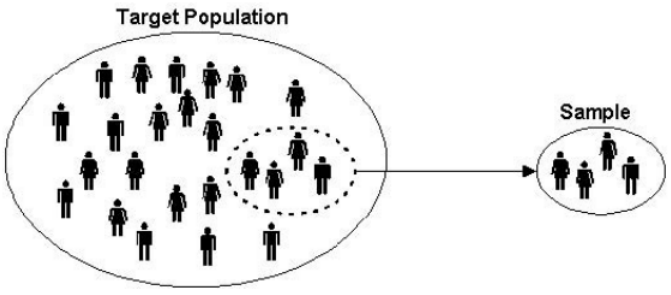

FIGURE 1.1: Sampling

1.1.3 Individual (statistical unit)

An individual, as a statistical unit, refers to the single, distinct entity being studied or observed in a statistical analysis. It represents the basic unit of data collection within a given population or sample.

Definition 1.1.5 *An individual is any element of the population* Ω, *denoted* ω $(\omega \in \Omega)$.

Remark 1.1.6

- The choice of an individual as the statistical unit depends on the research objective and context of the study.
- A statistical unit is an entity for which we aim to collect information.

1.1.4 Statistical Data (Variables or Characters)

Statistical Data refers to the information collected from individuals or units in a study to analyze patterns, trends, or relationships. This data is represented in the form of variables or characteristics.

Variable (character)

A variable is a measurable attribute or characteristic of an individual that can take different values.

Definition 1.1.7 *A statistical variable (denoted SV) is a function*

$$X : \Omega \to C$$

Where C is the set of values of the variable X (what is measured or observed on the individuals).

Example 1.1.8 *Height, temperature, nationality, eye color, socio-professional category, etc...*

1.1.5 Type of Variables

There are different ways variables can be described according to the ways they can be studied, measured, and presented. Variables are classified as follows:

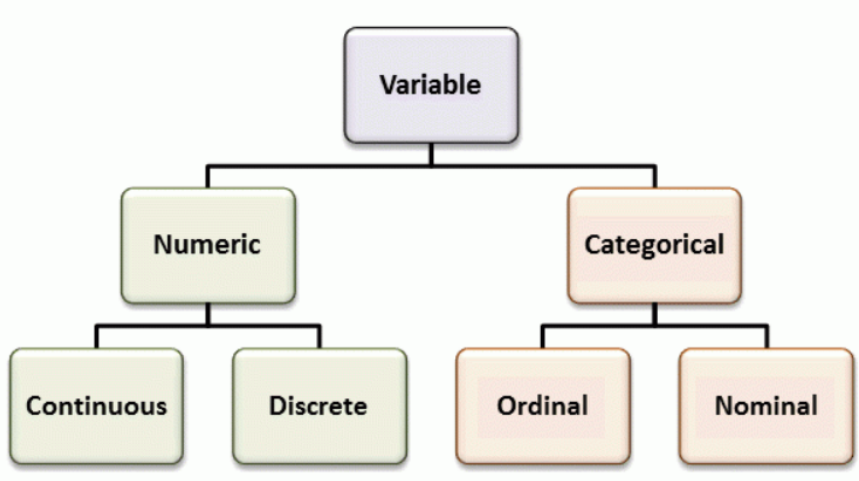

FIGURE 1.2: The type of a variable

1. Quantitative Variables (Numeric)

Represent numerical values. Can be discrete or continuous

- Discrete Variable: Takes only isolated values.

  Example: Number of children, number of accidents in a given period.

- Continuous Variable: Takes values within an interval.

  Example: Weight, height, blood glucose level.

2. Qualitative Variables (Categorical)

A qualitative variable is one whose values or modalities are expressed in a non-numeric manner or encoded without arithmetic meaning.

- **Nominal Scale:** Modalities are expressed as names without hierarchy.
- **Ordinal Scale:** Modalities indicate levels or degrees without specific numeric definition but with a clear hierarchy.

1.1.6 Data Representation

Data can be represented in various ways depending on the type of data and the purpose of the analysis, and it can be presented in various forms:

- Raw data presentation.
- Statistical tables.
- Graphical representations.
- Numerical representations.

1.1.7 Value

The value of a variable for an element is called **an observation** or **a measurement**.

Remark 1.1.9

- The value may be number or word.
- The set of all values is called data set.

1.2 Basic concepts

1.2.1 Frequency of value

The **frequency** (n_i) of a value x_i refers to the number of times x_i occurs in a dataset. It can be expressed as:

$$n_i = \text{number of occurrences of } x_i$$

The total **population size** (n) is given by the sum of all frequencies:

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

where:

- n: Total population size.
- k: Number of unique values in the dataset.

- n_i: Frequency of the i-th value.

Example 1.2.1 *For the dataset:* $\{2,2,3,3,3,4,5,5\}$*:*

- Unique values: $\{2,3,4,5\}$.
- Frequencies:

$$\begin{aligned} n_1 &= 2 \quad \text{(value } x_1 = 2 \text{ occurs 2 times)}, \\ n_2 &= 3 \quad \text{(value } x_2 = 3 \text{ occurs 3 times)}, \\ n_3 &= 1 \quad \text{(value } x_3 = 4 \text{ occurs 1 time)}, \\ n_4 &= 2 \quad \text{(value } x_4 = 5 \text{ occurs 2 times)}. \end{aligned}$$

- Total population size:

$$n = n_1 + n_2 + n_3 + n_4 = 2 + 3 + 1 + 2 = 8.$$

1.2.2 Relative frequency of value

The **relative frequency** (f_i) of a value x_i is the proportion of its frequency (n_i) compared to the total population size (n). It is given by:

$$f_i = \frac{n_i}{n}$$

It can also be expressed as a percentage:

$$f_i(\%) = \frac{n_i}{n} \times 100$$

Example 1.2.2 *We continue with the previous example*

- Total population size:$n = 8$
- Relative frequencies:

$$\begin{aligned} f_1 &= \frac{n_1}{n} = \frac{2}{8} = 0.25\,(25\%), \\ f_2 &= \frac{n_2}{n} = \frac{3}{8} = 0.375\,(37.5\%), \\ f_3 &= \frac{n_3}{n} = \frac{1}{8} = 0.125\,(12.5\%), \\ f_4 &= \frac{n_4}{n} = \frac{2}{8} = 0.25\,(25\%). \end{aligned}$$

Remark 1.2.3

- The sum of all relative frequencies in a dataset equals 1 (or 100% if it expressed as a percentage):

$$\sum_{i=1}^{k} f_i = 1 \quad \text{or} \quad \sum_{i=1}^{k} f_i(\%) = 100\%.$$

- A higher relative frequency indicates that the value is more common in the dataset.

1.2.3 Cumulative frequency

The **cumulative frequency** (N_i) of a value is the running total of frequencies up to that value. For a dataset with values $x_1, x_2, \ldots, x_k$, where n_i is the frequency of x_i, we have two types of cumulative frequency:

1. Increasing Cumulative Frequency

The increasing cumulative frequency is calculated as: Adding frequencies sequentially from the smallest to the largest value.

$$N_i \nearrow = \sum_{j=1}^{i} n_j$$

where:

- $N_i \nearrow$: Increasing cumulative frequency for the i-th value,
- n_j: Frequency of the j-th value.

2. Decreasing Cumulative Frequency

The cumulative frequency is calculated by subtracting frequencies sequentially, starting from the largest value.

$$N_i \searrow = n - \sum_{j=i}^{i-1} n_j$$

where:

- $N_i \searrow$: Decreasing cumulative frequency for the i-th value,
- n_j: Frequency of the j-th value.

Example 1.2.4 *For the dataset* $\{1, 1, 2, 2, 2, 3, 3, 4\}$*:*

Values (x_i)	1	2	3	4	$\sum$
Frequencies (n_i)	2	3	2	1	$n = 8$
ICF $(N_i \nearrow)$	2	5	7	8	–
DCF $(N_i \searrow)$	8	7	5	2	–

1.2.4 Relative frequency cumulative

The **relative cumulative frequency** represents the running total of relative frequencies in a dataset. The **cumulative frequency** (F_x) of a value is the running total of relative frequencies up to that value. For a dataset with values $x_1, x_2, \ldots, x_k$, where n_i is the frequency of n is size of the dataset, we have two types of relative frequency cumulative:

1. Increasing Relative Cumulative Frequency

Increasing relative cumulative frequency is the sum of relative frequencies from the smallest value up to the current value. It is given by:

$$F_i \nearrow = \sum_{j=1}^{i} \frac{n_j}{n} = \sum_{j=i}^{k} f_j$$

where:

- $F_i \nearrow$: Relative cumulative frequency for the i-th value,
- f_j: Relative frequency of the j-th value.

2. Decreasing Relative Cumulative Frequency

Decreasing relative cumulative frequency is the sum of relative frequencies from the largest value down to the current value. It is given by:

$$F_i \searrow = \sum_{j=i}^{k} \frac{n_j}{n} = \sum_{j=i}^{k} f_j$$

where:

- $F_i \searrow$: Decreasing relative cumulative frequency for the i-th value,
- k: Total number of unique values.

Example 1.2.5 *Consider the dataset* $\{1, 1, 2, 2, 2, 3, 3, 4\}$*:*

Values (x_i)	1	2	3	4	$\sum$
Frequencies (n_i)	2	3	2	1	$n = 8$
Relative Frequencies (f_i)	0.25	0.375	0.25	0.125	1
IRCF $(F_i \nearrow)$	0.25	0.625	0.875	1.0	–
DRCF $(F_i \searrow)$	1.0	0.875	0.625	0.25	–

1.3 Dataset and Frequency Table

A dataset is a collection of data points that can be organized in a structured manner, such as a list of numbers, names, or categories.
A frequency table is a summary of how often each value or category appears in a dataset. It helps in understanding the distribution of data.

1.3.1 Case of Quantitative Discrete Data

Quantitative discrete data consists of countable numerical values, meaning the values are whole numbers and cannot be subdivided (e.g., number of students, cars, or items).

Example 1.3.1 *Consider a survey of 20 students who reported the number of books they read in a month. The data collected is as follows:*

$$2, 3, 4, 5, 3, 2, 5, 6, 4, 3, 5, 7, 3, 4, 5, 6, 2, 3, 4, 5$$

Based on this data, we find the following frequency table:

Values(x_i)	2	3	4	5	6	7	$\sum$
Frequencies(n_i)	3	5	4	5	2	1	$n = 2$

- The population studied is the set (group) of students.
- The population size : $n = 20$.

- The variable X studied is the number of books readded per month.
- The type of X is a quantitative discrete.

We add the lines for calculating f_i, $N_i \nearrow$, $N_i \searrow$, $F_i \nearrow$ and $F_i \searrow$ as follows :

Values(x_i)	2	3	4	5	6	7	$\sum$
Frequencies(n_i)	3	5	4	5	2	1	20
Relative Frequencies(f_i)	0.15	0.25	0.20	0.25	0.10	0.05	1
ICF$(N_i \nearrow)$	3	8	12	17	19	20	–
DCF$(N_i \searrow)$	20	17	12	8	3	1	–
ICRF$(F_i \nearrow)$	0.15	0.40	0.60	0.85	0.95	1.00	–
DCRF$(F_i \searrow)$	1.00	0.85	0.60	0.40	0.15	0.05	–

1.3.2 Case of Quantitative Continuous Data

If x is a continuous quantitative variable, with classes: $[a_0, a_1[, \cdots, [a_{p-1}, a_p]$, individuals are grouped into intervals, partitioning the range of possible values. For p intervals, the data table is structured as follows:

Classes	$[a_0, a_1[$	$[a_1, a_2[$	$[a_2, a_3[$	$\cdots$	$[a_{p-1}, a_p[$	$\sum$
Frequencies n_i	n_1	n_2	n_3	$\cdots$	n_p	n
Classes Center c_i	$\frac{a_0+a_1}{2}$	$\frac{a_1+a_2}{2}$	$\frac{a_2+a_3}{2}$	$\cdots$	$\frac{a_{p-1}+a_p}{2}$	–
ICF $N_i \nearrow$	n_1	n_1+n_2	$n_1+n_2+n_3$	$\cdots$	n	–
DCF $N_i \searrow$	n	$n-n_1$	$n-n_1-n_2$	$\cdots$	$n-n_1-\cdots-n_{p-1}$	0
IRCF $F_i \nearrow$	$\frac{n_1}{n}$	$\frac{n_1+n_2}{n}$	$\frac{n_1+n_2+n_3}{n}$	$\cdots$	$\frac{n}{n}=1$	–
DRCF $F_i \searrow$	$\frac{n}{n}$	$\frac{n-n_1}{n}$	$\frac{n-n_1-n_2}{n}$	$\cdots$	$\frac{n-n_1-\cdots-n_{p-1}}{n}$	0

Definition 1.3.2 *Let* $i \in \{1, \ldots, p\}$

1. Class Amplitude (class width): $[a_{i-1}, a_i[$ is
$$l_i := a_i - a_{i-1}.$$
2. Proportion Density: $[a_{i-1}, a_i[$ is
$$d_i = \frac{f_i}{l_i}.$$

Classes	$[a_0, a_1[$	$[a_1, a_2[$	$\cdots$	$[a_{p-1}, a_p[$
Frequencies n_i	n_1	n_2	$\cdots$	n_p
Class Centers c_i	c_1	c_2	$\cdots$	c_p
Relative Frequencies f_i	$f_1 = \frac{n_1}{n}$	$f_2 = \frac{n_2}{n}$	$\cdots$	$f_p = \frac{n_p}{n}$
Amplitude l_i	$l_1 = a_1 - a_0$	$l_2 = a_2 - a_1$	$\cdots$	$l_p = a_p - a_{p-1}$
Proportion Density d_i	$d_1 = \frac{f_1}{l_1}$	$d_2 = \frac{f_2}{l_2}$	$\cdots$	$d_p = \frac{f_p}{l_p}$

Remark 1.3.3

- Proportion density helps compare frequencies across classes considering class width.

