PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

IBN KHALDOUN UNIVERSITY OF TIARET
FACULTY OF NATURE AND LIFE SCIENCES

Field: NATURAL AND LIFE SCIENCES
Specialization: BIOLOGICAL SCIENCES
Level: 2ND YEAR LICENSE

Pedagogical Handbook

Biostatistics

Presented by:

Mr. DELLAL Mohamed, MCA

Academic Year: 2024/2025

Contents

Preface

This handbook on **Biostatistics**, intended for students in **Bachelor's Degree in Biological Sciences** at Ibn Khaldoun University of Tiaret, has been designed to provide a clear and methodical introduction to the fundamental concepts of statistics applied to the natural and life sciences.

In a constantly evolving scientific context, biostatistics proves to be indispensable for the analysis and interpretation of biological data. This course aims to equip students with the necessary tools to understand and apply the main statistical techniques in their future research and practical work.

The content is structured around several chapters, covering essential themes such as:

- **Descriptive statistics**, useful for summarizing and interpreting data;
- **Probabilities** and statistical distributions, which model uncertainty;
- **Hypothesis testing** and **estimation**, fundamental in scientific decision-making;
- **ANOVA** and variance analysis for group comparison.

Each chapter is enriched with practical exercises to promote a better understanding of the concepts discussed and to encourage the independent application of the studied techniques.

We hope that this handbook will be a valuable resource throughout your academic journey and that it will contribute to the development of your skills in biological data analysis.

Dr. DELLAL Mohamed
Academic Year 2024/2025

Descriptive Statistics 1

Definition 1.0.1 *Statistics is the study of observable variations. It is a scientific method that involves gathering numerical data on large sets, then analyzing and interpreting it.*

- The set on which a statistical study is conducted is called the **population**.For example, the set of students in a lecture hall, the set of inhabitants of Algeria, etc.
- The elements of the population are called **individuals** or **statistical units**.
- The property studied for each individual is called **characteristics**. The population is studied according to one or more **characteristics**.
- The different values taken by the characteristic are called **modalities**.
- When the modalities are only numerical values, the characteristic is said to be **quantitative** (salary, height, temperature, etc.). There are two cases:
  1. The characteristic is said to be **discrete quantitative** if it can only take whole number values. For example, we can study the number of children, which can only take the modalities 0, 1, 2, 3, etc.
  2. The characteristic is said to be **continuous quantitative** if the values are taken from an interval of $\mathbb{R}$. These intervals are called **classes**.
- In cases where the modalities are not numerical values, the characteristic is said to be **qualitative**. For example, studying eye color or the sport practiced.
- A **statistical series** corresponds to the different modalities of a characteristic on a sample of individuals belonging to a given population.

- The number of individuals in the studied sample is called the **size** of the sample.

1.1 Statistical Distribution of a Variable

All data from the statistical series are grouped into a table indicating the distribution of individuals according to the studied characteristic.

1.1.1 Discrete Quantitative Variable

Definition 1.1.1 *Assume that the values of the characteristic are* $\{x_1, x_2, ..., x_p\}$*, let* n_i *denote the number of individuals for which the observed characteristic modality is* x_i*,* $i = \overline{1, p}$*. We then say that* $(x_i, n_i)_{i=1}^{p}$ *is a discrete statistical series.*

In general, each value x_i of a discrete quantitative variable corresponds to a number, denoted by n_i; this is in fact the number of individuals for whom the value x_i has been observed. The frequency f_i of the value x_i is calculated using the formula:

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

where n_i denotes the frequency corresponding to the value x_i and N the total frequency

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

We always have:

$$0 \leq f_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{i=p} f_i = 1.$$

Cumulative Frequencies and Cumulative Relative Frequencies

It can be useful to supplement the statistical table by adding either cumulative frequencies or cumulative relative frequencies. These quantities are defined as follows:

$$N_k = \sum_{i=1}^{i=k} n_i \qquad F_k = \sum_{i=1}^{i=k} f_i.$$

Score x_i	Frequency n_i	Relative frequency f_i	Cumulative frequency N_k	Cumulative relative frequency F_K
7	2	2/50	2	2/50
8	3	3/50	5	5/50
9	6	6/50	11	11/50
10	4	4/50	15	15/50
11	8	8/50	23	23/50
12	5	5/50	28	28/50
13	6	6/50	34	34/50
14	5	5/50	39	39/50
15	4	4/50	43	43/50
16	3	3/50	46	46/50
17	2	2/50	48	48/50
18	2	2/50	50	1

Table 1.1 – *Table of statistical data including observed values, frequencies, cumulative frequencies, and corresponding relative frequencies.*

In other words, N_i represents the number of observations less than or equal to x_i and F_i represents their frequency (or their percentage if we consider 100 F_i).

Example 1.1.1 *The list of scores obtained during a test by 50 candidates is:*

8	10	9	14	7	13	11	15	17	14
13	13	15	16	9	11	10	11	12	15
11	13	9	16	10	15	7	11	14	18
12	15	14	8	17	18	9	11	10	13
9	14	12	9	11	11	8	12	13	16

Table 1.1 displays the calculated frequencies, cumulative frequencies, and relative frequencies based on the data from Example 1.1.1. Readers are advised to reproduce these calculations to solidify their understanding of the process.

1.1.2 Continuous Quantitative Variable

Definition 1.1.2 *Let us assume that the values of the variable are grouped into intervals* $[a_0, a_1[, ..., [a_{p-1}, a_p[$*, We denote by* n_i *the number of individuals whose observed value of the variable falls within the interval* $[a_{i-1}, a_i[$*,* $i = \overline{1, p}$*. We then say that* $([a_{i-1}, a_i[, n_i)_{i=1}^{p}$ *is a continuous statistical series.*

The infinite observable values of a continuous quantitative variable make it impossible to generalize the bar chart. Establishing a frequency distribution table requires dividing the variable's range into p subintervals. $[a_0, a_1[, [a_1, a_2[, ..., [a_{p-1}, a_p[$.

- Each of these intervals is called a **class**.
- The quantities a_{i-1} and a_i are called the **boundaries** of the i^{th} class.
- $x_i = \frac{a_i + a_{i-1}}{2}$ is the **center** of the i^{th} class.
- $a_i - a_{i-1}$ is the **amplitude** of this class.

We continue to use a statistical table similar to the one seen in the previous section, but now with the classes arranged in ascending order in the first column. The concepts of frequency, relative frequency, cumulative frequency, and cumulative relative frequency are defined in the same way as for discrete data.

Example 1.1.2 *We consider the age of a population of 1000 people. To simplify data entry, we define 4 age groups:* $[0, 20[, [20, 40[, [40, 60[, [60, 80]$. *The distribution according to their age is:*

Age range	$[0, 20[$	$[20, 40[$	$[40, 60[$	$[60, 80]$
Number of people	*360*	*380*	*160*	*100*

1.2 Common Graphical Representations

For a discrete variable, there are essentially two types of graphical representations which are, in fact, complementary: the bar chart and the cumulative (step) chart.

1.2.1 The Bar Chart

Let $(x_i, n_i)_{i=1}^{p}$ be a discrete statistical series. The values $x_1, x_2, ..., x_p$ are plotted on the (ox) axis. To create the bar chart, from each x_i, a "bar" segment is drawn perpendicular to the (ox) axis with a length of f_i or n_i.

Example 1.2.1 *Figure 1.1 shows the bar chart of the data from Example 1.1.1. This chart includes a horizontal axis (the abscissa), which displays the observations of the*

variable considered (here, the scores), and a vertical axis (the ordinate), which shows the counts.

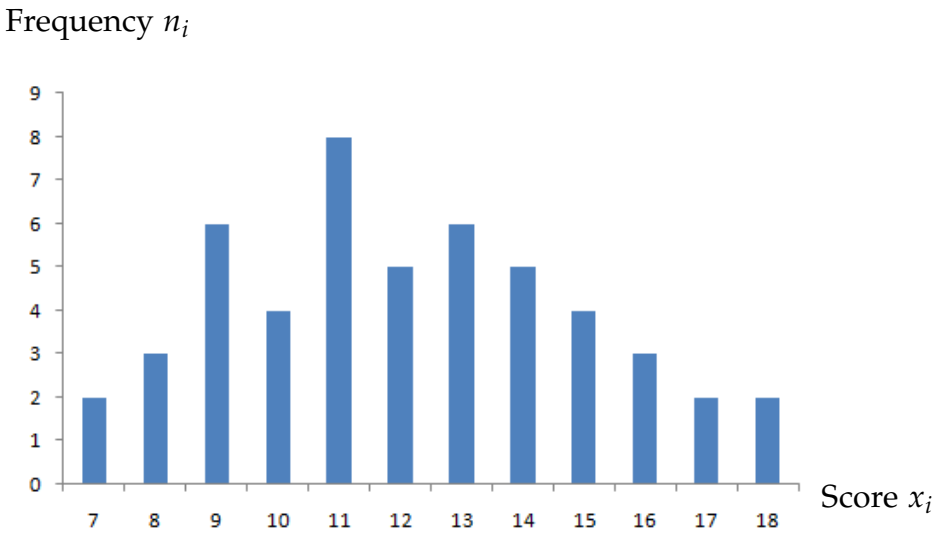


Figure 1.1 – *The Bar Chart.*

1.2.2 Cumulative Distribution Function

Let $(x_i, n_i)_{i=1}^{p}$ be a discrete statistical series. The values $x_1, x_2, ..., x_p$ are plotted on the (ox) axis according to their magnitudes. The cumulative distribution function is defined by:

$$F(x) = \begin{cases} 0 & \text{if} \quad x \in]-\infty, x_1[\\ F_i & \text{if} \quad x \in [x_i, x_{i+1}[, \ i = \overline{1, p-1} \\ 1 & \text{if} \quad [x_p, +\infty[\end{cases}$$

The cumulative curve (cumulative diagram) is the representative curve of the function F. This curve is used to visualize cumulative counts, or cumulative frequencies or percentages. It thus allows for the simple determination of the number or proportion of observations that are less than or equal to a given value in the series.

Example 1.2.2 *Figure 1.2 shows the cumulative curve related to Example 1.1.1. It illustrates what's known as a step function. On the x-axis, once again, are the observa-*

tions of the variable in question, while the y-axis now displays the cumulative counts, cumulative frequencies, or cumulative percentages.

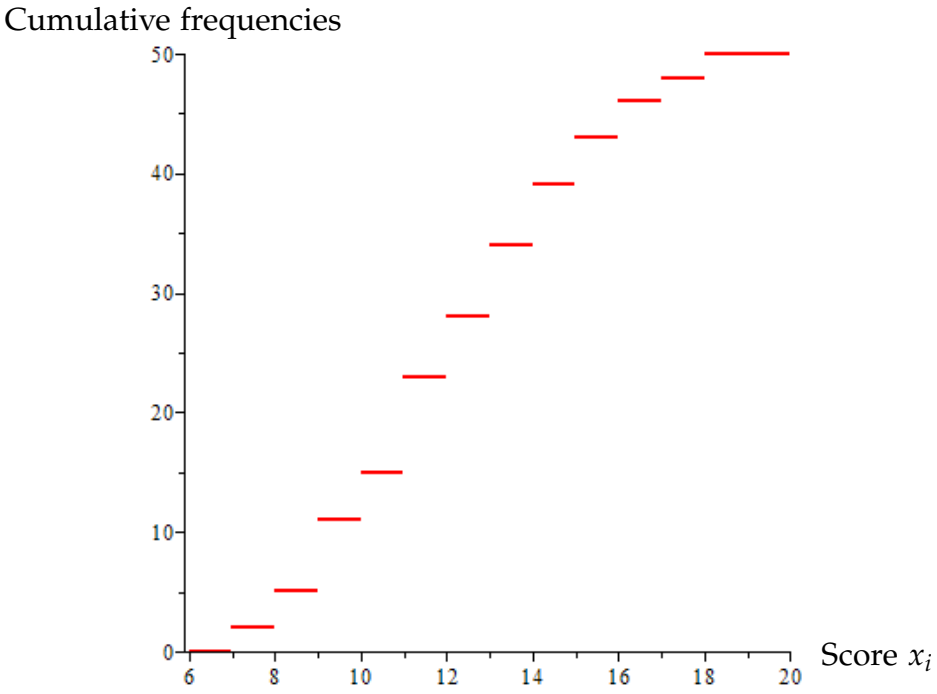


Figure 1.2 – *Cumulative Distribution Function.*

For a continuous quantitative variable, the two common graphs replacing the bar chart and the cumulative diagram are the histogram and the cumulative curve.