- If all classes have the same width, calculating proportion density is unnecessary.

Example 1.3.4 *A teacher collects the following tutorial marks for a group of students:*

8,25	14,25	9,5	14	6,25	11,75	10	17	14,75	11,25	19,25	10
8	6	12,75	10	18,75	11,5	13,5	12,25	13	18,5	15	17
6	12,25	4,75	16	10,75	9,75	15,5	12	16,5	15,25	10	3.

- Population: All students in the group.
- Population Size: $n = 36$.
- Variable x: tutorial marks.
- Type of variable is quantitative discrete.

Since there are many distinct values, we group the data into 5 classes with an amplitude of 4. We obtain the following table:

ClassInterval	$[0,4[$	$[4,8[$	$[8,12[$	$[12,16[$	$[16,20[$	$\sum$
Frequency	1	4	12	12	7	36

In the case of a continuous quantitative variable, constructing the frequency table requires first grouping the data into classes. This involves determining the expected number of classes and, consequently, the amplitude associated with each class or class interval.

Several empirical formulas allow the determination of the number of classes for a sample of size n:

- **Sturges's rule**: Number of classes: $k = 1 + 3,322(\log n)$.
- **Yule's rule**: Number of classes: $k = 2,5\left(\sqrt[4]{n}\right)$.
- The interval for each class is then obtained as follows:

$$C = \frac{(X_{\max} - X_{\min})}{k}.$$

where X_{max} and X_{min}, are, respectively, the largest and smallest values of X in the dataset.

Example 1.3.5 *We have the following dataset of heights (in cm):*

150	152	153	155	158	160	161	163	165	167
168	170	172	173	175	177	178	180	182	185

1. Determine the Number of Classes
Using **Sturges's Rule**:

$$k = 1 + 3.322 \log(20)$$

Approximating:

$$k = 1 + 3.322 \times 1.301 = 5.324$$

Rounded up: **k = 6** classes.

2. Compute the Class Interval

$$C = \frac{185 - 150}{6} = 5.83 \approx 6$$

3. Construct the Frequency Table

Class Interval	**[150-156[**	**[156-162[**	**[162-168[**	**[168-174[**	**[174-180[**	**[180-186[**	$\sum$
Frequency n_i	4	3	3	4	3	3	20
RF (f_i)	0,2	0,15	0,15	0,2	0,15	0,15	1
RF (%)	20%	15%	15%	20%	15%	15%	**100%**

1.3.3 Case of Qualitative Nominal Data

A nominal qualitative variable consists of categories that have no inherent order or ranking. Each category is distinct, and the only meaningful analysis is based on grouping and frequency counting.

Example 1.3.6

A study is conducted to analyse the preferred social media platforms of 500 users. The collected data is summarized in the following table:

Social Media Platform	**Facebook**	**Instagram**	**Twitter (X)**	**TikTok**	**LinkedIn**	$\sum$
Frequency (n_i)	120	150	80	90	60	**500**
Relative Frequency (f_i)	0,24	0,30	0,16	0,18	0,12	1
Relative Frequency (%)	24%	30%	16%	18%	12%	**100%**

The variable (social media platform) consists of categories that have no natural ranking.

- Population: Users of social medial platforms.
- Population Size: $n = 500$.
- Variable x: Social media platforms.
- Type of variable is qualitative nominal.

1.3.4 Case of Qualitative Ordinal Data

Example 1.3.7

A company conducted a survey to analyse the highest level of education attained by its 300 employees. The results are presented in the table below.

Educational Level	**High School**	**Diploma**	**Bachelor's**	**Master's**	**PhD**	$\sum$
Frequency (n_i)	50	70	100	60	20	**300**
Relative Frequency (f_i)	0,167	0,233	0,333	0,2	0,067	1
Relative Frequency (%)	16.7%	23.3%	33.3%	20%	6.7%	**100%**

- The variable (educational qualification) follows a ranked order:
  High School < Diploma < Bachelor's < Master's < PhD.
- The data allows for comparison and ordering, but the differences (e.g., between a Bachelor's and a Master's) are not necessarily equal.
- Population: Employees.
- Population Size: $n = 300$.
- Variable x: level of education.
- Type of variable is qualitative ordinal.

1.4 Exercises

Exercise 1 *A company conducted a survey to determine how many cups of coffee employees drink per day. The data collected from 40 employees is as follows:*

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

1) Identify the statistical series (population, sample size, studied variable, and its type).

2) Construct a frequency table including: frequency, Cumulative frequency, Relative frequency.

Exercise 2 *The weekly revision times (in hours) for a group of students are listed below in increasing order:*

4	7	8	9	10	11	12	12	12	13
14	14	14	15	16	16	17	17	19	21

1) Determine the population studied, the population size, the variable studied, and its type.

2) Using Sturge's rule (or Yule's rule), grouping the measures of the previous data set.

3) Construct the statistical table by calculating : Increasing Cumulative Frequency (ICF) and Increasing Cumulative Relative Frequency (ICRF).

Exercise 3 *The observed results of the sequence of a DNA strand are:*

G	G	A	T	A	G	C	T	A	G	G	A	T	G	C	C	T
A	G	T	A	G	A	T	C	G	A	G	C	T	G	C	T	A
C	C	G	A	T	C	G	C	T	C	C	T	C	T	G	C.	

$A : Adenine, \quad G : Guanine, \quad C : Cytosine, \quad T : Thymine.$

1) Identify the statistical series presented above (the population, its size, the studied variable, and its nature).

2) Provide the distribution of partial frequencies.

3) Calculate the relative frequencies.

Exercise 4 *Consider the following list of students' names, with each name followed by the number of books they read in a year (shown in parentheses):*

$$1 = few,\ 2 = moderate,\ 3 = many,\ 4 = exceptional$$

Ali (3), Salim (3), Foued (1), Soltane (2), Adel (1), Soufiane (2), Adnnane (3), Line (2), Souad (2), Farida (3), Ilhame (4), Abdelhakim (2), Djamil (1), Hocine (3), Zaher (3), Djamel (3), Taha (3), Fateh (4), Brahim (3), Jihane (3).

1) Define the distribution of these students according to their reading habits (population, variable, etc...).

2) Construct a representative table of this distribution.

3) Complete the table by calculating: Cumulative Frequency (ICF, DCF) and Cumulative Relative Frequency (ICRF, DCRF).

CHAPTER 2

Graphical Representation of Statistical Data

This chapter explores different methods of graphically representing statistical data, their applications, and how to choose the appropriate visualization technique depending on the nature of the data. Here's an overview of graphing ungrouped and grouped data:

2.1 Graphing Ungrouped Data

Ungrouped data consists of raw observations without classification into intervals. Some common ways to represent ungrouped data graphically include:

2.1.1 Bar Graph

- Used for categorical or discrete numerical data.
- Represents comparisons among different categories.
- Each category is displayed as a separate bar.

Example 2.1.1 *Consider the number of students achieving specific grades in a test given by:*

Grades	***A***	***B***	***C***	***D***	***F***	$\sum$
Frequency (n_i)	10	15	30	20	5	***80***

Graph Presentation:

- The x-axis represents different grades (A, B, C, D, F).
- The y-axis represents the number of students.
- Each bar's height corresponds to the number of students in that category.

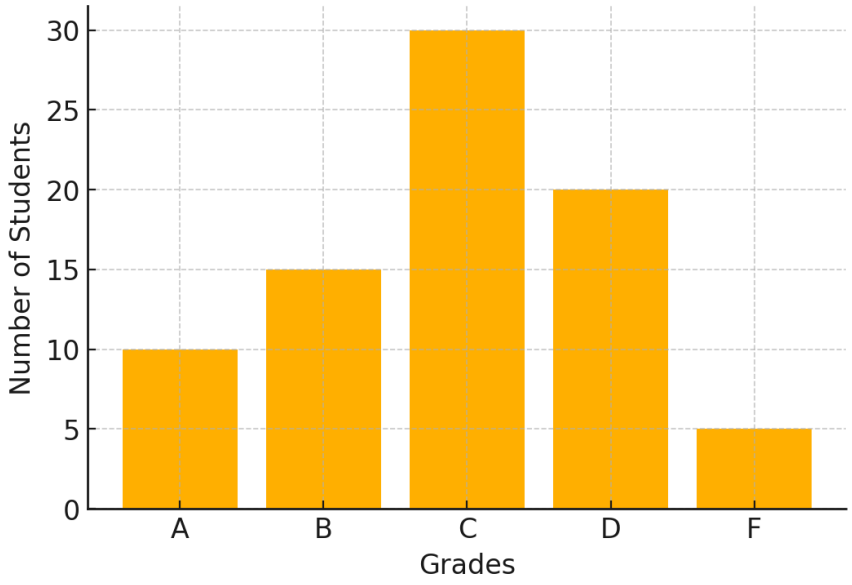

FIGURE 2.1: The number of students achieving specific grades in a test

2.1.2 Line Graph

- Best for showing trends over time.
- Represents continuous data with points connected by lines.
- Helps in identifying patterns and forecasting.

Example 2.1.2 *Consider the temperature changes monthly in a year:*

Months	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
Temperature	5	7	12	18	22	26	30	28	24	18	10	6

Graph Presentation:

- The x-axis represents months (January to December).
- The y-axis represents temperature values.
- Points are plotted for each month and connected by a line to show trends.

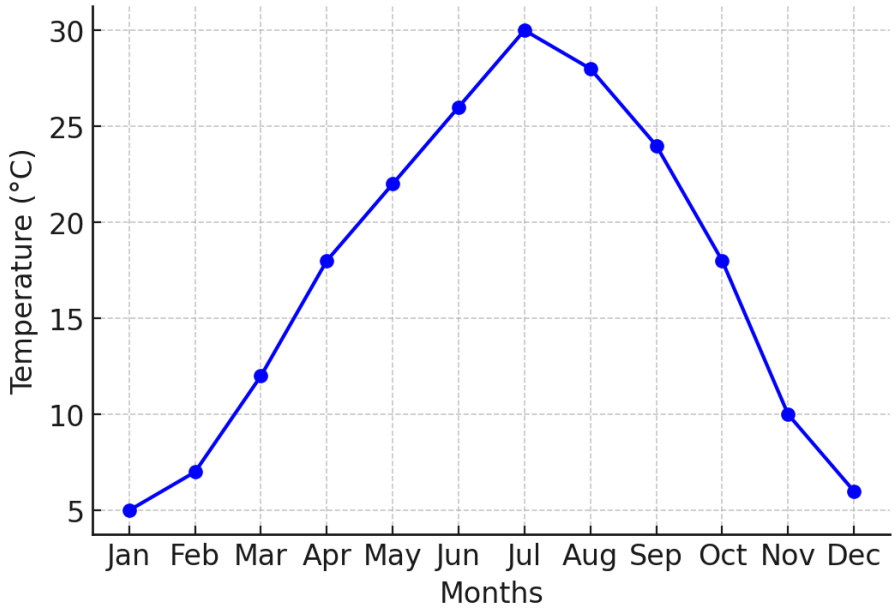

FIGURE 2.2: Monthly temperature changes in a year

2.1.3 Pie Chart

- Displays data as proportional slices of a circle.
- Best for showing percentages and relative proportions.
- Useful for showing part-to-whole relationships.

Example 2.1.3 *Consider the distribution of expenses in a household budget:*

Household budget	*Entertainment*	*Transport*	*Food*	*Rent*	*Saving*	$\sum$
Expenses %	10%	15%	25%	40%	10%	100

Graph Presentation:

- Each category (Rent, Food, Transport, etc.) is represented as a slice.
- The size of each slice is proportional to the percentage of total expenses.
- Labels and percentages help in better understanding.

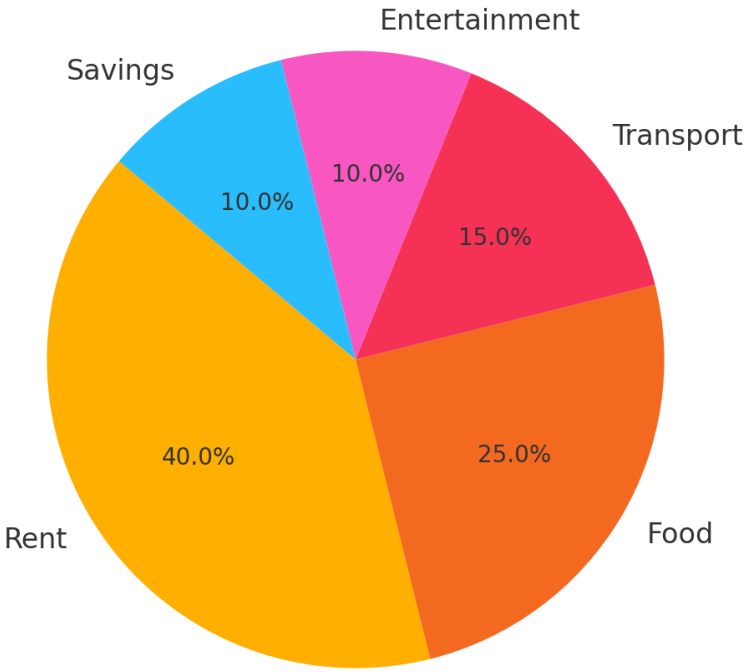

FIGURE 2.3: The distribution of expenses in a household budget

2.1.4 Scatter Plot

- Used to display relationships between two numerical variables.
- Helps identify correlations (positive, negative, or no correlation).
- Common in statistical regression analysis.

Example 2.1.4 *The relationship between study hours and exam scores is given by:*

Study hours	0.5	2	3	4	5	6	7	8	9	10
Exam scores	50	55	60	63	70	72	75	77	85	90

Graph Presentation:

- The x-axis represents study hours.
- The y-axis represents exam scores.
- Each point represents a student's study hours and corresponding score.
- The pattern of points shows the correlation (positive in this case).

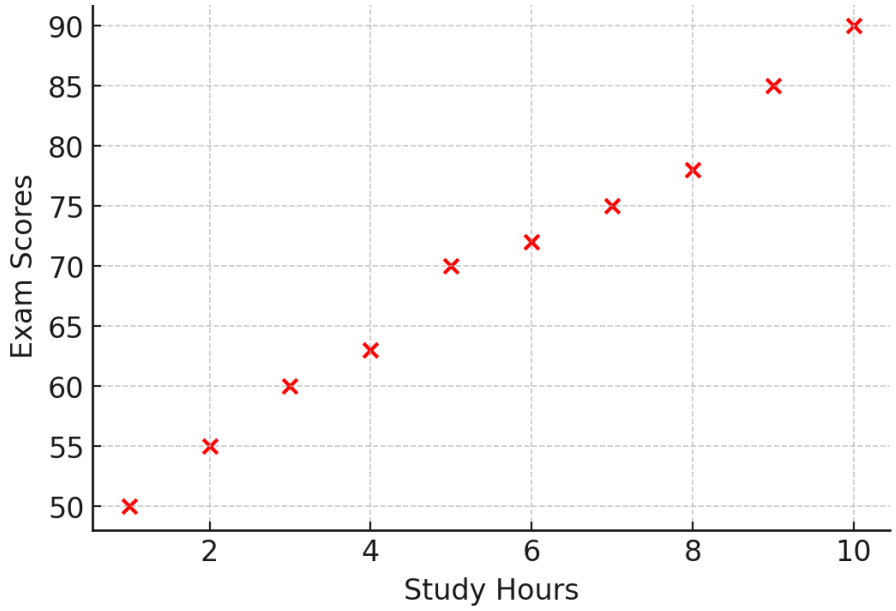

FIGURE 2.4: The relationship between study hours and exam scores

2.2 Graphing Grouped Data

Grouped data is organized into classes or intervals. The most common methods to represent grouped data are:

2.2.1 Histogram

A bar graph where bars represent frequency intervals. Used for continuous data.

Example 2.2.1 *Consider heights of students grouped into the following intervals:*

Height (cm)	[150-160[	[160-170[	[170-180[	[180-190[
Frequency	1	12	5	1

Graph Presentation:

- X-axis: Represents height intervals (150–160 cm, 160–170 cm, etc.).
- Y-axis: Represents the frequency (number of students) in each height interval.
- Bars: Each bar represents the number of students whose heights fall within each specific interval.

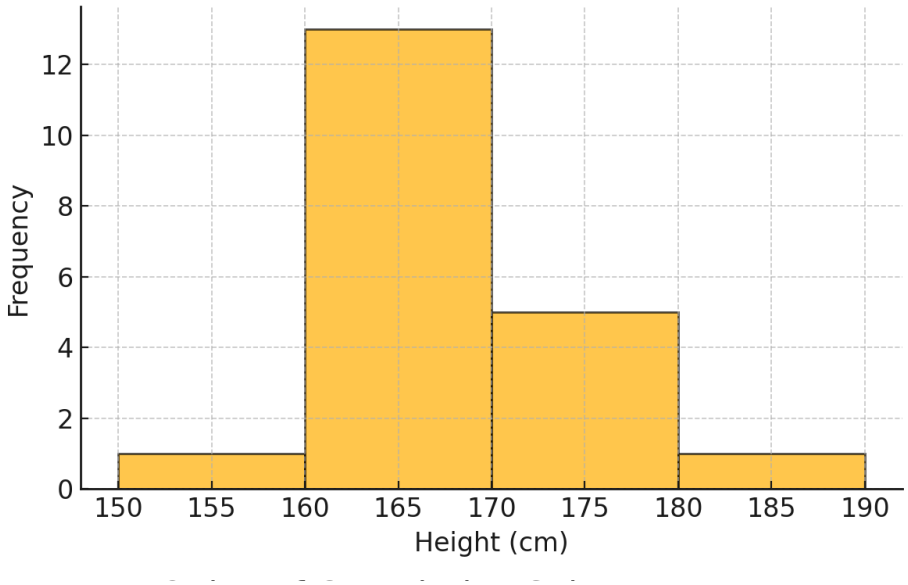

FIGURE 2.5: Histogram of students heights

Remark 2.2.2

The histogram gives an overview of the distribution of student heights, showing how students are distributed across the different height intervals.

2.2.2 Frequency Polygon

Similar to a histogram but uses lines instead of bars. Plots midpoints of class intervals and connects them.

Example 2.2.3 *Showing the following distribution of test scores in a large dataset:*

Score intervals	[40-50[	[50-60[	[60-70[	[70-80[	[80-90[	[90-100[
Frequency	1	2	2	3	3	2

Graph presentation:

- X-axis: Represents the midpoints of the test score intervals (e.g., 45,55,65, etc.).
- Y-axis: Represents the frequency (number of students) in each interval.
- Line: Each point on the line corresponds to the frequency of students in that particular interval.

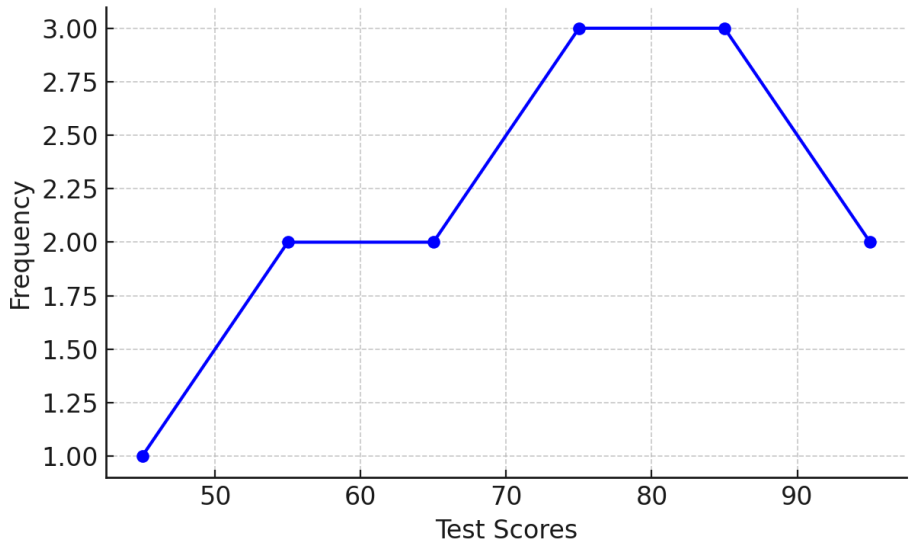

FIGURE 2.6: Frequency polygon of test scores

Remark 2.2.4

The frequency polygon is useful for showing the shape of the distribution and for comparing multiple data sets on the same graph.

2.2.3 Curve

curve generally refers to any smooth, continuous line or path, but in the context of statistics and data visualization, the term "curve" often refers to the line that connects the points plotted on a graph.

- **Curve:** A general term used to describe a smooth line that connects data points. It could refer to any type of graph or line on a plot, such as a bell curve for a normal distribution, a trend line, or the line of best fit.
- **Ogive Curve**: a specific type of curve used in cumulative frequency graphs. The ogive curve specifically shows the accumulation of data values across class intervals.

Example 2.2.5 *The cumulative number of people earning below a certain salary threshold:*

Salary (USD)	[1000-2000[	[2000-3000[	[3000-4000[	[4000-5000[	[5000-more[
Cumulative Frequency	0	3	5	7	7

Graph presentation:

- X-axis: Represents salary intervals (1000–2000, 2000–3000, etc.).
- Y-axis: Represents the cumulative number of people earning below each salary threshold.
- Cumulative line: The cumulative line rises as we move along the X-axis, indicating the total number of people below each salary level.

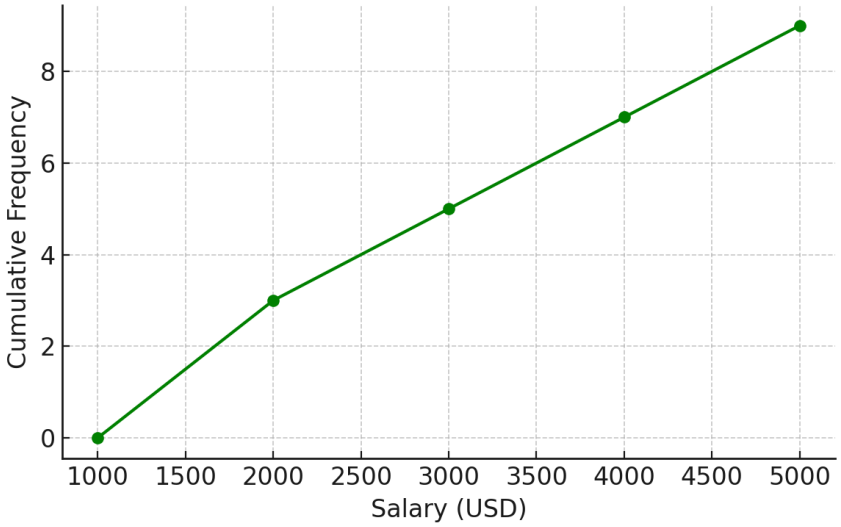

FIGURE 2.7: Curve of cumulative salary frequency

Remark 2.2.6

The ogive helps identify the percentage of the population below certain salary thresholds, providing insight into salary distribution.

2.2.4 Box Plot (Box-and-Whisker Plot)

Shows the median, quartiles, and outliers in the data set. Helps identify distribution and spread.