1.2.3 The Histogram

Let $([a_{i-1}, a_i[, n_i)_{i=1}^{p}$ be a continuous statistical series. A histogram consists of rectangles with a base of $[a_{i-1}, a_i[$ and the height of each rectangle is proportional to f_i or n_i. Each class is then represented by a rectangle whose base is delimited by the corresponding limits and whose height is what is called the the frequency density (or percentage). In the case the intervals of the classes are not equal, we make the following modification: the height is the ratio of the relative frequencies to the amplitude of the class

$$h_i = \frac{f_i}{a_i - a_{i-1}},\ i = 1, 2, ..., p$$

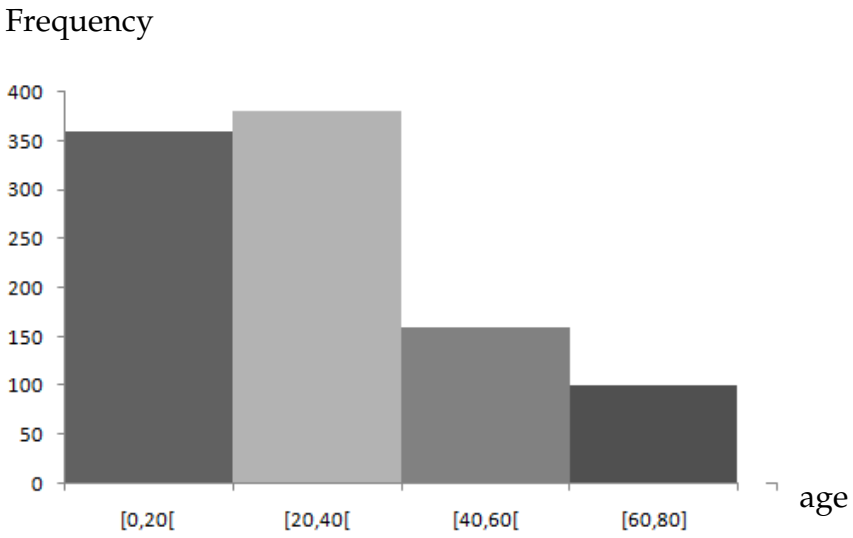


Figure 1.3 – *Histogram of frequencies.*

Example 1.2.3 *Figure 1.3 gives the histogram corresponding to the data of Example 1.1.2. We see that this is what is called a staircase function. On the abscissa are, once again, the observations of the variable considered, while on the ordinate are now the cumulative numbers, the cumulative frequencies or the cumulative percentages.*

1.2.4 The cumulative curve

Let $([a_{i-1}, a_i[, n_i)_{i=1}^p$ be a continuous statistical series. $(F_i)_{i=1}^p$ be the cumulative relative frequencies. To obtain the cumulative curve, we set $F_0 = 0$ and connect the points with coordinates (a_i, F_i), $i = 0, 1, ..., p$ by straight line segments. Each class considered must first be represented by a single point whose abscissa is the upper bound of the class a_i and whose ordinate is the cumulative relative frequencies of this class F_i. The cumulative curve is then the curve joining the points in question. It therefore represents the evolution of the cumulative frequencies (or Frequency, or percentages), as the cumulative diagram did in the discrete case.

Example 1.2.4 *Figure 1.4 gives the cumulative curve relative to Example 1.1.2.*

Cumulative relative frequencies

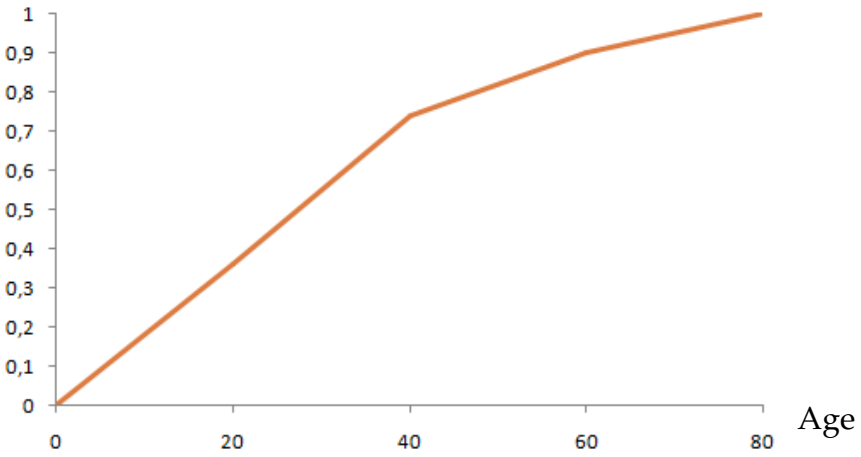


Figure 1.4 – *The cumulative curve*

1.3 Statistical Position Parameters

1.3.1 Arithmetic Mean

Definition 1.3.1

- *The mean of a discrete statistical series* $(x_i, n_i)_{i=1}^{p}$*, denoted* $\bar{x}$*, is the number*

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$$

where p *denotes the number of distinct values of* X*.*

- *The mean of a continuous statistical series* $([a_{i-1}, a_i[, n_i)_{i=1}^{p}$*, denoted* $\bar{x}$*, is the real*

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$$

where x_i *is the center of the class* $[a_{i-1}, a_i[$ *and* p *the number of classes.*

- *The formulation* $\sum_{i=1}^{p} f_i x_i$ *is called the* ***weighted arithmetic mean*** *of* X*, because each distinct value of* X *is weighted by its corresponding frequency.*

Proposition 1.3.1 *If the mean of the discrete statistical series* $(x_i, n_i)_{i=1}^{p}$, *is* $\bar{x}$ *then the mean of the discrete statistical series* $(\alpha x_i + \beta, n_i)_{i=1}^{p}$, *is* $\alpha\bar{x} + \beta$.

Property 1.3.1 *The sum of deviations from the mean is zero.*

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

1.3.2 The Mode

Definition 1.3.2 *Let* $(x_i, n_i)_{i=1}^{p}$ *be a discrete statistical series. The* ***mode****, denoted by* M_o, *is the value of the statistical variable for which the count (or frequency) is the greatest. That is, if k satisfies* $n_k = \max(n_1, n_2, ..., n_p)$, *then* $M_o = x_k$.

Example 1.3.1 *In the case of Example 1.1.1, the mode is the most frequent note:*

$$M_o = 11.$$

Definition 1.3.3 *Let* $([a_{i-1}, a_i[, n_i)_{i=1}^{p}$ *be a continuous statistical series and let* $n_k = \max(n_1, n_2, ..., n_p)$. *We then say that* $[a_{k-1}, a_k[$ *is* ***the modal class*** *(the class with the highest frequency) and we have:*

$$M_o = a_{k-1} + \frac{d\Delta_i}{\Delta_i + \Delta_s}$$

a_{k-1}*: the lower bound of the modal class.*
$d = a_k - a_{k-1}$*: the amplitude of the modal class.*
Δ_i*: difference in frequency between the modal class and the lower class.*
Δ_s*: difference in frequency between the modal class and the upper class.*

Example 1.3.2 *In the case of Example 1.1.2, the modal class is* $[20, 40[$. *We will then apply this formula using the frequencies and the class frequencies*

$$M_o = a_{k-1} + \frac{d\Delta_i}{\Delta_i + \Delta_s} = 20 + \frac{20(380 - 360)}{(380 - 360) + (380 - 160)} = 21.66$$

Remark 1.3.1 *The mode and the modal class are not necessarily unique.*

1.3.3 La médiane

Definition 1.3.4 *Let $(x_i, n_i)_{i=1}^{p}$ be a discrete statistical series. The **median** (denoted M_e) is the value of the statistical variable that divides the population into two equal parts:*

- *If N is odd: $N = 2k+1$ then the median is the $(\frac{N+1}{2})$-th value*

$$M_e = x_{k+1}.$$

- *If N is even: $N = 2k$ then the median is the mean of the two central values*

$$M_e = \frac{x_k + x_{k+1}}{2}.$$

Example 1.3.3 *Following Example 1.1.1, since $N = 50$ is even, then the median is the mean of the two central values: $M_e = \frac{x_{25} + x_{26}}{2} = \frac{12+12}{2} = 12$*

Definition 1.3.5 *Let $([a_{i-1}, a_i[, n_i)_{i=1}^{p}$ be a continuous statistical series. We seek to determine **the median class** $[a_{k-1}, a_k[$, the class containing the $(\frac{N}{2})$-th value, and we have*

$$M_e = a_{k-1} + d\frac{N/2 - N_k}{n_k}$$

a_{k-1}: the lower bound of the median class.
$d = a_k - a_{k-1}$: the amplitude of the median class.
N_k: cumulative frequency up to a_{k-1}.
n_k: frequency of the median class.
N: total frequency.

Example 1.3.4 *Following Example 1.1.2, let's use the cumulative frequency line to determine the median: there are 1000 people, 50% of the total headcount is 500, the median here is the age corresponding to the cumulative frequency 500.*

The median is therefore in the median class $[20, 40[$

$$M_e = a_{k-1} + d\frac{N/2 - N_k}{n_k} = 20 + 20\frac{500 - 360}{380} = 27.36$$

Remark 1.3.2 *The median M_e can be determined graphically as the x-coordinate of the point on the cumulative curve where the y-coordinate is equal to $\frac{1}{2}$.*

1.3.4 Quantile

The quantile of order α of a quantitative variable X is the value x_α of this variable that divides the studied population into two sub-populations, with the respective frequencies equal to α and $1-\alpha$ of the total population. When X is continuous, x_α can be determined by the equation:

$$F(x_\alpha) = \alpha.$$

The most commonly used quantiles correspond to specific values of α.

- The quartiles of X are its three quantiles $x_{1/4}$, $x_{2/4}$, $x_{3/4}$.
- $Q_1 = x_{1/4}$, is called the **first quartile**; one-fourth of the values taken by X are less than or equal to Q_1.
- $Q_2 = x_{1/2} = M_e$, is the median.
- $Q_3 = x_{3/4}$, is called the **third quartile**; one-fourth of the values taken by X are greater than or equal to Q_3.

1.4 Shape Parameters

In statistics, shape parameters are crucial for understanding the characteristics of a data distribution. These include measures such as skewness and kurtosis, which provide insights into the asymmetry and peakedness of the distribution. This section elaborates on these parameters with definitions, formulas, interpretations, and examples.

1.4.1 Skewness (Coefficient of Skewness)

Skewness is a measure of the asymmetry of a distribution. It quantifies how much a distribution deviates from being symmetric around the mean. Skewness can be:

- **Zero (Symmetric)**: The distribution has no asymmetry, resembling a normal or bell-shaped distribution.
- **Positive (Right-skewed)**: The tail on the right side is longer or fatter, indicating more high values in the data.
- **Negative (Left-skewed)**: The tail on the left side is longer or fatter, indicating more low values in the data.

The skewness coefficient S for a sample of n observations $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$ and standard deviation s is calculated by:

$$S = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Interpretation of Skewness:

- $S = 0$: The distribution is symmetric (e.g., normal distribution).
- $S > 0$: Right-skewed distribution, where values above the mean are more dispersed.
- $S < 0$: Left-skewed distribution, where values below the mean are more dispersed.

Example of Skewness: Consider the following two datasets:

- **Dataset A (Symmetric)**: $5, 6, 7, 8, 9$
- **Dataset B (Right-skewed)**: $2, 4, 6, 8, 20$

Calculating the skewness for each:

- **Dataset A**: $S \approx 0$ (symmetric)
- **Dataset B**: $S > 0$ (positively skewed)

This example demonstrates how skewness indicates the direction of asymmetry.

1.4.2 Kurtosis (Coefficient of Kurtosis)

Kurtosis is a measure of the "tailedness" or "peakedness" of a distribution, assessing the tendency of a distribution to produce outliers. Kurtosis compares the shape of the distribution to that of a normal distribution in terms of its peak and tails.

The kurtosis K for a sample is given by:

$$K = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4$$

We usually subtract 3 from the result to obtain the "excess kurtosis," which helps compare the shape to a normal distribution:

$$\text{Excess Kurtosis} = K - 3$$

The three types of kurtosis are:

- **Mesokurtic** ($K = 3$): The distribution has a similar kurtosis to a normal distribution.
- **Leptokurtic** ($K > 3$): The distribution has a sharper peak and fatter tails than a normal distribution, indicating more frequent extreme values.
- **Platykurtic** ($K < 3$): The distribution has a flatter peak and thinner tails, with values clustered around the mean.