Example 2.2.7 *Visualization of the income distribution in different regions.*

Graph presentation:

- X-axis: Represents the income data.
- Box: The box represents the interquartile range (IQR), which is the middle 50% of the data. The upper and lower edges of the box correspond to the first (Q1) and third (Q3) quartiles, while the line in the middle represents the median.
- Whiskers: The "whiskers" extend from the quartiles to the minimum and maximum values within a specified range (usually 1.5 times the IQR from the quartiles).
- Outliers: Data points that lie beyond the whiskers are considered outliers and are typically represented as individual points.

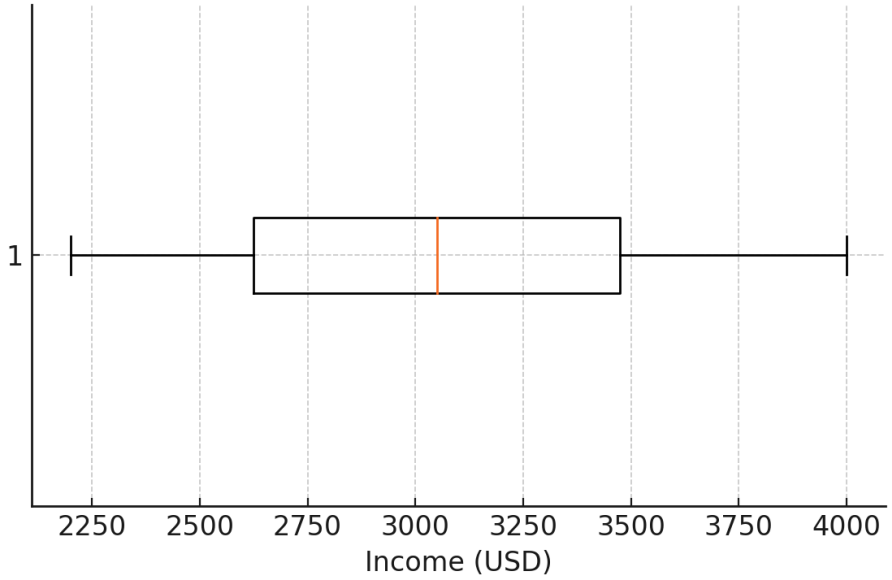

FIGURE 2.8: Box plot of income distribution

Remark 2.2.8

Box plots are useful for comparing distributions, especially when you want to quickly identify the range, central tendency, and spread of the data.

2.3 Exercises

Exercise 5 *The following table represents the number of absences of* 233 *students:*

x_i	0	1	2	3	4
n_i	90	45	60	20	18

1) Represent the statistical table showing the relative frequencies and the increasing and decreasing relative cumulative frequencies.

2) What is the proportion of students who have at least 3 absences?

3) What does the value represent: $1 - (f_1 + f_2)$?

4) Draw the relative cumulative curve of increasing frequencies.

Exercise 6 *The table below shows the blood group of* M_2 *Computer Science students:*

A	O	O	O	A	AB	B	B	A	A
B	AB	A	A	B	O	A	A	O	B
A	O	B	O	O	A	A	O	O	A
A	B	AB	O	A	B	O	O	A	A
O	O	A	A	AB	B	A	A	O	A
AB	B	A	A	AB	B	A	O	O	B

1) Construct a statistical table including relative frequencies in percentage.

2) Give two appropriate graphical representations for this type of data.

Exercise 7 *A car dealership wants to analyze its sales performance for the year 2024. The collected data concernthe number of cars sold each week over the 48 working weeks. These data are grouped into classes, as shown in the following table:*

x	$[0-3[$	$[3-6[$	$[6-9[$	$[9-12[$
n_i	9	14	13	10

1) Draw the histogram of this distribution, ensuring the correct scale for the axes.

2) Draw the polygon of cumulative frequencies and the cumulative curve.

Exercise 8 *The study of the weight of 48 people yielded the following results in kg, represented in an ordered statistical series:*

37 43 47 50 52 54 55 56 58 61 62 63 64 65 66 66 67 68 69 69 69 70 71 72

72 72 73 73 74 74 75 76 77 77 79 79 80 82 82 84 86 87 88 90 92 93 97 97

1) Determine the size of the population.

2) Group the data into classes according to Sturges' rule.

3) Represent the statistical table showing the frequencies and the increasing and decreasing cumulative frequencies.

4) Give the percentage proportion of people whose weight X is greater than or equal to 80 kg

5) Draw the histogram of frequencies and the frequency polygon.

6) Draw the cumulative curves of increasing and decreasing frequencies.

CHAPTER 3

Numerical Representation of Statistical Data

The statistical data can be represented in numerical form using various methods to summarize, analyse, and interpret data effectively. This chapter introduces the numerical representations that enable students to draw meaningful conclusions and make informed decisions. The following are the key numerical methods used to represent statistical data:

3.1 Measures of Central Tendency

The central tendency measures are statistical tools used to summarize a data set by identifying a single value that best represents the center of the distribution. These measures provide insights into the general trend of the data and are essential for understanding the overall distribution. The three primary measures of central tendency are mean, median, and mode. Each measure has its own characteristics and is used in different contexts depending on the nature of the data. The approach to computing these measures differs depending on whether the data are ungrouped (raw data) or grouped into classes.

3.1.1 Mean (Arithmetic Mean)

Mean For Ungrouped Data

The mean is the most commonly used measure of central tendency. Calculated by summing all the data values and dividing the sum by the total number of data points.

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{\sum n_i x_i}{n}$$

where:

- $\bar{x}$ is the mean,
- x_i represents each individual data point,
- n_i represents frequency for each individual data point,
- n is the total number of data points.

Characteristics of the Mean

- It is sensitive to extreme values (outliers). A very high or very low value can distort the mean.
- It provides a precise measure of the central point when the data is normally distributed.

Graphically

- The mean is the balance point of a bar chart. It can be visualized as the point where the distribution of the data is balanced.
- In a symmetric distribution, the mean is located at the center of the distribution.

Mean For Grouped Data

Let c_i denote the midpoint of each class, and n the frequency for that class. The mean is estimated by:

$$\bar{x} = \frac{\sum n_i c_i}{\sum n_i} = \frac{\sum n_i c_i}{n}$$

where:

- c_i is the midpoint of each class (class center),
- n_i is the frequency for that class,
- n is the total number of data class.

3.1.2 Median

Median For Ungrouped Data

The median is the middle value in an ordered data set, the data is arranged from smallest to largest (or vice versa). If the data set has an odd number of values, the median is the middle number. If the data set has an even number of values, the median is the average of the two middle values.

Steps to Find the Median

1. Order the data from smallest to largest.
2. If n is odd, the median is the value at position $\frac{n+1}{2}$
3. If n is even, the median is the average of the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$

$$X_{med} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Graphically

- Find Half the Total Frequency: If your total number of observations is n compute $\frac{n}{2}$ or $\frac{n+1}{2}$ when using the positional method for odd-sized data.
- Locate on the Graph: On the vertical axis, find the point corresponding to $\frac{n}{2}$.
- Draw a Horizontal Line: Draw a horizontal line from this point until it intersects the cumulative frequency curve.
- Drop a Vertical Line: From the intersection, drop a vertical line down to the x-axis.
- Read the Median: The x-coordinate where this vertical line meets the axis is the median value.

Median For Grouped Data

The median class is the class in which the $\frac{n}{2}$ th observation lies. The formula interpolates within this class to estimate the median value. Then the median is calculated by:

$$X_{med} = a + \left(\frac{\frac{n}{2} - N_{\frac{n}{2}-1}}{N_{\frac{n}{2}} - N_{\frac{n}{2}-1}} \right) \times l$$

where:

- $n = \sum n_i$ be the total frequency,
- a be the lower boundary of the median class,
- $N_{\frac{n}{2}-1}$ be the cumulative frequency of all classes before the median class,
- $N_{\frac{n}{2}}$ be the cumulative frequency of the median class,
- l be the class width (amplitude class).

3.1.3 Mode

Mode For Ungrouped data

The mode is the value that appears most frequently in a dataset. A dataset can have:

- One mode (unimodal),
- More than one mode (bimodal or multimodal),
- Or no mode if all values occur with equal frequency.

Graphically

- In a bar graph, the mode corresponds to the highest bar, indicating the most frequent value or range of values.
- In a frequency line graph, the mode is the point at which the graph reaches its highest peak.

Mode

The **modal class** is defined as the class interval that has the highest frequency. In other words, you simply look at your frequency distribution table and find the class interval with the maximum frequency. That interval is the modal class. Once the modal class is identified, the mode for grouped data can be estimated using the formula:

$$X_{Mod} = a + \left(\frac{n_{mod} - n_{mod-1}}{(n_{mod} - n_{mod-1}) + (n_{mod} - n_{mod+1})} \right) \times l$$

where:

- a is the lower boundary of the modal class,
- n_{mod} is the frequency of the modal class,
- n_{mod-1} is the frequency of the class preceding the modal class,
- n_{mod+1} is the frequency of the class succeeding the modal class,
- l is the class amplitude (width).

Example 3.1.1 ***Applied Central Tendency Measures in Ungrouped Data***
Consider a group of students achieving specific marks in a test as follows:

Data values	5	7	8	9	10	12	$\sum$
Frequency (n_i)	3	1	1	1	1	1	***8***

Compute the Measures:

- Mean:
$$\bar{x} = \frac{3 \times 5 + 7 + 8 + 9 + 10 + 12}{8} = \frac{61}{8} = 7.625$$
- Median: we have $n = 8$ is even, then
$$X_{med} = \frac{X_{\frac{8}{2}} + X_{\frac{8}{2}+1}}{2} = \frac{7+8}{2} = 7.5$$
- Mode: The value 5 occurs most frequently, $X_{mod} = 5$

Plotting the Bar Graph:

- The horizontal axis shows the data values.
- The vertical axis shows their frequency.
- We highlight the mode by coloring its bar differently.
- We add vertical dashed lines for the mean and median and annotate them.

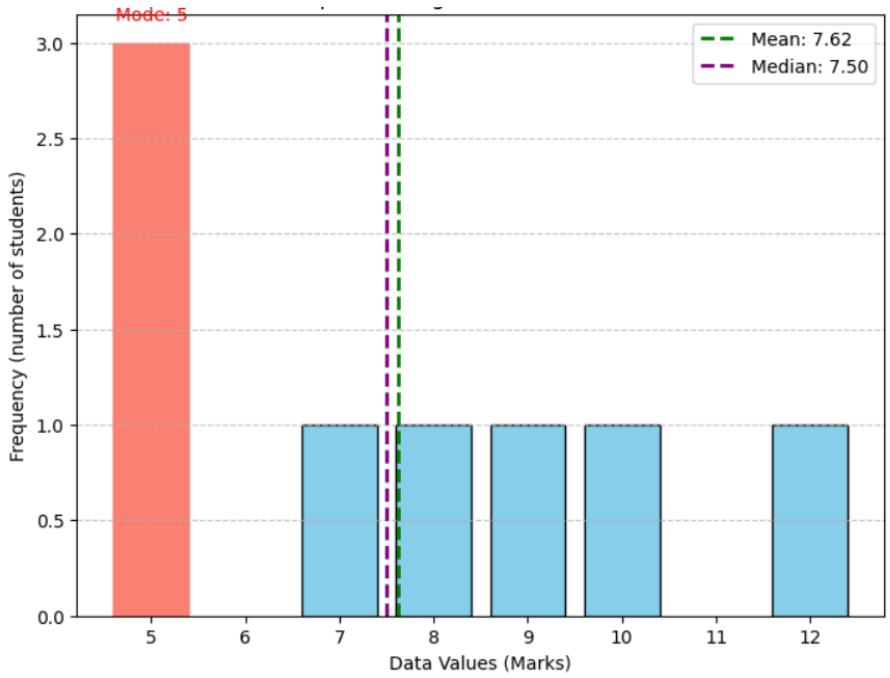

FIGURE 3.1: Bar Graph Showing Mean, Median and Mode

Example 3.1.2 ***Applied Central Tendency Measures in Grouped Data***
Consider heights of students grouped into the following intervals:

Height (cm)	$[150-160[$	$[160-170[$	$[170-180[$	$[180-190[$	$\sum$
Frequency	1	12	5	2	20

We first use the midpoints of the class intervals for our calculations:

Height (cm)	$[150-160[$	$[160-170[$	$[170-180[$	$[180-190[$	$\sum$
Frequency(n_i)	1	12	5	2	20
Midpoints(c_i)	155	165	175	185	–

Mean:

$$\bar{x} = \frac{1(155) + 12(165) + 5(175) + 2(185)}{20} = \frac{3380}{20} = 169$$

Then we calculate the cumulative frequencies:

Height (cm)	$[150-160[$	$[160-170[$	$[170-180[$	$[180-190[$	$\sum$
Frequency(n_i)	1	12	5	2	20
CF$(N_i \uparrow)$	1	13	18	20	–

Median:
The median is found by first identifying the median class: **The median class** is the class in which the $\frac{n}{2} = \frac{20}{2} = 10$ th observation lies which is [160-170[. Then the median is calculated by:

$$X_{med} = 160 + \left(\frac{10-1}{13-1}\right) \times 10 = 167.5$$

Mode:
The **modal class** is defined as the class interval that has the highest frequency, which is [160-170 [, the mode for the grouped data can be estimated using the formula:

$$X_{Mod} = 160 + \left(\frac{12-1}{2(12)-1-5}\right) \times 10 = 166.11$$

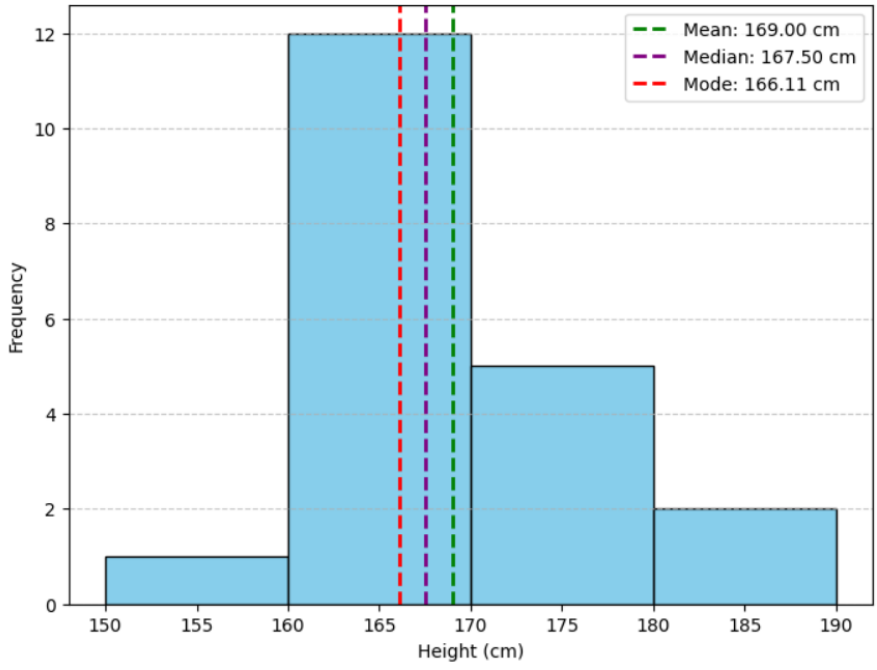

FIGURE 3.2: Histogram Showing Mean, Median and Mode

3.2 Measures of Dispersion

Measures of dispersion, also called spread or variability measures, describe the extent to which data values deviate from the central value (mean or median) of the distribution. These measures give an understanding of how data points are spread out in a dataset. Common measures of dispersion include range, variance, and standard deviation. Below is a breakdown of these measures for both ungrouped and grouped data.

3.2.1 Range

Range For Ungrouped Data

The range is the simplest measure of dispersion. It is the difference between the highest and lowest values in the data set.

$$\text{Range} = X_{\text{max}} - X_{\text{min}}$$

Where:

- X_{max} is the maximum value in the data set.
- X_{min} is the minimum value in the data set.

Range For Grouped Data

For grouped data, the range is still the difference between the maximum and minimum class limits. However, since we use class intervals, the exact range might be approximate.

$$\text{Range} = \text{Upper boundary of the last class} - \text{Lower boundary of the first class}$$

3.2.2 Variance

Variance For Ungrouped Data

Variance measures the average squared deviation of each data point from the mean. A high variance indicates that data points are spread around the mean, while a low variance indicates that data points are close to the mean.

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

Where:

- x_i represents each data point.
- $\bar{x}$ is the mean of data.
- n is the size of data.

Variance For Grouped Data

To calculate the variance, we first compute the midpoints c_i for each class interval. These midpoints are then used in the formulas.

1. Find the midpoints of each class interval, denoted by c_i.
2. Compute the squared deviation of each midpoint from the mean, $(c_i - \bar{x})^2$, where $\bar{x}$ is the mean calculated using midpoints.
3. Weight the squared deviations by the frequency of each class, n_i, and sum them up to get the total squared deviation.

Formula for Variance:

$$Var(x) = \frac{\sum n_i(c_i - \bar{x})^2}{n}$$

Where:

- n_i is the frequency of class i.
- c_i is the midpoint of class i.
- $\bar{x}$ is the mean of the midpoints.
- $n = \sum n_i$ is the total frequency (sum of all class frequencies).

3.2.3 Standard Deviation

Standard Deviation For Ungrouped Data

Standard deviation is the square root of variance. It gives a measure of the average distance of each data point from the mean, making it more interpretable than variance.

$$\sigma = \sqrt{Var(x)} = \sqrt{\frac{\sum n_i(x_i - \bar{x})^2}{n}}$$

Standard Deviation For Grouped Data

To calculate the standard deviation for grouped data, we first compute the midpoints c_i for each class interval. These midpoints are then used in the formulas.

1. Find the midpoints of each class interval, denoted by c_i.
2. Compute the squared deviation of each midpoint from the mean, $(c_i - \bar{x})^2$, where $\bar{x}$ is the mean calculated using midpoints.
3. Weight the squared deviations by the frequency of each class, n_i, and sum them up to get the total squared deviation.

Formula for Standard Deviation:

$$\sigma = \sqrt{\frac{\sum n_i(c_i - \bar{x})^2}{n}}$$

Where:

- n_i is the frequency of class i.
- c_i is the midpoint of class i.
- $\bar{x}$ is the mean of the midpoints.
- $n = \sum n_i$ is the total frequency (sum of all class frequencies).

3.2.4 Coefficient of Variation

Coefficient of Variation For Ungrouped Data

The coefficient of variation of a statistical variable x is defined as the ratio:

$$C_v(x) = \frac{\sigma_x}{\overline{x}}.$$

Remark 3.2.1

- The coefficient of variation is used to assess the **homogeneity** of the distribution.
- A higher coefficient of variation indicates more variability, while a lower coefficient of variation suggests consistency.

Coefficient of Variation For Grouped Data

To calculate the Coefficient of Variation for grouped data, we first compute the midpoints c_i for each class interval. These midpoints are then used in the formulas.

1. Find the midpoints of each class interval, denoted by c_i.
2. Compute the squared deviation of each midpoint from the mean, $(c_i - \bar{x})^2$, where $\bar{x}$ is the mean calculated using midpoints.
3. Weight the squared deviations by the frequency of each class, n_i, and sum them up to get the total squared deviation.

Formula for Coefficient of Variation:

$$C_v(x) = \frac{\sqrt{\frac{\sum n_i(c_i-\bar{x})^2}{n}}}{\frac{\sum n_i c_i}{n}} = \frac{\sigma_x}{\overline{x}}.$$

Where:

- σ_x is the standard deviation of x.
- $\bar{x}$ is the mean of the midpoints.

Example 3.2.2 *Applied Measures of Dispersion in Ungrouped Data*
Consider a dataset of test scores: $5, 7, 8, 9, 10, 12$.

Step 1: Calculate the mean

$$\bar{x} = \frac{5+7+8+9+10+12}{6} = \frac{51}{6} = 8.5$$

Step 2: Calculate the Squared Deviations from the Mean

$$(5-8.5)^2 = 12.25, \quad (7-8.5)^2 = 2.25, \quad (8-8.5)^2 = 0.25, \quad (9-8.5)^2 = 0.25$$

$$(10-8.5)^2 = 2.25, \quad (12-8.5)^2 = 12.25$$

Step 3: Calculate the Variance

$$Var(x) = \frac{12.25+2.25+0.25+0.25+2.25+12.25}{6} = \frac{29.5}{6} = 4.91$$

Step 4: Calculate the Standard Deviation

$$\sigma = \sqrt{4.91} \approx 2.22$$

Step 4: Calculate the Coefficient of Variation

$$C_v(x) = \frac{\sigma_x}{\overline{x}} = \frac{2.22}{8.5} = O.261.$$

Example 3.2.3 ***Applied Measures of Dispersion in Grouped Data***
Consider the following grouped data for the heights of students:

Height Range (cm)	$[150-160[$	$[160-170[$	$[170-180[$	$[180-190[$
Frequency(n_i)	3	5	7	2
Midpoint(c_i)	155	165	175	185

Step 1: Calculate the mean
First, calculate the sum of $n_i c_i$:

$$\sum n_i c_i = 3 \times 155 + 5 \times 165 + 7 \times 175 + 2 \times 185 = 465 + 825 + 1225 + 370 = 2885$$

Now, compute the total frequency n:

$$n = 3 + 5 + 7 + 2 = 17$$

The mean is:

$$\bar{x} = \frac{2885}{17} \approx 169.12$$

Step 2: Calculate the Squared Deviations from the Mean for Each Class
$(155 - 169.12)^2 = 198.29, \quad (165 - 169.12)^2 = 16.97, \quad (175 - 169.12)^2 = 34.80,$

$(185 - 169.12)^2 = 253.62$
Step 3: Calculate the Weighted Sum of Squared Deviations

$$\sum n_i (c_i - \bar{x})^2 = 3 \times 198.29 + 5 \times 16.97 + 7 \times 34.80 + 2 \times 253.62 = 1430.56$$

Step 4: Calculate the variation and standard deviation
Variance:

$$Var(x) = \frac{1430.56}{17} \approx 84.06$$

Standard Deviation:

$$\sigma = \sqrt{84.06} \approx 9.17$$

Coefficient of Variation:

$$C_v(x) = \frac{7.9}{169.12} = 0.047.$$

3.3 Measures of Position

Measures of position are used to describe the relative position of a particular data point within a dataset. These measures help us understand how a data point compares to the rest of the data. Common measures of position include percentiles, quartiles, and deciles. Below is a detailed breakdown of these measures for both ungrouped and grouped data.

3.3.1 Quartiles

Quartiles For Ungrouped Data

Quartiles divide a dataset into four equal parts. The three quartiles are:

- Q_1 (First quartile): 25th percentile, divides the lower 25% of the data.
- Q_2 (Second quartile): 50th percentile, also the median.
- Q_3 (Third quartile): 75th percentile, divides the top 25% of the data.

To find the quartiles, we use the formula of (Nearest Rank method) for finding the k-th quartile is:

$$X_{Q_k} = \frac{k}{4} \times n$$

Example 3.3.1 *Consider the following dataset of test scores:*

$$55, 60, 65, 70, 75, 80, 85, 90, 95, 100$$

Now, let's find the quartiles:

- Q_1 (First quartile): The median of the lower half of the data ({55, 60, 65, 70, 75}), which is 65.
- Q_2 (Second quartile): The median of the entire data, which is 75.
- Q_3 (Third quartile): The median of the upper half of the data ({80, 85, 90, 95, 100}), which is 90.

So, the quartiles are:

$$Q_1 = 65, \quad Q_2 = 75, \quad Q_3 = 90$$

Quartiles for Grouped Data

Once you have identified the class containing the quartile, use linear interpolation to calculate the exact quartile value within the class. The interpolation formula is:

$$X_{Q_k} = a + \frac{n_{Q_k} - N_{Q_{k-1}}}{N_{Q_k} - N_{Q_{k-1}}} \times l$$

Where:

- Q is the quartile value.
- a is the lower boundary of the quartile class.

- N_Q is the cumulative frequency of the class containing the quartile.
- N_{Q-1} is the cumulative frequency before the class containing the quartile.
- n_{Q_k} is the frequency of the quartile class.
- l is the class width.

Example 3.3.2

Suppose you want to calculate the 1^{st} quartile for the following grouped data:

Class Interval	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$\sum$
Frequency	5	8	12	15	40

The Cumulative Frequency Table:

Class Interval	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$\sum$
Frequency	5	8	12	15	40
Cumulative Frequency	5	13	25	40	–

The class which contains 1^{st} quartile is the class of position:

$$\frac{n}{4} = 10$$

Thus, the class of position 10 is $[10-20[$

Now apply the interpolation formula:

$$X_{Q_1} = 10 + \frac{(10-5)}{(13-5)} \times 10$$

$$X_{Q_1} = 10 + \frac{5}{8} \times 10 = 10 + 6.25 = 16.25$$

Thus, the 1st quartile is approximately 10.4.