Example of Kurtosis: Consider three datasets:

- **Dataset C (Mesokurtic)**: Normal distribution with values clustered around the mean.
- **Dataset D (Leptokurtic)**: More extreme values, leading to a high kurtosis.
- **Dataset E (Platykurtic)**: Few extreme values, with data concentrated around the mean.

In each case, the kurtosis indicates the shape of the distribution's tails and peak relative to a normal distribution.

1.4.3 Detailed Examples

To better understand skewness and kurtosis, consider the following example.

Example Dataset: The daily returns of a stock over a period are:

$$1.2, 0.5, -0.3, 0.7, 1.8, -2.0, 0.3, -0.4, 2.5, -1.2$$

Step 1: Calculate the Mean and Standard Deviation.

$$\bar{x} = \frac{1.2 + 0.5 + (-0.3) + 0.7 + 1.8 + (-2.0) + 0.3 + (-0.4) + 2.5 + (-1.2)}{10} = 0.29$$

$$s = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2} = 1.22$$

Step 2: Calculate the Skewness.

$$S = \frac{1}{10}\sum_{i=1}^{10}\left(\frac{x_i - \bar{x}}{s}\right)^3 = 0.52$$

Since $S > 0$, this dataset is positively skewed.

Step 3: Calculate the Kurtosis.

$$K = \frac{1}{10}\sum_{i=1}^{10}\left(\frac{x_i - \bar{x}}{s}\right)^4 = 2.89$$

Excess kurtosis:

$$K - 3 = -0.11$$

This indicates a slightly platykurtic distribution, as $K < 3$.

1.4.4 Interpretation of Shape Parameters

- **Skewness**: In this example, the positive skewness suggests a right-skewed distribution, meaning there are some higher values that stretch the tail to the right.
- **Kurtosis**: The negative excess kurtosis value suggests a distribution with

slightly thinner tails than a normal distribution, meaning fewer extreme values than expected.

1.4.5 Importance of Shape Parameters in Data Analysis

Shape parameters like skewness and kurtosis are essential in many statistical applications:

- **Identifying the Distribution Type**: Skewness and kurtosis help identify if the data follows a normal distribution or if it has asymmetry or heavy/light tails.
- **Choosing Appropriate Statistical Tests**: Many statistical tests assume normality. High skewness or kurtosis may necessitate non-parametric tests or transformations.
- **Outlier Detection**: Kurtosis can signal the presence of outliers or extreme values that may need special attention.

Understanding these shape parameters can guide data analysis, ensuring appropriate statistical methods and interpretations.

1.5 Dispersion Parameters

1.5.1 Range

- The range of a discrete statistical series $(x_i, n_i)_{i=1}^{p}$, denoted E, is the number

$$E = x_p - x_1,$$

  i.e., the difference between the largest and smallest observed values.

- The range of a continuous statistical series $([a_{i-1}, a_i[, n_i)_{i=1}^{p}$ is the real number

$$E = a_p - a_0.$$

1.5.2 Interquartile Range

The interquartile range, denoted I, is the difference between the third quartile Q_3 and the first quartile Q_1; it is given by:

$$I = Q_3 - Q_1.$$

This range contains 50% of the population. This characteristic is valuable because it is entirely independent of extreme values (and thus very reliable).

1.5.3 Variance

Variance is, by definition, the average of the squares of deviations from the mean; it is a quadratic measure of dispersion around the mean.

$$V(x) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2,$$

where p denotes the number of distinct values of X if the statistical series is discrete or denotes the number of classes if the series is continuous. This formula can also be rewritten as:

$$V(x) = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2.$$

It can also be expressed as:

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

1.5.4 Standard Deviation

The standard deviation of the variable X, denoted by $\sigma(X)$, is, by definition, the square root of the variance of this variable:

$$\sigma(X) = \sqrt{V(X)}.$$

Note that the standard deviation is the most commonly used measure of dispersion.

- If σ is low, it indicates that the values are quite concentrated around the mean.
- If σ is high, it indicates that the values are more dispersed around the mean.

Example 1.5.1 *Using the example 1.1.2, we have:*

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \frac{1}{1000} \times 30000 = 30 \text{ years}.$$

Age Group	$[0,20[$	$[20,40[$	$[40,60[$	$[60,80]$	*Total*
Number of People	*360*	*380*	*160*	*100*	*1000*
Class Center x_i	*10*	*30*	*50*	*70*	/
$n_i.x_i$	*3600*	*11400*	*8000*	*7000*	*30000*
$n_i.x_i^2$	*36000*	*342000*	*400000*	*490000*	*1268000*

$$V(x) = \left(\frac{1}{N}\sum_{i=1}^{p} n_i x_i^2\right) - \bar{x}^2 = \frac{1268000}{1000} - (30)^2 = 368;$$

$$\sigma(X) = \sqrt{V(X)} = 19.18 \simeq 19 \text{ years}.$$

1.6 Exercises

Exercise 1

We measure the diameters of tree trunks of the same species. We study 400 specimens and obtain the following results:

Diameter in cm	25	26	27	28	29	30
Frequency n_i	40	60	120	140	20	20

1. What characteristic is being studied? Is it quantitative or qualitative?
2. Construct the table of frequencies f_i, and cumulative frequencies F_i.
3. What is the proportion of specimens with a diameter less than or equal to 27 cm?

4. What is the average diameter of these trunks? What is the median of this series?

5. Determine the variance, then the standard deviation, rounded to 0.01, of the statistical series.

Answer:

1. The variable: the diameter of the tree trunk. (it is a quantitative variable)

2. Frequency table (f_i) and cumulative frequencies (F_i):

Values x_i	25	26	27	28	29	30	**Total**
Frequencies n_i	40	60	120	140	20	20	400
Relative frequencies f_i	0.1	0.15	0.3	0.35	0.05	0.05	1
Cumulative frequencies F_i	0.1	0.25	0.55	0.9	0.95	1	//

3. According to the cumulative frequency table, the proportion of specimens with a diameter less than or equal to 27 cm is 0.55.

4. Mean diameter and Median:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{6} n_i x_i = \sum_{i=1}^{6} f_i x_i = 27.25$$

Thus, the average diameter of these tree trunks is 27.25 cm.

Median:

Total frequency $N = 400$ is even, so the median is the average of the two central values:

$$M_e = \frac{x_{200} + x_{201}}{2} = \frac{27 + 27}{2} = 27.$$

5. Variance and standard deviation:

$$V(x) = \left(\frac{1}{N}\sum_{i=1}^{6} n_i x_i^2\right) - \bar{x}^2 = 744.05 - 742.56 = 1.49.$$

We conclude that the standard deviation is $\sigma(X) = \sqrt{V(X)} = 1.22$

Exercise 2

The wingspan of 100 individuals of a butterfly species was measured, with the results grouped in the table:

Wingspan x_i	73	74	75	76	77	78	79	80
Frequency n_i	4	12	6	14	18	16	22	8
Frequency f_i								
Cumulative Frequency F_i								

1. Complete the table above with the frequency line f_i and the cumulative frequency line F_i.
2. What proportion of butterflies has a wingspan strictly less than or equal to 77 mm?
3. Calculate the average wingspan, the median, and the mode.
4. Determine the variance, then the standard deviation. Interpret the result.

Answer:

1. Frequency table (f_i) and cumulative frequencies (F_i):

Wingspan: x_i	73	74	75	76	77	78	79	80
Frequencies: n_i	4	12	6	14	18	16	22	8
Relative frequencies: f_i	0.04	0.12	0.06	0.14	0.18	0.16	0.22	0.08
Cumulative frequencies F_i	0.04	0.16	0.22	0.36	0.54	0.7	0.92	1
$n_i.x_i$	292	888	45	1064	1386	1248	1738	640
$n_i.x_i^2$	21316	65712	33750	80864	106722	97344	137302	51200

2. According to the cumulative frequency table, the proportion of butterflies with a wingspan strictly less than or equal to 77 mm is 0.54, representing 54% of the total, or $\frac{54 * 100}{100} = 54$ butterflies.

3. Average wingspan, median, and mode:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{8} n_i x_i = \frac{7706}{100} = 77.06$$

Thus, the average wingspan is 77.06 mm.

Median:

Total count $N = 100$ is even, so the median is the average of the two central values:

$$M_e = \frac{x_{50} + x_{51}}{2} = \frac{77 + 77}{2} = 77$$

Thus, the median of the data series is 77 mm.

Mode:

The mode of this series is the value of the variable corresponding to the maximum frequency. Here, it is 79 mm.

4. Variance and standard deviation:

$$V(x) = \left(\frac{1}{N}\sum_{i=1}^{8} n_i x_i^2\right) - \bar{x}^2 = \frac{594210}{100} - (77.06)^2 = 3.85$$

We conclude that the standard deviation is $\sigma(X) = \sqrt{V(X)} = 1.96$
Interpretation: Since the standard deviation is high, the values are dispersed around the mean.

Exercise 3

The bone density X was measured in 400 women with osteoporosis. The results are provided in the table below:

Bone Density	[40,50[	[50,60[	[60,70[	[70,80[	[80,90]	[90,100]
Number of Subjects	12	52	136	144	54	2

1. Represent this series with a histogram.

2. Determine the Mode and provide its graphical representation.
3. Calculate the average bone density.
4. Find the median value, explain its significance, and describe how to represent it graphically.

Answer:

1. Histogram:

A histogram consists of rectangles with a base $[a_{i-1}, a_i[$, and the height of each rectangle is proportional to f_i or n_i.

2. Mode:

First, we determine the modal class (the class with the highest frequency): $[70, 80[$

$$M_o = a_{k-1} + d\frac{\Delta_i}{\Delta_i + \Delta_s} = 70 + 10\frac{8}{8+90} = 70.81$$

3. Average bone density:

To determine the average bone density $\bar{x}$, we consider the midpoint x_i of each interval.

Bone density	[40,50[	[50,60[	[60,70[	[70,80[	[80,90]	[90,100]	**Total**
Number of subjects	12	52	136	144	54	2	400
Class midpoint x_i	45	55	65	75	85	95	//
$n_i.x_i$	540	2860	8840	10800	4590	190	27820

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{6} n_i x_i = \frac{27820}{400} = 69.55$$

4. Median:

First, we determine the median class (the class containing the $(\frac{N}{2})$-th value): $[60, 70[$

$$M_e = a_{k-1} + d\frac{N/2 - N_k}{n_k} = 60 + 10\frac{200 - 64}{136} = 70$$

The median is the value that divides the population into two equal parts; it is the x-coordinate of the point on the cumulative curve with an ordinate of 0.5.

Exercise 4

The temperature is recorded every hour over 4 days in a forest. The sorted results are gathered in the following table:

Temperature	14.5	15	15.5	16	16.5	17	17.5	18	18.5
Frequency	5	7	10	12	15	10	11	9	8

1. What is the studied statistical variable and its nature?
2. Calculate the mean of this series.
3. Determine the median and mode of this series.
4. Determine the variance, then the standard deviation. Interpret the result.

Exercise 5

The table below summarizes a study on height in cm:

Class	Frequency	Relative Frequency
[140 ;150[	1	0.10
[150;155[	α	β
[155;160[	2	0.20
[160;170[	4	0.40

For each statement, mark true or false:

1. The studied variable is discrete quantitative.
2. The total sample frequency cannot be determined.

3. The value of β is equal to 0.25.
4. The total sample frequency is equal to 10.
5. The median is located between 155 and 160.

Exercise 6

The following statistical table gives the total skull length of 60 mice.

Total Skull Length (mm)	22.5	23	23.5	24	24.5	25	25.5
Frequency	6	6	21	9	10	6	2

1. What is the studied statistical population? What is the studied variable and its nature?
2. Calculate the average total skull length and the median.
3. Determine the variance, then the standard deviation of the series.

Exercise 7

We consider the age of a population of 1000 people. To simplify data entry, we define 4 age groups: [0, 20[, [20, 40[, [40, 60[, [60, 80[. The distribution based on their age is

Age Group	[0, 20[	[20, 40[	[40, 60[	[60, 80[
Number of people n_i	360	380	160	100

1. Construct the table of relative frequencies (f_i) and cumulative frequencies (F_i).
2. What is the proportion of people who are at least 40 years old? Less than 60 years old? Between 20 and 60 years old?
3. Construct the frequency histogram.
4. Construct the cumulative frequency polygon (cumulative curve).