3.3.2 Deciles

Deciles for Ungrouped Data

Deciles divide the data into ten equal parts, each containing 10% of the data. The k-th decile corresponds to the $k \times 10$-th decile.

The formula for finding the k-th decile is:

$$X_{D_k} = \frac{k}{10} \times n$$

Where:

- D_k is the k-th decile.
- k is the decile number (e.g., D_1 is the first decile, D_2 is the second decile, and so on).
- n is the number of data points.

Example 3.3.3 *Given the dataset:*

$$55, 60, 65, 70, 75, 80, 85, 90, 95, 100$$

The formula for calculating the deciles is:

$$D_k = \frac{k(n)}{10}$$

where:

- k is the decile number ($k = 1, 2, 3, ..., 9$)
- n is the number of data points ($n = 10$)

Step 1: Compute the Positions Since $n = 10$, the formula simplifies to:

$$D_k = \frac{k(10)}{10} = k$$

Thus, the decile positions correspond to the k-th data point in the ordered list.
Step 2: Extract Decile Values
From the sorted dataset:

$$55, 60, 65, 70, 75, 80, 85, 90, 95, 100$$

$$\begin{aligned} D_1 &= \text{1st value} = 55 \\ D_2 &= \text{2nd value} = 60 \\ D_3 &= \text{3rd value} = 65 \\ D_4 &= \text{4th value} = 70 \\ D_5 &= \text{5th value} = 75 \\ D_6 &= \text{6th value} = 80 \\ D_7 &= \text{7th value} = 85 \\ D_8 &= \text{8th value} = 90 \\ D_9 &= \text{9th value} = 95 \end{aligned}$$

Final Decile Values

$$D_1 = 55, \quad D_2 = 60, \quad D_3 = 65, \quad D_4 = 70, \quad D_5 = 75$$

$$D_6 = 80, \quad D_7 = 85, \quad D_8 = 90, \quad D_9 = 95$$

Since the dataset has exactly 10 values, each decile aligns perfectly with an existing data point without requiring interpolation.

Deciles for Grouped Data

Deciles are found using a similar approach to quartiles. Once you have identified the class containing the quartile, use linear interpolation to calculate the exact decile value within the class. The interpolation formula is:

$$X_{D_k} = a + \frac{n_{D_k} - N_{D_{k-1}}}{N_{D_k} - N_{D_{k-1}}} \times l$$

Where:

- D is the decile value.
- a is the lower boundary of the decile class.
- N_D is the cumulative frequency of the class containing the decile.
- N_{D-1} is the cumulative frequency before the class containing the decile.
- n_{D_k} is the frequency of the decile class.
- l is the class width.

Example 3.3.4

Suppose you want to calculate the 9^{th} decile for the following grouped data:

Class Interval	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$\sum$
Frequency	5	8	12	15	40

The Cumulative Frequency Table:

Class Interval	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$\sum$
Frequency	5	8	12	15	40
Cumulative Frequency	5	13	25	40	–

The class which contains 9^{th} decile is the class of position:

$$\frac{9n}{10} = 36$$

Thus, the class of position 36 is $[30-40[$

Now apply the interpolation formula:

$$X_{D_9} = 30 + \frac{(36-25)}{(40-25)} \times 10$$

$$X_{D_9} = 30 + \frac{11}{25} \times 10 = 30 + 7.33 = 37.33$$

Thus, the 9th decile is approximately 37.3.

3.3.3 Percentiles

Percentiles For Ungrouped Data

A percentile is a measure that indicates the relative rank of a particular value in a data set. The k-th percentile is the value below which k% of the data fall. For example, the 50th percentile is the median, as it divides the data into two equal halves.

The formula of (Nearest Rank method) for finding the k-th percentile is:

$$X_{P_k} = \frac{k}{100} \times n$$

Where:

- P_k is the k-th percentile.

- k is the percentile (e.g., 25th, 50th, 75th).
- n is the number of data points.

Example 3.3.5 *Consider the following data set representing the scores of 10 students on a test:*

$$56, 61, 65, 70, 73, 75, 80, 82, 85, 90$$

Step 1: Sort the data in ascending order (if not already sorted).

$$56, 61, 65, 70, 73, 75, 80, 82, 85, 90$$

Step 2: Apply the formula for the 25th percentile.
For the 25th percentile, $k = 25$ and $n = 10$ (the number of data points).

$$X_{P_{25}} = \frac{25}{100} \times 10 = 2.5$$

Step 3: Find the position of the percentile. Thus, the 25th percentile of the data set is 63.

Percentiles for Grouped Data

Once you have identified the class containing the percentile, use linear interpolation to calculate the exact percentile value within the class. The interpolation formula is:

$$X_{P_k} = a + \frac{n_{P_k} - N_{P_{k-1}}}{N_{P_k} - N_{P_{k-1}}} \times l$$

Where:

- P is the percentile value.
- a is the lower boundary of the percentile class.
- N_p is the cumulative frequency of the class containing the percentile.
- N_{p-1} is the cumulative frequency before the class containing the percentile.
- n_{P_k} is the frequency of the percentile class.
- l is the class width.

Example 3.3.6

Suppose you want to calculate the 60th percentile for the following grouped data:

Class Interval	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$\sum$
Frequency	5	8	12	15	40

The Cumulative Frequency Table:

Class Interval	$[0-10[$	$[10-20[$	$[20-30[$	$[30-40[$	$\sum$
Frequency	5	8	12	15	40
Cumulative Frequency	5	13	25	40	–

The class which contains 60^{th} percentile is the class of position:

$$\frac{60n}{100} = 24$$

Thus, the class of position 24 is $[20 - 30[$

Now apply the interpolation formula:

$$X_{P_{60}} = 20 + \frac{(24 - 13)}{(25 - 13)} \times 10$$

$$X_{P_{60}} = 20 + \frac{11}{12} \times 10 = 20 + 9.17 = 29.17$$

Thus, the 60th percentile is approximately 29.17.

3.4 Exercises

Exercise 9 *We observe 90 times the number of arrivals (variable X) of clients at a post office during a time interval (10 minutes), and we get the following values:*

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

1) Construct the statistical table for the distribution of variable X.

2) Determine the mode, the arithmetic mean, and the median.

3) Determine the first and third quartiles Q_1 and Q_3.

Exercise 10 *Given the data set:*

$$23, 23, 23, 34, 39, 39, 39, 39, 39, 39, 39, 39, 45, 45, 45,$$
$$45, 48, 48, 52, 52, 52, 52, 52, 55, 62, 62, 62, 62, 62, 62.$$

1) Give the distribution of the partial frequencies (calculate the frequencies and increasing cumulative frequencies).

2) Determine the mode, the median, and the quartiles for this data set.

3) Calculate the third decile D_3 and the 90th percentile C_{90}.

Exercise 11 *At an Algeria Telecom agency, we recorded the amounts of phone bills issued on a specific day. The results are shown in the table below:*

X	$[500, 800[$	$[800, 1100[$	$[1100, 1400[$	$[1400, 1700[$	$[1700, 2000[$	$[2000, 2300[$
n_i	3	10	27	12	26	22

1) Determine the modal class, then calculate the mode.

2) Find the median of this data set.

3) Calculate the 75th percentile C_{75}, the arithmetic mean $\overline{x}$, the variance $Var(x)$.

Exercise 12 *We weighed the olives harvested from 150 farms. The following table represents the results, expressed in tons:*

Weight (tons)	*Frequency* n_i
$[5.00, 5.01[$	*4*
$[5.01, 5.02[$	*18*
$[5.02, 5.03[$	*25*
$[5.03, 5.04[$	*36*
$[5.04, 5.05[$	*30*
$[5.05, 5.06[$	*22*
$[5.06, 5.07[$	*11*
$[5.07, 5.08[$	*3*
$[5.08, 5.09[$	*1*

1) Identify the statistical series x presented above (the population, the character studied, and its nature).

2) Consider a new statistical variable $y = \frac{(x-5.045)}{0.01}$, calculate the arithmetic mean and variance of y.

3) Deduce the mean and variance of x.

4) Calculate the coefficient of variation of y, and deduce the coefficient of variation of x.

CHAPTER 4

Combinatorial Analysis (Counting)

In this chapter, we explore the fundamental principles of combinatorial analysis, focusing on the methods used to count and arrange objects. Combinatorics is essential for solving problems related to counting permutations, combinations, and selections in various contexts. We will examine key concepts such as the counting principle, factorials, binomial coefficients, and the inclusion-exclusion principle.

4.1 Random Experiment

Definition 4.1.1 *Random experiments are those that lead to random outcomes when the experiment is repeated under the same conditions.*

- Outcomes cannot be predicted in advance.
- Results are directly dependent on chance.

Example 4.1.2 *Coin toss, dice roll.*

FIGURE 4.1: Random experiment of tossing coin

FIGURE 4.2: Random experiment of rolling die

4.2 Probability Theory

Probability theory does not allow predicting which result will occur, but it determines the chance each result (outcome) has of occurring.

Definition 4.2.1 (Probability) *is the study of chance and uncertainty.*

Probability theory contains three essential elements:

1) The Universe.

2) Events.

3) Probability Measure.

4.2.1 The Universe

Definition 4.2.2 *It is the set of all possible outcomes in a random experiment. Also known as a fundamental set of a random experiment, it is denoted by Ω with ω_i representing elements of Ω.*

Example 4.2.3

1) A die is rolled, so the possible results are the 6 elements $1, 2, 3, 4, ..., 6$; in this experiment, a fundamental set is:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

2) A coin and a die are tossed simultaneously (at the same time). A fundamental set in this case consists of the set of ordered pairs:

$$\begin{aligned}\Omega &= \left\{(x, y) \,/\, x \in \{Heads, Tails\}, y = \overline{1, 6}\right\} \\ &= \{(Heads, 1), (Heads, 2), (Heads, 3), (Heads, 4)\,(Heads, 5)\,(Heads, 6) \\ &\quad (Tails, 1), (Tails, 2), (Tails, 3), (Tails, 4)\,(Tails, 5)\,(Tails, 6)\}.\end{aligned}$$

3) A coin is tossed twice in succession. A fundamental set in this case consists of a set of sequences:

$$\begin{aligned}\Omega &= \{(xy) \,/\, x \in \{Heads, Tails\}, y \in \{Heads, Tails\}\} \\ &= \{Heads\ Heads, Heads\ Tails, Tails\ Heads, Tails\ Tails\}\end{aligned}$$

4.2.2 Event

Definition 4.2.4 *An event is any subset A_i of a fundamental set Ω that has a given property.*

Example 4.2.5

1) A die is rolled.

A: "obtaining an even number."
B: "obtaining an odd number."
C: "obtaining a number less than 6."
Then:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\begin{aligned}A &= \{2, 4, 6\} \\ B &= \{1, 3, 5\} \\ C &= \{1, 2, 3, 4, 5\}\end{aligned}$$

2) Toss two coins simultaneously,

$$\Omega = \{(Heads, Heads), (Heads, Tails), (Tails, Heads), (Tails, Tails)\}$$

If the event $A = \{(\text{Tails}, \text{Tails}), (\text{Tails}, \text{Heads})\}$, then A is defined as follows:
A: "The first coin shows Tails."

3) Tossing two dice, then

$$\begin{aligned}\Omega &= \left\{(x, y) / x, y = \overline{1, 6}\right\} \\ &= \{(1,1), (1,2), ..., (6,6)\}\end{aligned}$$

such that Card $(\Omega) = 6 \times 6 = 36$.

If the event $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$ then
A: "The sum of the results of the two dice is equal to 7."

Realization of an Event

Let A be a given event; if the outcome of the experiment belongs to the set of possible outcomes defined by A, we say that an event A is realized. In other words, the event A has "occurred" because the actual result of the experiment matches one of the outcomes included in A. Combinatorial analysis is the counting of arrangements or groupings that can be formed from the elements of a finite set.

4.3 Disposition

A disposition is the set formed by choosing elements from n elements of a finite set.

Example 4.3.1 *Let $E = \{1, 2, 3\}$. How many dispositions of two elements can be constructed from this set?*

$$\underbrace{\boxed{?}}_{3}\ \underbrace{\boxed{?}}_{3}$$

$$= 3 \times 3$$
$$= 9\text{dispositions}$$

- We can construct 9 numbers from the digits 1, 2, and 3, which are:

$$11, 12, 21, 13, 31, 22, 23, 32, 33$$

4.3.1 Types of dispositions

Disposition without repetition

This is an disposition where an element can appear 0 or 1 time.