5. What is the proportion of people who are less than 30 years old (using graphical determination)?

6. Calculate the median and the mean.

Probabilities 2

2.1 Elements of Probability Calculus.

A random experiment is an experiment for which the outcome cannot be predicted before it is carried out. We denote

- ω a possible outcome of this random experiment.
- Ω The set or universe of all possible outcomes.
- *Event* A: any subset or part of the universe Ω.
- *Elementary event* $\{\omega\}$: subset of the universe consisting of a single element (in other words, a singleton).
- *Impossible event*: an event that contains none of the elements of Ω.
- *Certain event*: the set Ω of all possibilities.

Example 2.1.1 *If two coins are tossed, the possible outcomes are*

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$$

where T stands for "tails" and H for "heads." An event is a logical assertion about a random experiment, such as "getting heads twice" or "getting heads at least once."

- *The event "getting heads twice" is the subset* $\{(H,H)\}$.
- *The event "getting heads at least once" is the subset*

$$\{(H,H),(H,T),(T,H)\}$$

- *The set Ω is called the certain event, and the empty set $\varnothing$ is called the impossible event.*

2.2 Operations on Events.

Union: The event $A \cup B$ occurs as soon as either A or B occurs.

Intersection: The event $A \cap B$ occurs as soon as A and B both occur in the same experiment.

Difference: The event $A \setminus B$ occurs when A occurs and B does not.

Complement: The complement of A is the event $\Omega \setminus A$, denoted ${}^cA$.

2.3 Relationships between Events.

Mutually exclusive events: Events A and B are said to be mutually exclusive if and only if their intersection is the empty set: $A \cap B = \varnothing$

Inclusion: If A is included in B, we write $A \subset B$. We say that A implies B.

Example 2.3.1 *A die is rolled, with faces numbered from 1 to 6. Then*
- The set of all possible outcomes of the experiment is: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Event A: "obtaining an even number" is: $A = \{2, 4, 6\}$
- Event B: "obtaining a multiple of 3" is: $B = \{3, 6\}$
- Event C: "obtaining an odd number" is: $C = \{1, 3, 5\}$
- Event $A \cup B$: "obtaining an even number OR a multiple of 3" is:
$A \cup B = \{2, 3, 4, 6\}$
- Event $A \cap B$: "obtaining an even number AND a multiple of 3" is: $A \cap B = \{6\}$
- Complement of A: "obtaining an odd number" is: ${}^cA = \{1, 3, 5\} = C$
- Events A and B are mutually exclusive because $A \cap B = \varnothing$.

2.4 Probability

Definition 2.4.1 *A probability space (Ω, P) consists of*

1) Ω, a set (the set of outcomes of a random experiment)

2) P, a probability on Ω.

*An element $\omega \in \Omega$ is called an **outcome**, which is a possible result of a random experiment.*
*A subset $A \subset \Omega$ is called an **event**. It is a set of outcomes (for example, those that satisfy a certain condition). The set of events is thus the set $\mathcal{P}(\Omega)$ of parts (or subsets) of Ω.*

Definition 2.4.2 *A probability on Ω is a function $P : \mathcal{P}(\Omega) \longrightarrow [0, 1]$, defined on the events, such that:*

1) $P(\Omega) = 1$;

2) if A and B are two mutually exclusive events then

$$P(A \cup B) = P(A) + P(B)$$

Special Case (Uniform Probability).

If all elementary events $\omega_1, \omega_2, \cdots, \omega_n$ comprising Ω (with cardinality n) have the same probability p, we have: $P(\{\omega_i\}) = p$ for all $1 \leq i \leq n$, and furthermore:

$$P(\Omega) = P(\{\omega_1\} \cup ... \cup \{\omega_n\}) = P(\{\omega_1\}) + ... + P(\{\omega_n\}) = np$$

whence $p = \dfrac{1}{n}$. If A is an event, that is, a subset of Ω, then:

$$P(A) = \frac{card(A)}{card(\Omega)}$$

where $card(A)$ denotes the number of elementary events in A and $card(\Omega)$ that in Ω.

Properties of a Probability.

1) $P(\varnothing) = 0$;

2) For any event A, $P({}^cA) = 1 - P(A)$.

3) ***Monotonicity***: If $A \subset B$, then $P(A) \leq P(B)$ and $P(B \setminus A) = P(B) - P(A)$.

4) If $A_1, \ldots, A_n$ are mutually exclusive events, then

$$P(A_1 \cup A_2 \cup ... \cup A_n) = \sum_{k=1}^{n} P(A_k)$$

5) ***Inclusion-Exclusion Principle***: If A and B are two events then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

6) ***Conditional Probabilities***: The probability of A given that B occurs is defined by

$$P_B(A) = \frac{P(A \cap B)}{P(B)}, \quad \text{if} \quad P(B) > 0$$

7) ***Independence***: Two events A and B are said to be independent if and only if $P(A \cap B) = P(A)P(B)$. This is equivalent to saying that $P_B(A) = P(A)$ or $P_A(B) = P(B)$.

2.5 Concept of a Random Variable

2.5.1 Definition of a Random Variable

The concept of a random variable formalizes the association of a value to the outcome of a random experiment.

Definition 2.5.1 *A real random variable X is a function associating a real number to an elementary event*

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

Example 2.5.1 *We flip two coins and consider the random variable representing the number of heads obtained. Three outcomes are possible: 0, 1, 2*

$$X(\omega) = \begin{cases} 0 & \text{if} \quad \omega = (F, F) \\ 1 & \text{if} \quad \omega \in \{(P, F), (F, P)\} \\ 2 & \text{if} \quad \omega = (P, P) \end{cases}$$

2.5.2 Probability Distribution

Definition 2.5.2 *Let Ω be a sample space with a probability P, and let X be a random variable. The probability distribution of X, denoted P_X, is the function that associates to every subset A of $\mathbb{R}$*

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

For simplicity, we will write in the following course

$$P(\{\omega \in \Omega : X(\omega) \in A\}) = P(X \in A)$$

Similarly, we will denote

$$P(\{\omega \in \Omega : X(\omega) = x\}) = P(X = x)$$

P_X can also be seen as a probability on $X(\Omega)$, the set of values taken by X, also called the **support** of P_X. Sometimes we write $X \rightsquigarrow P_X$ to indicate that X follows the distribution P_X.

2.5.3 Cumulative Distribution Function

Definition 2.5.3 *The cumulative distribution function of the random variable X is defined by*

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}$$

Properties of F_X:

- $0 \leq F_X \leq 1$
- $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$
- F_X is always increasing and right-continuous.
- $P(a < X \leq b) = F_X(b) - F_X(a), \quad \forall a < b.$

2.5.4 Discrete Real Random Variables

Definition 2.5.4 *A real random variable X taking values in a finite or countable set $\mathcal{X}$ is called a discrete real random variable. In this case, the distribution of X is determined*

by the set of probabilities:

$$P_X(x) = P(X = x), \quad x \in \mathcal{X}.$$

Thus, for any subset A of $\mathcal{X}$, we then have:

$$P_X(A) = P(X \in A) = \sum_{x \in A} P(X = x)$$

*To characterize a discrete distribution, it is therefore sufficient to provide the **elementary probabilities** $P_X(x)$ for all $x \in X(\Omega)$.*

2.5.5 Continuous Real Random Variables

Continuous quantitative random variables take values in $\mathbb{R}$, or, more often in biology, within an interval included in $\mathbb{R}^+$ (biometric measurements, concentrations, etc.). The probability that a random variable X (for example, blood glucose level) takes the exact value $x = 0.3846 mg/L$ is almost zero. However, we can calculate the probability $F(x)$ that X is less than a certain value x:

$$F(x) = P(X \leq x)$$

which is called the **cumulative distribution function** of X.

Remark 2.5.1 *The probability associated with the event $X = a$ is zero, as it is impossible to observe this exact value. We then consider the probability that the random variable X takes values within an interval $[a, b]$. As this interval tends towards o, the value taken by X tends towards a function called the probability density function or probability density.*

2.5.6 Probability Density

The **density function**, denoted $f(x)$, is the derivative of F. Thus, the cumulative distribution function of the random variable X can be defined as the area under the curve f between $-\infty$ and x (Fig. 1). This is the probability that X is less than x:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

Note that the total area under the density is 1 (this is the sum of probabilities):

$$F(+\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

Similarly, the probability that X is between a and b ($b > a$) is

$$P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx.$$

The probability density f_X characterizes the distribution of a continuous random variable.

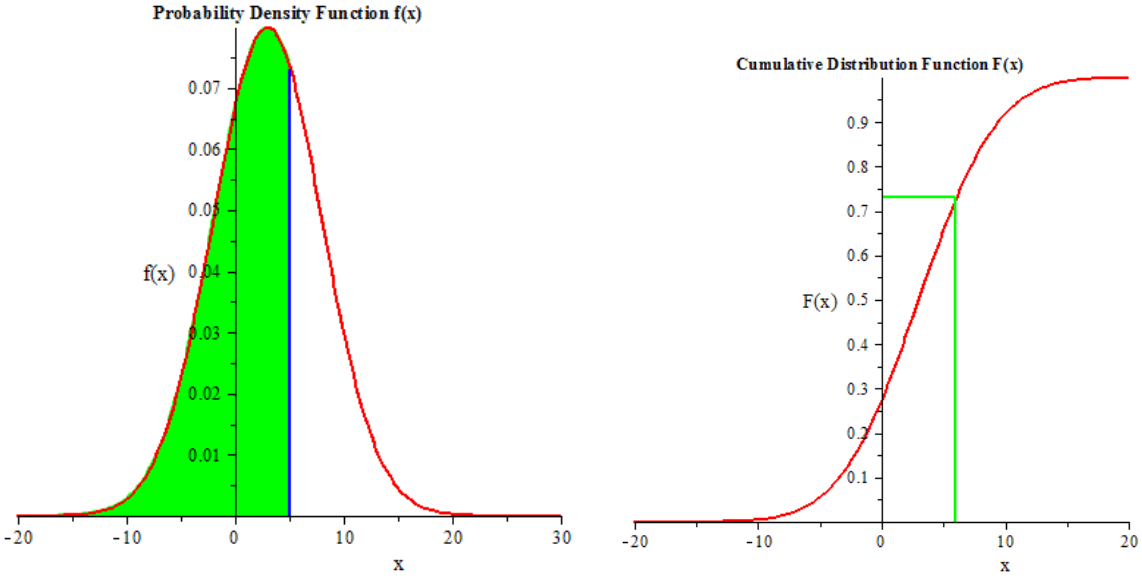


Figure 2.1 – *In this example, X is the random variable associated with a measurement of fasting blood glucose in humans. $F(6)$ is the probability that this measurement is less than* 0.6.

2.5.7 Quantile

For each value x, we can calculate the cumulative distribution function $F(x)$. Conversely, we may be interested in the value of x such that $F(x)$ has a certain

value. The quantile of order α, denoted q_α, is the value such that

$$F(q_\alpha) = \alpha$$

Commonly used quantiles for statistical tests are $q_{0.025}$, $q_{0.05}$, and $q_{0.95}$.

The median of a distribution is by definition the quantile at 50%: half of the population has a value less than the median.

2.6 Expectation of a Random Variable

Definition 2.6.1 *The **expectation** or **mean** of a random variable X, denoted $E[X]$, is the average of its values, weighted by their probabilities.*

- *If X is discrete,*
$$E[X] = \sum_{x \in X(\Omega)} xP(X = x).$$
- *If X is continuous, with density f,*
$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx.$$

The expectation is not always defined. It requires that the series or the integral above converges absolutely.

Properties

- The expectation is linear: for all random variables X and Y, and all real numbers a and b
$$E[aX + b] = aE[X] + b \text{ and } E[X + Y] = E[X] + E[Y]$$
- The expectation is increasing: if $X \leq Y$, then $E[X] \leq E[Y]$

Proposition 2.6.1 *Let X be a random variable, and $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a function.*

- *If X is discrete, then*
$$E[\varphi(X)] = \sum_{x \in X(\Omega)} \varphi(x)P(X = x).$$

- *If X is continuous, with density f, then*

$$E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) f(x) dx.$$

Provided that the series and the integral are well-defined.

2.7 Variance, Standard Deviation

Definition 2.7.1 *Let X be a random variable. The **variance** of X is the expectation of the squares of the deviations of X from its mean:*

$$V(X) = E\left([X - E(X)]^2\right)$$

*For a random variable X having a variance, the **standard deviation** $\sigma(X)$ is the square root of the variance:*

$$\sigma(X) = \sqrt{V(X)}$$

Properties

For all random variables X and Y, and all real numbers a and b

- $V(X) = E[X^2] - E[X]^2$
- If X has a variance, then $aX + b$ has a variance:

$$V(aX + b) = a^2 V(X)$$

- $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$, where covariance is defined by

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Proposition 2.7.1 *For any random variable X possessing a variance, the random variable $Y = \frac{X - E[X]}{\sigma(X)}$ is centered $(E[Y] = 0)$ and standardized $(V(Y) = 1)$.*

2.8 Common Discrete Distributions

2.8.1 Bernoulli Distribution

We say that a real random variable X taking values in $\{0,1\}$ follows a Bernoulli distribution with parameter $p \in (0,1)$, denoted $\mathcal{B}(p)$, if

$$P(X=1) = 1 - P(X=0) = p.$$

We have

$$E(X) = p \quad \text{and} \quad V(X) = p(1-p).$$

A Bernoulli trial is a random experiment with two outcomes (for example, flipping a coin): success (for instance heads) with probability p and failure (tails in this example) with probability $1-p$.

2.8.2 Binomial Distribution

We say that a real random variable X taking values in $\{0,1,...,n\}$ follows a binomial distribution with parameters (n,p), denoted $\mathcal{B}(n,p)$, if

$$P(X=k) = C_n^k p^k (1-p)^{n-k}, \;\; 0 \leq k \leq n.$$

We have

$$E(X) = np \quad \text{and} \quad V(X) = np(1-p).$$

The binomial distribution consists of repeating n independent Bernoulli trials with parameter p.

Example 2.8.1 *We are interested in the infection of trees in a forest by a parasite. Let p be the proportion of infected trees. We study 4 trees. If a tree is infected, we say we have a success; otherwise, we have a failure. Let X be the number of infected trees among the 4. Then X follows the binomial distribution $\mathcal{B}(4,p)$.*

$$P(X=0) = C_4^0(1-p)^4 = (1-p)^4, \qquad P(X=1) = C_4^1 p(1-p)^3 = 4p(1-p)^3,$$
$$P(X=2) = C_4^2 p^2(1-p)^2 = 6p^2(1-p)^2, \quad P(X=3) = C_4^3 p^3(1-p) = 4p^3(1-p),$$
$$P(X=4) = C_4^4 p^4 = p^4.$$

2.8.3 Geometric Distribution

We say that a real random variable X taking values in $\mathbb{N}^*$ follows a geometric distribution with parameter $p \in (0,1)$, denoted $\mathcal{G}(p)$, if

$$P(X = 1) = p(1-p)^{k-1}, \quad k \in \mathbb{N}^*.$$

We have

$$E(X) = 1/p \quad \text{and} \quad V(X) = (1-p)/p^2.$$

The random variable X representing the rank of the first success when independently repeating a Bernoulli trial with parameter p follows a geometric distribution with parameter p.

2.8.4 Poisson Distribution

We say that a real random variable X taking values in $\mathbb{N}$ follows a Poisson distribution with parameter $\lambda > 0$, denoted $\mathcal{P}(\lambda)$, if

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}.$$

We have

$$E(X) = \lambda \quad \text{and} \quad V(X) = \lambda.$$

It is also used to model "rare events." Suppose that over a period T, an event occurs on average λ times. Let X be the random variable representing the number of times the event occurs in the period T. X takes integer values: 0, 1, 2,$\cdots$ and follows a Poisson distribution with parameter λ.

2.9 Common Continuous Distributions

2.9.1 Normal Distribution or Gaussian Distribution

A real random variable X follows a normal (or Gaussian) distribution with mean μ and standard deviation $\sigma > 0$. The distribution of X has the probability density

function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

A normal distribution will be denoted as $\mathcal{N}(\mu, \sigma)$. We have

$$E(X) = \mu \quad \text{and} \quad V(X) = \sigma^2.$$

The normal distribution applies to many phenomena in biology, physics, and economics.

2.9.2 Standard Normal Distribution

When the mean μ is 0 and the standard deviation is 1, the distribution will be denoted $\mathcal{N}(0,1)$ and will be called the standard normal distribution. Its probability density function is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Only the distribution $\mathcal{N}(0,1)$ is tabulated since other distributions (with different parameters) can be derived from this one using the following theorem:

If X follows the distribution $\mathcal{N}(\mu, \sigma)$, then $z = \frac{X-\mu}{\sigma}$ follows the distribution $\mathcal{N}(0,1)$.

Property 2.9.1 *Let X and Y be two independent random variables following the distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$, respectively. Then,*

- *the random variable $Z = \frac{X-\mu_i}{\sigma_i}$, $i = 1, 2$, follows a standard normal distribution $\mathcal{N}(0,1)$.*
- *the random variable $X + Y$ follows the normal distribution $\mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.*

Using Tables: To calculate $P[a \leq X \leq b]$ or $P[X \leq x]$, we can use numerical calculations on a computer or, more simply, a table that provides $P[X \leq x]$ for any positive decimal x up to two decimal places. Then we note that

$$P[a \leq X \leq b] = P[X \leq b] - P[X \leq a]$$

and we can read the probabilities from the table if a and b are positive.
To find $P[X \leq -x]$ when $x > 0$, we use the fact that X and $-X$ have the same

distribution:

$$P[X \leq -x] = P[-X \geq x] = P[X \geq x] = 1 - P[X \leq x]$$

Example 2.9.1 *We want to calculate* $P[-1 \leq X \leq 1]$.

$$\begin{aligned} P[-1 \leq X \leq 1] &= P[X \leq 1] - P[X \leq -1] = P[X \leq 1] - (1 - P[X \leq 1]) \\ &= 2P[X \leq 1] - 1 = 2 \times 0.8413 - 1 = 0.6826 \end{aligned}$$

2.9.3 Approximation of a Binomial Distribution by a Normal Distribution

The normal distribution can be obtained as the limit of a binomial distribution when the number n is large enough and the probability p is not too close to 0 or 1: otherwise, if X follows the binomial distribution $\mathcal{B}(n, p)$ with $n \geq 30$, $np \geq 5$, and $n(1-p) \geq 5$, then practically X follows the normal distribution $\mathcal{N}(np; np(1-p))$.

2.9.4 Chi-Squared Distribution

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of independent random variables each following the same distribution $\mathcal{N}(0, 1)$. Then the random variable $\sum_{i=1}^{n} Z_i^2$ follows a distribution called **chi-squared distribution** with n degrees of freedom, denoted $\chi^2(n)$.

- The distribution of X has the probability density:

$$f(x) = \frac{(x/2)^{n/2-1}}{2\Gamma(n/2)} e^{-x/2}, \quad x > 0$$

  where Γ is the Gamma function defined by $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$.

- The expectation of the $\chi^2(n)$ distribution is equal to the number n of degrees of freedom, and its variance is $2n$.

- The sum of two independent random variables following $\chi^2(n_1)$ and $\chi^2(n_2)$ respectively also follows a χ^2 distribution with $n_1 + n_2$ degrees of freedom.

2.9.5 Student's t-Distribution

Let Z and Q be two independent random variables such that Z follows $\mathcal{N}(0,1)$ and Q follows $\chi^2(n)$. Then the random variable

$$T = \frac{Z}{\sqrt{Q/n}}$$

follows a distribution called **Student's t-distribution** with n degrees of freedom, denoted $St(n)$.

- The density of Student's t-distribution with n degrees of freedom is

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n/2)} \frac{1}{(1+x^2/n)^{\frac{n+1}{2}}}, \quad x > 0$$

- The expectation is not defined for $n = 1$ and equals o if $n \geq 2$. Its variance does not exist for $n \leq 2$ and equals $n/(n-2)$ for $n \geq 3$.
- The Student's t-distribution converges in distribution to the standard normal distribution.
- For $n = 1$, the Student's t-distribution is called the Cauchy distribution, or Lorentz distribution.

2.9.6 Fisher-Snedecor Distribution

Let Q_1 and Q_2 be two independent random variables such that Q_1 follows $\chi^2(n_1)$ and Q_2 follows $\chi^2(n_2)$; then the random variable

$$F = \frac{Q_1/n_1}{Q_2/n_2}$$

follows a Fisher-Snedecor distribution with (n_1, n_2) degrees of freedom, denoted $\mathcal{F}(n_1, n_2)$.

- The density of the Fisher distribution with (n_1, n_2) degrees of freedom is

$$f(x) = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{x^{n_1/2-1}}{(1+n_1 x/n_2)^{\frac{n_1+n_2}{2}}}, \text{ if } x > 0 \ (0 \text{ otherwise})$$

- The expectation is not defined for $n_2 < 3$ and equals $\frac{n_2}{n_2-2}$ if $n_2 \geq 3$. Its variance exists only if $n_2 \geq 5$ and equals $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$.
- If F follows a Fisher distribution $\mathcal{F}(n_1, n_2)$, then $1/F$ follows a Fisher distribution $\mathcal{F}(n_2, n_1)$.
- If T follows a Student's t-distribution with n degrees of freedom, then T^2 follows a Fisher distribution $\mathcal{F}(1, n)$.

2.10 Exercises

Exercise 1

Consider a family with 2 children. We assume that the arrival of a girl is as certain as that of a boy.

1. What is the probability that both children are boys given that the older one is a boy?
2. What is the probability that both children are boys given that at least one of the children is a boy?

Exercise 2

We roll 2 fair dice.

1. What is the probability that at least one of them shows a 6, given that the 2 results are different?
2. What is the probability that at least one of them shows a 6, given that their sum is i? Calculate the result for all possible values of i.

Exercise 3

A student takes a multiple-choice test (MCQ) that consists of 10 questions. For each question, four answers are proposed, only one of which is correct. Strangely, he has never attended class and will therefore choose at random. We denote by X the number of correct answers he provides.

1. Specify the probability distribution followed by X.
2. What is the probability that he has exactly 3 correct answers?
3. What is the probability that he has at most 2 correct answers?

Exercise 4 [Geometric Distribution]

We consider a series of independent trials. In each trial, we observe a "success" with probability p and a "failure" with probability $1-p$. Let X be the following discrete random variable:

$$X = \text{the number of trials needed to achieve the } 1^{st} \text{ "success".}$$

1. Calculate the probability distribution of X, i.e., $P(X = k), k \in \mathbb{N}$. This distribution is called the geometric distribution.
2. Verify that $E(X) = 1/p$.
3. Verify the "memoryless" property of the geometric distribution

$$P_{(X>j)}(X > k) = P(X > k - j), \;\; k > j.$$

Exercise 5

Let X be a random variable whose density function is

$$f(x) = \begin{cases} c(1-x^2) & \text{if} \quad -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

1. Calculate the value of c.
2. What is the cumulative distribution function of X?
3. Calculate $E(X)$.

Exercise 6

We say that the random variable X follows the exponential distribution with parameter λ, denoted $\varepsilon(\lambda)$, if its density is

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if} \quad x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

1. Calculate the cumulative distribution function $F(x)$.
2. What is the α-quantile $q = F^{-1}(\alpha)$?

Exercise 7

We assume that the height, in centimeters, of a 25-year-old pygmy is a normal random variable with parameters $\mu = 140$ and $\sigma = 6$.

1. What percentage of 25-year-old pygmies have a height greater than 150 cm?
2. Among pygmies measuring more than 145 cm, what percentage exceed 150 cm?

Exercise 8

Let $X \rightsquigarrow \mathcal{N}(0,1)$ and define $Y = X^2$.
Find the cumulative distribution function of Y and its density.
Graphically represent this density.
This distribution is called a $\mathcal{X}^2(1)$ distribution (chi-squared with one degree of freedom).

Exercise 9

In a given country, the serum cholesterol level of a randomly selected individual is modeled by a normal distribution with a mean of 200 mg/100 mL and a standard deviation of 20 mg/100 mL.

1. What is the probability that a randomly selected individual in this country has a cholesterol level less than 160 mg/100 mL?

2. What proportion of the population has a cholesterol level between 170 and 230 mg/100 mL?

3. In another country, the average serum cholesterol level is 190 mg/100 mL, with the same standard deviation. Revisit the previous questions.

4. We select an individual at random from the first country, then from the second. What is the probability that the first individual has a higher cholesterol level than the second?

Exercise 10

The height X of men in France is modeled by a normal distribution $\mathcal{N}(172, 196)$ (unit: cm).

1. What proportion of French people have a height less than 160 cm?

$$P(X < 160) = P\left(\frac{X - 172}{\sqrt{196}} < \frac{160 - 172}{\sqrt{196}}\right) = \pi(-0.857) = 1 - \pi(0.857) = 0.1957$$

where π denotes (as in the rest of the exercise) the cumulative distribution function of the law $\mathcal{N}(0, 1)$.