Example 4.3.2 *Let the set $E = \{1, 2, 3\}$, how many dispositions of two elements can be made from this set without repetition?*

In this case, the possible dispositions are:

$$12, 21, 13, 31, 23, 32$$

.

Disposition with repetition

This is an diposition where an element can appear more than once.

Example 4.3.3 *Let the set $E = \{1, 2, 3\}$, how many dispositions of two elements can be made from this set?*

In this case, the possible dispositions are:

$$11, 12, 21, 13, 31, 22, 23, 32, 33$$

.

Ordered disposition

In an ordered disposition, the order of elements matters.

Example 4.3.4 *Let the set $E = \{1, 2, 3\}$, how many ordered dispositions of two elements can be created from this set?*

In this case, the possible ordered dispositions are:

$$11, 12, 21, 13, 31, 22, 23, 32, 33$$

.

Unordered disposition

In an unordered disposition, the order of the elements does not matter.

Example 4.3.5 *suppose we have 10 people. How many ways can we form a group of 6 people?*

In this case, there are 210.

4.4 Classical combinatorial formulas

Multiplets

Consider an ordered disposition of λ elements $(x_1, x_2, ..., x_\lambda)$ where:
The 1^{st} element comes from set A, with $card\,(A) = \alpha$
The 2^{nd} element comes from set B, with $card\,(B) = \beta$
.
.
.
The λ^{th} element comes from set S, with $card\,(S) = \gamma$
The number of possible ordered dispositions (or sequences) is given by: $\alpha \times \beta \times \cdots \times \gamma$.

Example 4.4.1 *See the practical examples provided in the exercises.*

4.5 Arrangement

4.5.1 Arrangement without Repetition

An arrangement is an **ordered** and without repetition **disposition**; it is a way of choosing p $(p \leq n)$ elements from n, arranging them in a particular order. The number of arrangements is denoted A_n^p and is given by:

$$A_n^p = \frac{n!}{(n-p)!}$$

Example 4.5.1 *How many different ways can we elect a president and a vice-president from 10 people?*

$$\underbrace{\boxed{?}}_{10}\underbrace{\boxed{?}}_{9}$$
$$= 10 \times 9$$
$$= A_{10}^2 = \frac{10!}{8!}$$
$$= 90.$$

So, there are 90 different ways to elect a president and a vice-president from 10 people.

4.5.2 Arrangement with Repetition

An arrangement is ordered and with repetitions, that is, it is a way of choosing p elements from n elements with the possibility of selecting an element more than once. The number of arrangements is denoted A_n^p and is given by::

$$A_n^p = n^p$$

.

Example 4.5.2 *How many subsets of two letters can be formed from the set $\{A, B, C, D\}$?*

$$\underbrace{\boxed{?}}_{4}\underbrace{\boxed{?}}_{4}$$
$$= 4 \times 4$$

For each letter, there are two possibilities: either it is included, or it is not included. This allows us to form $\mathcal{A}_4^2 = 4^2 = 16$ subsets.

4.6 Permutation

When $p = n$, we obtain a special case of arrangements called **permutations**.

4.6.1 Permutation without Repetition

This is an ordered arrangement without repetition of these n elements, that is, it is a way to arrange n distinct elements in a sequence. The number of permutations of n elements is denoted P_n, and is given by:

$$P_n = n!$$

.

Example 4.6.1 *How many different ways are there to arrange 3 students?*

$$\underbrace{\boxed{?}}_{3}\ \underbrace{\boxed{?}}_{2}\ \underbrace{\boxed{?}}_{1}$$

$$= 3 \times 2 \times 1$$

also, you can use the formula for permutations without repetition: $P_3 = 3! = 6$. Thus, there are 6 different ways to arrange 3 students.

4.6.2 Permutation with Repetition

A permutation with repetition involves arranging n elements where:
n_1 are identical,
n_2 are identical,
.
.
.
n_r are identical,, with $n_1 + n_2 + \cdots + n_r = n$

$$P_n^{(n_1,n_2,\ldots,n_r)} = \frac{n!}{n_1!n_2 \ldots n_k!}$$

.

Example 4.6.2 *How many ways can we arrange* 2 *red books,* 5 *green books, and* 1 *white book on a shelf? (Only the color differentiates the books!)*

To determine how many ways we can arrange 2 red books, 5 green books, and 1 white book on a shelf (where only the color differentiates the books), we use the formula for permutations with repetition:

- We notice that $2 + 5 + 1 = 8$

$$P_8^{(2,5,1)} = \frac{8!}{2!5!1!} = 168$$

We can arrange these 8 books in 168 different ways.

4.7 Combination

4.7.1 Combination without repetition

This is an **unordered** and **without repetition** arrangement, meaning it is a way of choosing p elements from n elements, where each element can be chosen only once, without arranging them in a particular order.

The number of combinations without repetition is denoted as C_n^p and given by:

$$C_n^p = \frac{A_n^p}{p!} = \frac{n!}{(n-p)!p!}$$

Example 4.7.1 *We draw* 5 *cards from a deck of* 32 *cards. How many possible results are there??*

To find the number of possible results when drawing 5 cards from a deck of 32 cards, you use the formula for combinations without repetition:

$$C_{32}^5 = \frac{32!}{27!5!} = 201376$$

So, there are 201376 possible results when drawing 5 cards from a deck of 32 cards.

4.7.2 Combination with repetition

This is an **unordered** and **with repetition** disposition, meaning it is a way to choose p elements from n elements with the possibility of selecting an element multiple times.

The number of combinations with repetition is denoted $\mathcal{C}_n^p$ and is given by:

$$\mathcal{C}_n^p = C_{n+p-1}^p = \frac{(n+p-1)!}{(n-1)!p!}$$

Example 4.7.2 *In a dominoes game, there are 7 possible values* $\{blank, 1, 2, 3, 4, 5, 6\}$*, Each domino has 2 values on it, and the order does not matter (flipping the domino does not change what it represents).*

- The total number of dominoes we can form is choosing $p = 2$ elements from $n = 7$ values, which is calculated as:

$$C_{7+2-1}^2 = \frac{8!}{6!2!} = 28 \text{ dominoes}$$

Property 4.7.3 *For any p elements chosen from n, the following properties hold::*

1) $\forall n \geq 1, C_n^0 = C_n^n = 1$ and $C_n^1 = C_n^{n-1} = n$.

2) $\forall n \geq 0, \forall 0 \leq p \leq n, C_n^p = C_n^{n-p}$.

3) $\forall n \geq 0, \forall 0 \leq p \leq n-1, C_n^p = C_{n-1}^{p-1} + C_{n-1}^p$.

Pascal's Triangle

Definition 4.7.4 *is a triangular array of numbers where each entry is the sum of the two numbers directly above it. Named after the French mathematician Blaise Pascal, it has applications in algebra, probability, and number theory. Here's how it's constructed:*

- The top row of Pascal's Triangle contains a single 1.
- Each subsequent row starts and ends with 1, and every other number is the sum of the two numbers directly above it.

Example 4.7.5 *The last property allows us to construct the following* ***Pascal's triangle****:*

	0	1	2	3	4	⋯
0	1					
1	1	1				
2	1	2	1			
3	1	3	3	1		
4	1	4	6	4	1	
⋮						

Newton's Binomial

Is a method to expand any power of a binomial expression, i.e., an expression of the form $(a+b)^n$. The general formula for expanding these expressions into a sum involving terms with binomial coefficients.

$$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k} = \sum_{k=0}^{n} C_n^k a^{n-k} b^k.$$

Where:

- $C_n^k$is the binomial coefficient, also known as "n choose k," and is given by:

  $$C_n^k = \frac{n!}{(n-k)!k!}$$

  It represents the number of ways to choose k elements from a set of n elements.
- $a^k b^{n-k}$ represents the powers of a and b in each term.
- The index k starts from 0 and goes up to n.

Example 4.7.6

1) $(1+x)^n = \sum_{k=0}^{n} C_n^k x^k 1^{n-k} = C_n^0 x^0 1^n + C_n^1 x^1 1^{n-1} + C_n^2 x^2 1^{n-2} + C_n^3 x^3 1^{n-3} + ... + C_n^n x^n 1^0$.

2) $(x+y)^2 = \sum_{k=0}^{n} C_n^k x^k y^{n-k} = C_2^0 x^0 y^2 + C_n^1 x^1 y^1 + C_n^2 x^2 y^0$.

4.8 Exercises:

Exercise 13 *A coin and a die are tossed, give the sample space.*

1) Express the following events explicitly:

$A =$ Heads and an even number appear,
$B =$ A prime number appears,
$C =$ Tails and an odd number appear,

2) Express explicitly the event:

a) A or B occurs,

b) B and C occur,

3) Which of the events A, B and C are mutually exclusive?

Exercise 14 *Starting from the word "Barika"*

1) How many four-letter words can be formed?

2) What is the minimum number of distinct four-letter words that can be formed?

Exercise 15 *What is the total number of possible codes consisting of* 4 *symbols, where:*

- *The first two symbols are alphabet letters.*
- *The last two symbols are digits.*

Exercise 16 *A phone number consists of 5 digits. It must start with* 0*, the second digit is between* 1 *and* 5*, and the other digits are free.*

How many different phone numbers can be formed based on these conditions?

Exercise 17 *When rolling 12 dice successively, the outcome is called an ordered sequence of 12 numbers.*

How many possible sequences exist where the face 1 appears 5 times, the face 2 appears 3 times, the face 3 appears 3 times, and the face 4 appears once,

Exercise 18 *Knowing that Algerian landline phone numbers contain* 9 *digits:*

1) How many different phone numbers can be constructed?

2) How many phone numbers can be constructed specifically for the commune of Barika, knowing that the first three digits are fixed as 0, 3, and 3?

Chapter 5

Probability and Conditional Probability

This chapter introduces the concept of a probability space, the mathematical framework used to define probabilities. A probability space consists of three key elements: the sample space, events, and a probability function that assigns likelihoods to events. Understanding these components is essential for calculating probabilities in a structured and consistent way.

The chapter also explores conditional probability, which focuses on the probability of an event occurring given that another event has already taken place. This concept is critical for analyzing dependent events and forms the basis for many advanced statistical methods. Through examples and exercises, students will learn how to define probability spaces and apply conditional probability

5.1 Probability space

5.1.1 Set of Subsets of Ω

Let Ω be the universe of a random experiment. $\mathcal{P}(\Omega)$ denotes the set of all possible subsets of Ω.

$$A \in \mathcal{P}(\Omega) \Leftrightarrow A \subset \Omega$$

Example 5.1.1 *Let* $\Omega = \{a, b, c\}$ *then,*

$$\mathcal{P}(\Omega) = \left\{ \emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \underbrace{\{a, b, c\}}_{\Omega} \right\},$$

If $\Omega = \{1, 2, 3, 4, 5, 6\}$, then

$$\mathcal{P}(\Omega) = \left\{ \begin{array}{c} \emptyset, \{1\}, \{2\}, ..., \{6\}, \{1, 2\}, ..., \{1, 6\}, ..., \{6, 6\}, ..., \{1, 2, 3\}, ..., \\ \{2, 3, 4\}, ..., \{1, 2, 3, 4\}, ..., \{1, 2, 3, 4, 5\}, ..., \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\Omega} \end{array} \right\}.$$

5.1.2 Set of Events $(\mathcal{F})$

A $(\sigma - \text{algebra})$ on Ω is any subset $\mathcal{F}$ of $\mathcal{P}(\Omega)$ that satisfies the following properties:

1) $\Omega \in \mathcal{F}$.

2) $A \in \mathcal{F} \Rightarrow \overline{A} \in \mathcal{F}$.

3) For any sequence $(A_i)_{i \geq 0}$ of elements in $\mathcal{F}$, $\bigcup_{i=0}^{\infty} A_i \in \mathcal{F}$.

5.1.3 Probability measure:

A probability $\mathbf{P}$ on Ω is a function (mapping) from $\mathcal{F}$ to the interval $[0, 1]$:

$$\mathbf{P} : \mathcal{F} \longrightarrow [0, 1]$$

Satisfying the following properties:

1) For every event A in $\mathcal{F}$, $0 \leq \mathbf{P}(A) \leq 1$.

2) $\mathbf{P}(\Omega) = 1$.

3) Let $(A_i)_{i \in \mathbb{N}}$ be a sequence of mutually disjoint (incompatible, i.e., $A_i \cap A_j = \emptyset$ and $i \neq j$), then $\mathbf{P}\left(\bigcup_{i \in \mathbb{N}} A_i \right) = \sum_{i \in \mathbb{N}} \mathbf{P}(A_i)$.