2. What proportion of French people are taller than two meters?

$$P(X > 200) = P\left(\frac{X - 172}{\sqrt{196}} > \frac{200 - 172}{\sqrt{196}}\right) = P(Z > 2) = 1 - \pi(2) = 0.0227$$

3. What proportion of French people measure between 165 and 185 centimeters?

$$\begin{aligned} P(165 < X < 185) &= P\left(\frac{165 - 172}{\sqrt{196}} < \frac{X - 172}{\sqrt{196}} < \frac{185 - 172}{\sqrt{196}}\right) \\ &= \pi(0.928) - \pi(-0.5) = 0.8234 - 0.3085 = 0.5149. \end{aligned}$$

4. If we ranked ten thousand randomly selected French people in ascending order of height, what would be the height of the 9000th?

Answer:

This question amounts to finding the height such that 90% of French people have a height below it, namely the 0.9 quantile, or the ninth decile. Let x be this height.

$$P(X < x) = P\left(\frac{X - 172}{\sqrt{196}} < \frac{x - 172}{\sqrt{196}}\right) = 0.90$$

Therefore, $\frac{X-172}{\sqrt{196}}$ is the quantile function of the distribution $\mathcal{N}(0,1)$ at the point 0.9, which is 1.2816. Thus, we deduce:

$$x = 172 + 1.2816 \times \sqrt{196} \simeq 190 \text{ cm}$$

5. The height Y of women in France is modeled by a normal distribution $\mathcal{N}(162, 144)$ (in centimeters). What is the probability that a randomly chosen man is taller than a randomly chosen woman?

   If X denotes the height of the man and Y the height of the woman, assumed to be independent, then $X - Y$ follows the normal distribution $\mathcal{N}(10, 340)$. The probability that X is greater than Y is the probability that the difference is positive:

$$P(X - Y > 0) = P\left(\frac{(X - Y) - 10}{\sqrt{340}} < \frac{0 - 10}{\sqrt{340}}\right) = 1 - \pi(0.5423) = 0.7062.$$

Exercise 11

In a given country, the serum cholesterol level of a randomly selected individual is modeled by a normal distribution with a mean of 200 mg/100 mL and a standard deviation of 20 mg/100 mL.

1. What is the probability that a randomly selected individual in this country has a cholesterol level less than 160 mg/100 mL?

2. What proportion of the population has a cholesterol level between 170 and 230 mg/100 mL?

3. In another country, the average serum cholesterol level is 190 mg/100 mL, with the same standard deviation. Revisit the previous questions.
4. We select an individual at random from the first country, then from the second. What is the probability that the first individual has a higher cholesterol level than the second?

Exercise 3: Let X be a random variable following the distribution $\mathcal{N}(3, 25)$

1. Express using the cumulative distribution function of the law $\mathcal{N}(0, 1)$, and then calculate the following probabilities using the table.

$$(a)\ \ P(X < 6) \qquad (b)\ \ P(X > -2) \qquad (c)\ \ P(-1 \leq X \leq 1.5)$$

2. Determine the value of u in the following cases.

$$(a)\ P(X < u) = 0.36 \quad (b)\ P(X > u) = 0.36 \quad (c)\ P(|X - 3| \leq u) = 0.36$$

Exercise 12

It is known from experience that a certain surgical operation has a 90% chance of success. This operation is performed in a clinic 400 times each year. Let N be the number of successes in a year. We will use the normal approximation for N.

1. Calculate the expectation and variance of N.

   The expectation is $400 \times 0.9 = 360$, and the variance is $400 \times 0.9 \times 0.1 = 36$.

2. Calculate the probability that the clinic succeeds in at least 345 operations in the year.

$$P(N \geq 345) = P\left(\frac{N - 360}{\sqrt{36}} \geq \frac{345 - 360}{\sqrt{36}}\right) = 1 - \pi(-2.5) = \pi(2.5) = 0.9938$$

3. Calculate the probability that the clinic fails more than 28 operations in the year.

$$P(N \leq 372) = P\left(\frac{N - 360}{\sqrt{36}} \leq \frac{372 - 360}{\sqrt{36}}\right) = \pi(2) = 0.9772$$

Exercise 13

We estimate the probability that a person of vaccination age against the flu actually requests to be vaccinated at 0.4. In a population of 150,000 people of vaccination age, let N be the number of people who will request to be vaccinated.

1. What model would you propose for N?
2. If we prepare 60,500 vaccines, what is the probability that there are not enough?
3. Calculate the number m of vaccines that should be planned so that the probability of running out is equal to 0.1.

Exercise 14

We measure the height in cm of 2500 bacteria; the resulting distribution follows a normal law with a mean of 169 μm and a standard deviation of 5.6 μm.

1. What is the percentage of bacteria whose size is less than 155 μm?
2. What is the percentage of men whose height is between 155 cm and 175 μm?
3. What is the interval, centered on the average height, that contains 60% of the population in question?

Sampling Fluctuation 3

From a population P, multiple samples E can be drawn in various ways. The random selection of successive samples E_1, E_2, ..., E_I usually results in different values for quantities such as f (frequency), m (mean), and s^2 (variance).

The characteristics provided are not the exact characteristics of the population. They deviate to some extent due to the randomness of sampling.

This is referred to as sampling fluctuations.

Calculating an interval of fluctuation is thus another way to represent the dispersion of a variable.

3.1 Sampling Fluctuations of a Mean

We consider a population whose elements possess a measurable characteristic represented by the realization of a random variable X that follows a probability distribution with expectation μ and standard deviation σ_{pop}. We assume that the population is infinite, or if it is finite, that sampling is done with replacement.

- A random sample of size n is taken, and the values of X are measured for each element in the sample, yielding a sequence of values x_1, x_2,....,x_n.
- If a second sample of the same size n is taken, the sequence of values obtained is x'_1, x'_2,....,x'_n, then x''_1, x''_2,....,x''_n... etc. for additional samples.

The values x_1, x'_1,x''_1.... can be considered as values of a random variable X_1 that follows the distribution of X. Similarly, x_2, x'_2, x''_2... can be considered as values of a random variable X_2 that also follows the distribution of X,.... and x_n, x'_n, x''_n... as values of a random variable X_n that still follows the same distribution, that of X.

- X_1 could be named the "value of the first element of a sample."
  X_2 could be named the "value of the second element of a sample."
  ... X_n could be named the "value of the nth element of a sample."

- The assumption of an infinite population or sampling with replacement allows us to assert that these n random variables are independent.

Notation reminder: By convention, random variables are always denoted by uppercase letters (X_i), and the values they take in a realization by lowercase letters (x_i).

If the values taken by X in a sample are x_1, x_2,....,x_n, the sample mean $\bar{x}$ is given by $\bar{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$. This value is simply the value taken in this sample by the random variable $\dfrac{X_1 + X_2 + ... + X_n}{n}$.

Definition 3.1.1 *Let X be a random variable on a sample space Ω. A **sample** of X of size n is an n-tuple $(X_1, ..., X_n)$ of independent random variables with the same distribution as X. The distribution of X will be called the parent distribution. A realization of this sample is an n-tuple of real values $(x_1, ..., x_n)$ where $X_i(\omega) = x_i$.*

Definition 3.1.2 *The sample mean or empirical mean, denoted $\bar{X}$, is the statistic defined by*

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Proposition 3.1.1 *Let X be a random variable with mean μ and standard deviation σ_{pop}. Then:*

$$E(\bar{X}) = \mu, \qquad V(\bar{X}) = \frac{\sigma_{pop}^2}{n}.$$

Remark 3.1.1

- *The mean of the sampling distribution of means equals the population mean.*
- *As n increases, $Var(\bar{X})$ decreases.*

3.1.1 Distribution of the Sample Mean

Case of Large Samples: $n \geq 30$

Proposition 3.1.2 *Let X be a random variable with mean μ and standard deviation σ_{pop}.*

By the central limit theorem, $\bar{X}$ converges in distribution to $\mathcal{N}(\mu, \frac{\sigma_{pop}}{\sqrt{n}})$ when n is large enough.

Remark 3.1.2

- *This proposition is very powerful because it imposes no restrictions on the distribution of X in the population.*
- *If the variance is unknown, a large sample ($n \geq 30$) allows a reliable estimate for σ^2_{pop} by calculating the sample variance σ^2_{sample} and using*

$$\sigma^2_{pop} = \frac{n}{n-1}\sigma^2_{sample} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Case of Small Samples: $n < 30$

We then exclusively consider the case where X follows a normal distribution in the population.

Proposition 3.1.3 *Let X be a random variable with mean μ and standard deviation σ_{pop}.*

- *If σ_{pop} is **known** and $n < 30$, then $\bar{X}$ converges in distribution to $\mathcal{N}(\mu, \frac{\sigma_{pop}}{\sqrt{n}})$.*
- *If σ_{pop} is **unknown** and $n < 30$, then $T = \frac{\bar{X}-\mu}{\sigma_{sample}/\sqrt{n-1}}$ follows a Student's t-distribution with $n-1$ degrees of freedom, denoted T_{n-1}.*

3.2 Sampling Fluctuations of a Variance

Definition 3.2.1 *The Empirical Variance is the statistic denoted by $\sigma^2_{sample}(X)$, defined by*

$$\sigma^2_{sample} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Proposition 3.2.1 *Let X be a random variable with standard deviation σ_{pop}. We have:*

$$E(\sigma^2_{sample}) = \frac{n-1}{n}\sigma^2_{pop}.$$

Conclusion. The mean of sample variances is not equal to the population variance, but rather to the population variance multiplied by $\frac{n}{n-1}$. Naturally, if n is very large, these two numbers will be very close to each other.

3.3 Sampling Fluctuations of a Proportion

F is the frequency of occurrence of the characteristic in a sample of size n. Thus, $F = X/n$ where X is the number of times the characteristic appears in the n-sample.

By definition, X follows $\mathcal{B}(n, p)$. Therefore, $E(X) = np$ and $V(X) = npq$. It follows that

$$E(F) = p \qquad \text{and} \qquad V(F) = \frac{pq}{n}$$

Consequences.

- The expectation of the sample frequency is equal to the theoretical probability of occurrence in the population.
- When the sample size increases, the variance of F decreases, which is logical: the more information we have, the more likely it is that the observed proportion in the sample is close to the population proportion.

We know that if $n \geq 30$, $np \geq 15$, and $nq \geq 15$, we can approximate the binomial distribution by a normal distribution with the same mean and standard deviation. Therefore, F approximately follows $\mathcal{N}(p, \sqrt{\frac{pq}{n}})$.

3.4 Exercises

Exercise 1

In a given country, the mean and standard deviation of cholesterol levels in the population P are equal to 2.2 g/L and 0.52 g/L, respectively. A random sample of 42 women is taken.

1. What is the probability that the mean cholesterol level of the sample is less than 2.6 g/L?

Exercise 2

According to a study on consumer behavior, 25% of consumers are influenced by the brand when purchasing a product. If we randomly survey 100 consumers, what is the probability that at least 35 of them report being influenced by the brand?

Answer:

Let F be the random variable: "sample proportion in a sample of size 100." Here, it represents the proportion of consumers in the sample who report being influenced by the brand. We seek to calculate $P(F > 0.35)$.

We need to determine the distribution of F. Since $np = 100 * 0.25 = 25 > 15$ and $nq = 100 * 0.75 = 75 > 15$, we can consider that F follows $\mathcal{N}(p, \sqrt{\frac{pq}{n}}) = \mathcal{N}(0.25, 0.043)$.

We use the variable $T = \frac{F-0.25}{0.0433}$, which follows the $\mathcal{N}(0, 1)$ distribution. Thus, $P(F > 0.35) = P(T > 2.31) = 0.5 - P(0 < T < 2.31) = 0.5 - 0.4896 = 0.0104$.

Conclusion. There is approximately a 1 in 100 chance that more than 35 consumers in a sample of 100 will report being influenced by the brand when 25% of the entire population are such consumers.

Estimation 4

When studying a characteristic in a population, it is often necessary to examine this characteristic in a group of subjects (sample) before generalizing the results to the population. Estimation is the theory that allows this generalization from the sample to the population. This is the reverse problem of sampling.

Based on the characteristics of a sample, what can we infer about the characteristics of the population from which it is drawn?

Estimation involves providing approximate values for the parameters of a population using a sample of n observations drawn from this population. We may be wrong about the exact value, but we provide the "best possible value" that we can assume.

Point estimation, which provides a single value that we hope is as close as possible to the true value of the parameter.

Interval estimation, which gives an interval, called the confidence interval, that has a probability fixed in advance of containing the true value of the parameter.

4.1 Point Estimation

We want to estimate a parameter θ of a population (this could be its mean μ, its standard deviation σ, a proportion p). An estimator of θ is a statistic T (thus a function of $(X_1, ..., X_n)$) whose realization is considered a "good value" of the

parameter θ. The estimate of θ associated with this estimator is the observed value during the experiment, i.e., the value taken by the function at the observed point $(x_1, ..., x_n)$.

Example: To estimate the expectation $E(X)$ of the law of X, a natural estimator is the empirical mean $\bar{X}$, which produces an estimate $\bar{x}$, the descriptive mean of the series of observed values.

Unbiased estimators: The bias of an estimator is evaluated by the difference between the estimates of a parameter obtained on successive samples and the true value of the parameter. **An unbiased estimator** is a random variable whose expectation is equal to the exact value of the quantity we want to estimate.

Otherwise, we say that we have a **biased estimator**.

Efficient estimators: When estimating the mean, the sampling distribution of two statistics has the same expectation; the statistic with the lowest variance is called the "efficient estimator" of the mean, and the other statistic is called the "inefficient estimator." Sometimes, the efficient estimator is called the "best estimator."

4.1.1 Some classic estimators

- $\bar{X}$ is an unbiased estimator of the mean μ (since $E(\bar{X}) = \mu$). Its estimate $\bar{x}$ is the mean observed in a realization of the sample.
- $\sigma^2_{sample}(X)$ is a biased estimator for σ^2_{pop} (since $E(\sigma^2_{sample}) = \frac{n-1}{n}\sigma^2_{pop}$).
- As an unbiased estimator of the variance, we can propose:

$$S^2 = \frac{n}{n-1}\sigma^2_{sample} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  also denoted $\widehat{\sigma}^2$. Its estimate is $s^2 = \frac{n}{n-1}\sigma^2_{sample}$, where σ_{sample} is the observed standard deviation in a realization of the sample. s^2 is called the corrected empirical variance of the sample.

- If p is the frequency of a characteristic, F constitutes an unbiased estimator of p. Its estimate is denoted f.

Exercise: Consider the statistical sample (1, 0, 2, 1, 1, 0, 1, 0, 0).

1. Calculate its empirical mean and variance.
   We find:
   $$\bar{x} = \frac{6}{9} = \frac{2}{3} \qquad \text{and} \qquad s^2 = \frac{4}{9}$$

2. Assuming that the data in this sample are realizations of an unknown law variable, provide an unbiased estimate of the expectation and variance of this law.

   The empirical mean (2/3) is an unbiased estimate of the expectation. We obtain an unbiased estimate of the variance by multiplying s^2 by 9/8: we find 1/2.

3. We choose to model the values of this sample by a binomial law $\mathcal{B}(2, p)$. Use the empirical mean to propose a point estimate for p.

   The expectation of the law $\mathcal{B}(2, p)$ is $2p$. It is estimated by the empirical mean (here: 2/3). Thus, the probability p can be estimated by
   $$\frac{2/3}{2} = \frac{1}{3}$$

4. With the same model, use the empirical variance to propose another estimate of p.

   The variance of the law $\mathcal{B}(2, p)$ is $2p(1-p)$. It is estimated by 1/2. We obtain an estimate of p by solving the equation $2p(1-p) = 1/2$, whose solution is $p = 1/2$.

4.2 Confidence Intervals

Instead of using a function (estimator) that gives a point estimate of a parameter, we seek an interval in which the studied parameter is located with a controlled (and generally high) probability.

4.2.1 Estimation of a mean by confidence interval

Consider a random variable X following $\mathcal{N}(\mu, \sigma_{pop})$ and $X_1, ..., X_n$, n independent variables with the same law as X (or when the law of X is unknown, and we have a sample of size $n \geq 30$). We recall that the definitions of the empirical mean and corrected empirical variance (or modified) are respectively given by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Let $z_{\alpha/2}$ be the positive real number such that $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. We know that the random variable $\bar{X}$ follows the normal law $\mathcal{N}(\mu, \sigma/\sqrt{n})$, hence

$$\begin{aligned} 1 - \alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= P\left(-z_{\alpha/2} < \frac{X - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \\ &= P(\bar{X} - z_{\alpha/2}.\sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}.\sigma/\sqrt{n}) \end{aligned}$$

The confidence interval for the mean of a population with known variance σ^2 is given by

$$\bar{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$$

The probability that the mean $\bar{x}$ is within the interval $I = \left]\bar{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}\right[$ is: $P(I) = 1 - \alpha$.

Error Risk α: Here we call the interval I a confidence interval, $(1 - \alpha)$ is called the confidence level, and α is the error risk.

$z_{\alpha/2}$ is a value given by the table of the standard normal distribution.
According to the properties of the normal distribution, we generally choose the error risk ($\alpha = 5\%$), and in certain cases, we set ($\alpha = 1\%$):

1) for $\alpha = 5\%$, we choose $z_{\alpha/2} = 1.96$, and in this case $P(I) = 0.95$.

2) for $\alpha = 1\%$, we choose $z_{\alpha/2} = 2.6$, and in this case $P(I) = 0.99$.

Proposition 4.2.1 *The variable* $\dfrac{(n-1)S^2}{\sigma^2}$ *follows a* χ^2 *distribution with* $n-1$ *degrees of freedom*

Case where σ_{pop} is unknown:

When the variance σ^2_{pop} is unknown, it is necessary to replace this quantity in the previous formulas with the sample variance, which is a consistent estimator. Thus, we need to consider not $\frac{X-\mu}{\sigma/\sqrt{n}}$ but rather

$$\frac{X-\mu}{S/\sqrt{n}}$$

which no longer follows a normal distribution but rather a distribution known as the Student's t-distribution with $n-1$ degrees of freedom, denoted T_{n-1}. We have tables to obtain the quantiles of this distribution. We therefore conclude that

$$1-\alpha = P(-t_{\alpha/2} < Z < t_{\alpha/2}) = P\left(-t_{\alpha/2} < \frac{X-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right)$$

Thus, this interval is given by:

$$\bar{x} - t_{\alpha/2}.\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}.\frac{s}{\sqrt{n}}$$

where $\bar{x}$ and s^2 are the point estimates of the mean μ and the variance σ^2_{pop}, respectively. $t_{\alpha/2}$ will be read from the Student's distribution for the risk $\alpha/2$ with $n-1$ degrees of freedom.

4.2.2 Estimating a Proportion by Confidence Interval

We consider a population such that the proportion p for a certain category of the observed trait is unknown. We want to estimate this proportion p of this population from a sample of size n with the frequency of the studied category being f. Let F be the random variable that associates the frequency of elements belonging to the chosen category to each sample of size n. It is known that F approximately follows the distribution $\mathcal{N}(p, \sigma)$ with $\sigma = \sqrt{pq/n}$ for sufficiently

large n ($n > 30$). We have

$$\sigma' = \sqrt{\frac{f(1-f)}{n}}$$

the standard deviation associated with the frequency f of the sample of size n. Since p is unknown, we use the point estimate of σ:

$$\sigma = \sigma' \sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n}} \sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n-1}}$$

Therefore, the random variable Z defined by:

$$Z = \frac{F-p}{\sigma}$$

approximately follows a standard normal distribution $\mathcal{N}(0,1)$. We seek a confidence interval for the proportion p, i.e., an interval such that the probability of the proportion p not belonging to this interval is equal to α, where $\alpha \in [0;1]$. We call this interval a confidence interval with error risk α or with confidence coefficient $c = 1-\alpha$. The risk we take in saying that p belongs to this interval is therefore α, or equivalently, the probability that p does not belong to this interval is the risk α.

Let's determine this confidence interval:

$$1-\alpha = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\left(-z_{\alpha/2} < \frac{F-p}{\sigma} < z_{\alpha/2}\right)$$

$$= P(F - z_{\alpha/2}.\sigma < p < F + z_{\alpha/2}.\sigma)$$

$$= P\left(F - z_{\alpha/2}.\sqrt{\frac{f(1-f)}{n-1}} < p < F + z_{\alpha/2}.\sqrt{\frac{f(1-f)}{n-1}}\right)$$

The confidence interval for the proportion p with a confidence coefficient of $1-\alpha$ is:

$$I = \left]f - z_{\alpha/2}.\sqrt{\frac{f(1-f)}{n-1}}, f + z_{\alpha/2}.\sqrt{\frac{f(1-f)}{n-1}}\right[$$

Remark: when n is large, the difference between n and $n-1$ becomes negligible,

so the formula becomes

$$\left]f - z_{\alpha/2}\cdot\sqrt{\frac{f(1-f)}{n}}, f + z_{\alpha/2}\cdot\sqrt{\frac{f(1-f)}{n}}\right[$$

This is the most commonly used formula.

4.2.3 Estimation of a Variance by Confidence Interval

We consider the modified sample variance S^2. According to Proposition 4.2.1, we know that $\frac{(n-1)S^2}{\sigma^2}$ follows a χ^2 distribution with $n-1$ degrees of freedom. Let us determine this confidence interval:

$$1-\alpha = P(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}) = P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right)$$

$$= P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right)$$

where $\chi^2_{\alpha/2}$ will be read from the χ^2 table with $n-1$ degrees of freedom. We will seek values such that $P(K^2 > \chi^2_{\alpha/2}) = \alpha/2$ and $P(K^2 < \chi^2_{1-\alpha/2}) = \alpha/2$.

4.3 Exercises

Exercise 1

In a population P, we are interested in the blood cholesterol level (g/L). We randomly select a sample of 32 women. The mean and standard deviation of the cholesterol level in the sample are equal to 2.2 and 0.52, respectively.

- Calculate the 99% confidence interval for the mean cholesterol level of the entire female population.

Answer:

Here, $\alpha = 0.01$ and $1 - \alpha/2 = 0.995$. The 0.995 quantile of the $\mathcal{N}(0,1)$ distribution is 2.57. The confidence interval is:

$$\left]\bar{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n-1}}; \bar{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n-1}}\right[= \left]2.2 - 2.57.\frac{0.52}{\sqrt{31}}; 2.2 + 2.57.\frac{0.52}{\sqrt{31}}\right[= [1.96; 2.44]$$

Exercise 2

In a cancer center, a random sample of 100 women suspected of uterine cancer is examined. In fact, 25% of these women have uterine cancer.

- What is the 5% risk confidence interval for the frequency of uterine cancer F in the population of suspected women received at the cancer center?

Answer:

We have $n = 100$, $f = 0.25$. We assume that the sample is large. The application conditions $n > 30$, $np > 15$, and $nq > 15$ are assumed to be satisfied.

Here, $\alpha = 0.05$ and $1 - \alpha/2 = 0.975$. The 0.975 quantile of the $\mathcal{N}(0,1)$ distribution is 1.96. The confidence interval is:

$$\left]f - z_{\alpha/2}.\sqrt{\frac{f(1-f)}{n}}, f + z_{\alpha/2}.\sqrt{\frac{f(1-f)}{n}}\right[$$

$$= \left]0.25 - 1.96.\sqrt{\frac{0.25 * 0.75}{100}}, 0.25 + 1.96.\sqrt{\frac{0.25 * 0.75}{100}}\right[=]0.16, 0.34[$$

Exercise 3

The grape weight produced per vine was measured on 10 randomly selected vines in a vineyard. The following results, in kilograms, were obtained:

2.4 3.4 3.6 4.1 4.3 4.7 5.4 5.9 6.5 6.9.

We model the grape weight produced by a vine in this vineyard by a random variable following a $\mathcal{N}(\mu, \sigma)$ distribution.

1. Calculate the empirical mean and variance of the sample.
2. Provide a 0.95 confidence interval for μ.
3. Provide a 0.95 confidence interval for σ.
4. Now assume the standard deviation of production per vine is known and equal to 1.4. Provide a 0.95 confidence interval for μ.
5. What minimum number of vines should be observed to estimate μ with a confidence level of 0.99 and a precision of plus or minus $500g$?

Exercise 4

A sample of size $n = 500$ of 15-year-old adolescents was observed, in which 210 are overweight. Let p be the proportion of 15-year-old adolescents who are overweight. Provide a confidence interval for p at confidence levels of 0.95 and 0.99.