5.1.4 General Properties of a Probability

Any probability **P** in$(\Omega, \mathcal{F})$ space satisfies the following propositions:

1) $P(\emptyset) = 0$.

2) $\forall A \in \mathcal{F}$ then $P\left(\overline{A}\right) = 1 - P(A)$

3) $\forall A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4) If $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$

5) $\forall A, B \in \mathcal{F}$, if $A \subset B \Rightarrow P(A) \leq P(B)$

5.1.5 Probability of an Event

The probability of an event $A \in \mathcal{P}(\Omega)$, is a measure of how likely the event is to occur within the sample space Ω, It is denoted as $P(A)$, and is defined as:

$$\begin{aligned} P(A) &= \frac{\text{card } (A)}{\text{card } (\Omega)} \\ &= \frac{\text{Number of favorable outcomes}}{\text{Number of possible outcomes}} \end{aligned}$$

Here's a breakdown of the formula:

- Number of favorable outcomes: The count of outcomes in which the event A occurs.
- Number of possible outcomes: The total number of outcomes in the sample space Ω.

Example 5.1.2 *If you roll a fair six-sided die, the sample space* Ω *is* $\{1, 2, 3, 4, 5, 6\}$. *If the event* A *is rolling a number greater than 4, then the favorable outcomes are* $\{5, 6\}$. *Thus,*

$$P(A) = \frac{2}{6} = \frac{1}{3}$$

This fraction represents the probability of event A occurring.

5.1.6 Probability Space

A probability space is a triplet $(\Omega, \mathcal{F}, \mathbf{P})$ where:

1) Ω is the sample space (universe).

2) $\mathcal{F}$ is a σ-algebra on Ω (the set of events)

3) **P** is a probability measure.

Example 5.1.3 *Consider rolling two dice successively.*

The sample space is:

$$\begin{aligned} \Omega &= \{(1,1), (1,2), (1,3), \ldots, (1,6), (2,1), (2,2), (2,3), ..., (2,6), .., (6,6)\} \\ &= \{(i,j), i = 1, ..., 6; j = 1, ..., 6\} \end{aligned}$$

$$\text{card}\,(\Omega) = 6^2 = 36$$

The σ-algebra associated with this experiment: $\mathcal{F} = \mathcal{P}(\Omega)$
Let A be the event "the sum of the dice is $i + j = 4$", then:

$$A = \{(2,2), (3,1), (1,3)\}$$

and

$$P(A) = \frac{\text{card}\ (A)}{\text{card}\ (\Omega)} = \frac{3}{36}.$$

5.2 Conditional Probability and Independence

5.2.1 Conditional Probability

Conditional probability is the probability of an event occurring given that another event has already occurred. It's a way to calculate the likelihood of an event based on the knowledge that a related event has taken place.

Definition 5.2.1

Let B be an event with non-zero probability $P(B) \succ 0$ The conditional probability of an event A given that event B has occurred, is denoted as $P(A/B)$ or $P_B(A)$ such that:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Remark 5.2.2 *The events A and B are said to be related.*

Example 5.2.3 *We roll two fair dice. Then, the sample space Ω is:*

$$\begin{aligned}\Omega &= \{1,2,,3,4,5,6\} \times \{1,2,3,4,5,6\} \\ &= \{(1,1), (1,2), \ldots, (6,6)\} \\ &= \{(i,j), i = 1,...,6; j = 1,...,6\}\end{aligned}$$

$$\text{card}\ (\Omega) = 6^2 = 36$$

Let the following events be:
A :" The first result is greater than or equal to 4".
B :"$i = j$".
Then,
$A = \{(5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$.
$B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$
$A \cap B = \{(4,4), (5,5), (6,6)\}$

$$\begin{aligned}P(B) &= \frac{\text{card}\ (B)}{\text{card}\ (\Omega)} \\ &= \frac{6}{36} = \frac{1}{6}\end{aligned}$$

and

$$P(A \cap B) = \frac{\text{card } (A \cap B)}{\text{card } (\Omega)} = \frac{3}{36}$$

so,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{3/36}{6/36} = \frac{3}{6} = \frac{1}{2}$$

Properties 5.2.4 *Let $P_B : \mathcal{P}(\Omega) \to [0,1]$ such that P_B is a conditional probability.*

1) $P_B(\Omega) = 1$..

2) If $(A_i)_{i=1}^n \in \mathcal{A}$, $\forall i,j = 1,...,n$, $i \neq j : A_i \cap A_j = \emptyset$ (pairwise disjoint), then: $P_B\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P_B(A_i)$

5.2.2 The Independence of Events

Independence is a probabilistic concept that intuitively qualifies random events as having no influence on each other.

Definition 5.2.5

Two events A and B are said to be independent if and only if:

$$P(A \cap B) = P(A) \times P(B).$$

Remark 5.2.6

Dependence: $P_B(A) = \frac{P(A \cap B)}{P(B)}$

Independence: $P_B(A) = P(A)$ because $P(A \cap B) = P(A) \times P(B)$

Remark 5.2.7 *It is important to distinguish between independent events and incompatible events.*

Proposition 5.2.8

If A and B are two events and P is a probability, then the following propositions are equivalent:

1) A and B are independent.

2) A and $\overline{B}$ are independent.

3) $\overline{A}$ and B are independent.

4) $\overline{A}$ and $\overline{B}$ are independent.

Remark 5.2.9

If three events A, B et C are pairwise independent, this does not imply that:

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$$

5.2.3 Formula of Compound Probabilities

Theorem 5.2.10 *Let $A_1, A_2, ..., A_n$ be events, then:*

$$\begin{aligned}\mathbf{P}(A_1 \cap A_2 \cap ... \cap A_n) = \mathbf{P}(A_1) \times \mathbf{P}(A_2/A_1) \times \mathbf{P}(A_3/A_1 \cap A_2) \\ \times \cdots \times \mathbf{P}(A_n/A_1 \cap A_2 \cap ... \cap A_{n-1}).\end{aligned}$$

Example 5.2.11 *Among* 10 *mechanical pieces,* 4 *are defective. Two pieces are successively drawn at random from the set (without replacement). What is the probability that both pieces are correct?*

Method 1

Let: A_1 :" the first piece is not defective"
A_2 :" the second piece is not defective"then,

$$\begin{aligned}\mathbf{P}(A_1 \cap A_2) &= \mathbf{P}(A_1) \times \mathbf{P}(A_2/A_1) \\ &= \frac{3}{5} \times \frac{5}{9} \\ &= \frac{1}{3}\end{aligned}$$

Method 2

$\Omega = \{$ represents the set of all possible sequences of 2 pieces from the set of 10 pieces$\}$
$A = \{$ represents the sequences of 2 pieces from the set of6 correct pieces $\}$ In this case, the order is important because the pieces are drawn successively, without repetition, since the drawing is done without replacement. The first piece is drawn without being put back into the lot, and the same applies to the second piece, which leads to an arrangement without repetition.

$$P(A) = \frac{\text{card } (A)}{\text{card } (\Omega)} = \frac{A_6^2}{A_{10}^2} = \frac{\frac{6!}{4!}}{\frac{10!}{8!}} = \frac{6!8!}{10!4!} = \frac{1}{3}$$

$$P(B) = \frac{\text{card } (B)}{\text{card } (\Omega)} = \frac{2}{4} = \frac{1}{2}.$$

5.2.4 Bayes Formula:

Bayes formula (also known as Bayes Theorem) describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It is given as following definition.

Definition 5.2.12

Let $(\Omega, \mathcal{A}, P)$ be a probability space associated with a given experiment. Let $(A_i)_{1\leq i\leq p}$ be a partition of Ω and assume that $\forall i \in [1,p], P(A_i) \neq 0$ and $B \subset \Omega\ P(B) \neq 0$, Given the probabilities $P(B/A_i)$ andés $P(A_i)$ Bayes' formula is:

$$P\left(A_j/B\right) = \frac{P\left(B/A_j\right) \times P\left(A_j\right)}{\sum_{i=1}^{p} P(B/A_i) \times P(A_i)}.$$

Remark 5.2.13

This formula is also called the probability of the cause, as we are looking for the event A_j which is the cause of the realization of B.

Example 5.2.14

Three workshops produce electronic components of the same kind:
In workshop A_1 1% of the pieces are defective.
In workshop A_2 10% of the pieces are defective.
In workshop A_3 3% of the pieces are defective.
A random piece is chosen, and it is defective. What is the probability that it comes from A_1, A_2 or A_3
Let
A_1 : " the piece comes from A_1".
A_2 : " the piece comes from A_2".
A_3 : " the piece comes from A_3".
$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$
Let B : " be the event that the piece is defective".
$P(B/A_1) = 0,01, P(B/A_2) = 0,1$ et $P(B/A_3) = 0,03$
The probability that the defective piece comes from A_1 :

$$\begin{aligned} P(A_1/B) &= \frac{P(B/A_1).P(A_1)}{P(B/A_1).P(A_1) + P(B/A_2).P(A_2) + P(B/A_3) \times P(A_3)} \\ &= \frac{0,01 \times 1/3}{0,01 \times 1/3 + 0,1 \times 1/3 + 0,03 \times 1/3} \\ &= \frac{1}{14}. \end{aligned}$$

The probability that the defective piece comes from A_2 :

$$\begin{aligned} P(A_2/B) &= \frac{P(B/A_2).P(A_2)}{P(B/A_1).P(A_1) + P(B/A_2).P(A_2) + P(B/A_3) \times P(A_3)} \\ &= \frac{0,1 \times 1/3}{0,01 \times 1/3 + 0,1 \times 1/3 + 0,03 \times 1/3} \\ &= \frac{10}{14}. \end{aligned}$$

The probability that the defective piece comes from A_3 :

$$\begin{aligned} P(A_3/B) &= \frac{P\left(B/A_3\right).P\left(A_3\right)}{P\left(B/A_1\right).P\left(A_1\right)+P\left(B/A_2\right).P\left(A_2\right)+P\left(B/A_3\right)\times P\left(A_3\right)} \\ &= \frac{0,03\times 1/3}{0,01\times 1/3+0,1\times 1/3+0,03\times 1/3} \\ &= \frac{3}{14}. \end{aligned}$$

5.3 Exercises

Exercise 19

An urn contains one black ball and one white ball. Balls are drawn (one at a time) until the black ball appears. Each time the white ball is drawn, it is returned to the urn along with two additional white balls.

What is the probability that the process:

a) Stops after five draws?

b) Does not stop after four draws?

Exercise 20

Two urns, U_1 and U_2, contain 70% and 80% white balls, respectively. Urn U_1 contains three times as many balls as urn U_2.

All the balls are placed into a third urn, and a ball is drawn at random from this urn. If the ball is white, what is the probability that it came from urn U_1?

Exercise 21

In a workshop, there are three batches of parts with defect rates of 5%, 8%, and 10%, respectively.

A part is drawn from one of the batches, chosen at random, and it is found to be defective.

What is the probability that this batch contains 5% defective parts?

Exercise 22

Consider three urns U_1, U_2 and U_3. The first urn U_1 contains 9 white balls, 4 red balls, and 2 black balls. The second urn contains 8 white balls, 5 red balls, and 2 black balls. The third urn contains 6 white balls, 4 red balls, and 5 black balls. One of the three urns is chosen at random, and then three balls are drawn simultaneously from that urn.

Given that the three drawn balls are: "2 white and 1 red".

1) Calculate the probability that they came from urn U_2.

2) Calculate the probability that they did not come from urn U_2.

Given that the three drawn balls are: "1 white, 1 red, and 1 black".

1) Calculate the probability that they came from urn U_1.

2) Calculate the probability that they came from neither urn U_1 nor urn U_2.

Bibliography

[1] Theophilos Cacoullos. *Exercises in probability*. Springer Science & Business Media, 2012.

[2] George Casella and Roger Berger. *Statistical inference*. CRC press, 2024.

[3] Jay L Devore, Kenneth N Berk, and Matthew A Carlton. *Modern mathematical statistics with applications*. Springer Nature, 2021.

[4] David Forsyth. *Probability and statistics for computer science*. Vol. 13. Springer, 2018.

[5] Narayan C Giri. *Introduction to probability and statistics*. CRC Press, 2019.

[6] Zealure Holcomb. *Fundamentals of descriptive statistics*. Routledge, 2016.

[7] Jean-Pierre Lecoutre. *Statistique et probabilités*. Dunod, 2002.

[8] Philippe Tassi and Sylvia Legait. *Théorie des probabilités en vue des applications statistiques*. Editions Technip, 1990.