Exercise 5

Assume that the birth weight of a newborn sheep is a normal variable with a standard deviation of 0.5 kg. The average weight of the 49 lambs born on a farm in the spring was 3.6 kg.

a) Determine a 95% confidence interval for the average weight of a newborn lamb on this farm.

b) What would be the confidence level of an interval of length 0.1 kg centered at 3.6 for this average weight?

c) Compare this population with another one with the following parameters: (Mean = 4.2; SD = 0.8; $n = 36$; $\alpha = 0.05$).

Concept of Hypothesis Testing 5

The principle of hypothesis testing is to establish a working hypothesis and predict the consequences of this hypothesis for the population or sample. We compare these predictions with observations and conclude by accepting or rejecting the working hypothesis based on objective decision rules. Therefore, it is essential to carefully formulate the hypotheses to ensure they are appropriate and that the conclusions of the hypothesis test provide the information desired by the researcher.

To test a hypothesis, different steps must be followed:

1. Formulate the test hypotheses H_0 and H_1.
2. Set the significance level α.
3. Choose a statistical test or statistic to test H_0.
4. Define the distribution of the statistic under the hypothesis "H_0 is true".
5. Define the test's significance level or critical region, denoted α.
6. Calculate, from the data provided by the sample, the value of the statistic.
7. Make a decision regarding the hypothesis and provide an interpretation.

The null hypothesis: This states that there is no difference between the compared parameters or that the observed difference is not significant and is due to sampling fluctuations. This hypothesis is formulated with the intention of being

rejected. It is denoted H_0

The alternative hypothesis: denoted H_1, is the "negation" of H_0, meaning "H_0 is false." Rejecting H_0 means that H_1 holds or H_1 is true.

Consider a population in which we wish to study a characteristic (random variable) X following a distribution (known or unknown) with an unknown parameter θ and let $X_1, ..., X_n$ be an independent sample of X.

In this case, the hypothesis test can take two forms:

$$\begin{cases} H_0 & : & \theta = \theta_0 \\ H_1 & : & \theta \neq \theta_0 \end{cases} \quad : \text{This is a two-tailed test.}$$

$$\begin{cases} H_0 & : & \theta \leq \theta_0 \\ H_1 & : & \theta > \theta_0 \end{cases} \quad : \text{This is a right-tailed test.}$$

Example: A beverage manufacturer claims that two-liter bottles contain, on average, at least 2.028 liters of their product. A sample of two-liter bottles is selected, and their contents are measured to test the manufacturer's claim. In this type of hypothesis test, it is generally assumed that the manufacturer's claim is true unless the sample proves otherwise.
The formulated hypotheses are: $\begin{cases} H_0 & : & \mu \leq 2.028 \\ H_1 & : & \mu > 2.028 \end{cases}$

5.1 Errors, Significance Level, and Power of a Test

The error made by rejecting the null hypothesis H_0 when it is actually true is called a Type I error with probability α, defined by:

$$\alpha = P(\text{rejecting } H_0 \text{ given that } H_0 \text{ is true})$$

With α: called the significance level.

The error made by accepting the null hypothesis H_0 when it is actually false

is called a Type II error with probability β, defined by:

$$\beta = P(\text{accepting } H_0 \text{ given that } H_0 \text{ is false})$$

The test power, denoted π, is the probability of rejecting the hypothesis H_0 when this hypothesis is false, given by:

$$\pi = P(\text{rejecting } H_0 \text{ given that } H_0 \text{ is false}) = 1 - \beta$$

5.2 Test of Conformity

5.2.1 Test Related to Means

Case of a Large Sample or Known Variance:

We assume we have a sample following a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where the variance is known.

Assuming H_0 is true, the test used is the Z test

$$\text{Test statistic} \quad Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad \text{follows a standard normal distribution} \quad \mathcal{N}(0,1)$$

Null Hypothesis H_0	Alternative Hypothesis H_1	Decision Rule
$\mu = \mu_0$	$\mu \neq \mu_0$	Accept H_0 if $Z \in]-z_{\alpha/2}, z_{\alpha/2}[$ with $z_{\alpha/2}$ such that $P(\lvert Z\rvert > z_{\alpha/2}) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	Accept H_0 if $Z \leq z_\alpha$ with z_α such that $P(Z < z_\alpha) = 1 - \alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	Accept H_0 if $Z \geq -z_\alpha$ with z_α such that $P(Z < z_\alpha) = 1 - \alpha$

Case of Small Samples (Unknown Variance):

We assume we have a sample following a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where the variance is now unknown. Since the variance is unknown, we estimate it using the sample's empirical variance.

$$\text{Test statistic} \quad T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad \text{follows a Student's t-distribution}$$

with $(n-1)$ degrees of freedom.

Null Hypothesis H_0	Alternative Hypothesis H_1	Decision Rule
$\mu = \mu_0$	$\mu \neq \mu_0$	Accept H_0 if $T \in]-t_\alpha, t_\alpha[$ with t_α such that $P(\lvert T\rvert > t_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	Accept H_0 if $T \leq t_\alpha$ with t_α such that $P(\lvert T\rvert > t_\alpha) = 2\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	Accept H_0 if $T \geq -t_\alpha$ with t_α such that $P(\lvert T\rvert > t_\alpha) = 2\alpha$

5.2.2 Test Related to Proportions

The mathematical model is as follows. We have a population in which each individual either possesses or does not possess a certain characteristic, with the proportion of individuals possessing the characteristic denoted by p. We also have a random sample of size n drawn from this population. The proportion f calculated from the sample is considered as a realization of a random variable with a binomial distribution $\mathcal{B}(n; p)$, which can be approximated, if n is large enough, by a normal distribution $\mathcal{N}(p, \sqrt{pq/n})$.

$$\text{The test used is:} \quad Z = \frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$\text{follows a standard normal distribution} \quad \mathcal{N}(0,1)$$

Null Hypothesis H_0	Alternative Hypothesis H_1	Decision Rules
$p = p_0$	$p \neq p_0$	accept H_0 if $Z \in]-z_{\alpha/2}, z_{\alpha/2}[$ with $z_{\alpha/2}$ such that $P(\|Z\| > z_{\alpha/2}) = \alpha$
$p \leq p_0$	$p > p_0$	accept H_0 if $Z \leq z_\alpha$ with z_α such that $P(Z < z_\alpha) = 1 - \alpha$
$p \geq p_0$	$p < p_0$	accept H_0 if $Z \geq -z_\alpha$ with z_α such that $P(Z < z_\alpha) = 1 - \alpha$

5.3 Exercises

Exercise 1

The normal blood glucose level is 1 g/l of blood. Glucose levels were measured in 17 diabetic subjects who had not eaten for four hours, with an average of 1.2 g/l and a standard deviation of 0.1 g/l. Can we say, with a 5% risk, that these subjects are hyperglycemic, assuming that blood glucose levels follow a normal distribution?

Answer:

We have $n = 17 < 30$, $\bar{X} = 1.2$, $S = 0.1$, $\mu_0 = 1$, and $\alpha = 0.05$. We can perform the following statistical test:

$$\begin{cases} H_0 & : \quad \mu = 1 (\leq) \\ H_1 & : \quad \mu > 1 \end{cases}$$

Calculating

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{1.2 - 1}{0.1/\sqrt{17}} = 7$$

For $\alpha = 0.05$, we have $t_\alpha = 1.74$. Since $T > t_\alpha$, we reject H_0, meaning the group of individuals is hyperglycemic with a 5% risk.

Exercise 2

A drug manufacturer claims it is at least 90% effective in curing an allergy within 8 hours. In a sample of 200 people with this allergy, 160 were cured by the drug. Can we say that the cure proportion is lower than p_0 at the 5% risk level?

Answer:

We have $n = 200$, $p_0 = 0.9$, $F = \frac{160}{200} = 0.8$, and $\alpha = 0.05$. We can perform the following statistical test:

$$\begin{cases} H_0 & : & p_0 = 0.9 (\geq) \\ H_1 & : & p_0 < 0.9 \end{cases}$$

Calculating

$$Z = \frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.8 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{200}}} = -4.71$$

For $\alpha = 0.05$, we have $z_\alpha = 1.64$. Since $Z < z_\alpha$, we reject H_0, meaning that at a 5% risk, the proportion is lower than the manufacturer's claim.

Exercise 3

It is assumed that the average height of full-term newborns is 50 cm, with a standard deviation of 3. In a sample of 50 representative preterm newborns, an average height of 45 cm was observed.

Is the difference significant? In other words, can we state that preterm newborns are shorter than full-term newborns?

Answer:

In population P, $\mu_0 = 50$ cm and $\sigma = 3$ cm. For the sample, $\bar{X} = 45$ cm.

This is a comparison of an observed mean to a theoretical mean.

1. Null hypothesis H_0: There is no significant difference between the height

of preterm and full-term newborns.

$$\begin{cases} H_0 & : \quad \mu = 50 \\ H_1 & : \quad \mu \neq 50 \end{cases}$$

2. Under the null hypothesis H_0: $N > 30$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{45 - 50}{3/\sqrt{50}} = -11.7 \quad \text{which follows} \quad \mathcal{N}(0,1).$$

3. At the $\alpha = 0.05$ level, $z_\alpha = 1.96$. Since $Z < -z_\alpha$, we reject H_0, meaning there is a significant difference between the two means.

   Preterm newborns are significantly shorter than full-term newborns.

Exercise 4

Out of 10,000 children born between 1968 and 1973, 5300 were girls.

Is the proportion of girls compatible with the hypothesis of an equal probability of having a girl or a boy at the 1% risk level?

Answer:

We use the test for comparing an observed frequency to a theoretical frequency $p_0 = 0.5$.

1. Null hypothesis: equal probability of having a girl or a boy.

$$\begin{cases} H_0 & : \quad p_0 = 0.5 \\ H_1 & : \quad p_0 \neq 0.5 \end{cases}$$

2. Under the null hypothesis H_0: $np > 15$ and $nq > 15$

$$Z = \frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.53 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{10000}}} = 6 \quad \text{which follows} \quad \mathcal{N}(0,1).$$

3. At the $\alpha = 0.01$ level, $z_\alpha = 2.567$. Since $Z > z_\alpha$, we reject H_0, meaning there is no equal probability of having a girl or a boy.

Exercise 5

A rare disease in goats (present in 5% of the population) was the subject of an epidemiological investigation. In a sample of 1354 goats, a serological study found 116 infected females. Is the hypothesis verified?

Exercise 6

In a sample of 300 lambs treated with an anti-diarrheal medication, 243 were cured.

Test, specifying the probability level, the hypothesis that the cure proportion is 75%.

Exercise 7

A genetic anomaly affects 1 in 1000 individuals in France. In a given region, 57 people were affected out of 50,000 births. Is this region representative of the entire.

Exercise 8

In order to test a toxic solution, injections are given to a group of 80 mice. It is assumed that the injection is lethal in 80% of cases. Is the fact that 22 mice did not die compatible with this hypothesis at the 5% significance level?

5.4 Comparison of Two Independent Samples

5.4.1 Test for Comparing Two Means

Case of Large Samples or Known Variance:

Assume we have two randomly and independently drawn samples that follow the normal distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ where the variance is known.

We want to test H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$.

Assuming H_0 is true, the test used is that of Z:

The test statistic $Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ follows a standard normal distribution $\mathcal{N}(0,1)$

Null Hypothesis H_0	Alternative Hypo. H_1	Decision Rules
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	Accept H_0 if $Z \in]-z_{\alpha/2}, z_{\alpha/2}[$ with $z_{\alpha/2}$ such that $P(\lvert Z\rvert > z_{\alpha/2}) = \alpha$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	Accept H_0 if $Z \leq z_\alpha$ with z_α such that $P(Z < z_\alpha) = 1 - \alpha$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	Accept H_0 if $Z \geq -z_\alpha$ with z_α such that $P(Z < z_\alpha) = 1 - \alpha$

Case of Small Samples (Unknown Variances):

Assume we have two small samples ($n_1 < 30$) and/or ($n_2 < 30$) drawn randomly and independently from normal populations with unknown variances but assumed to be equal to a common value.

Since the variance is unknown, we first calculate the estimated common variance for the two samples:

$$S_c^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

The test statistic $T_{n_1+n_2-2} = \dfrac{\bar{X}_1 - \bar{X}_2}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ follows a Student's t-distribution.

Null Hypothesis H_0	Alternative Hypothesis H_1	Decision Rules
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	Accept H_0 if $T \in]-t_\alpha, t_\alpha[$ with t_α such that $P(\lvert T\rvert > t_\alpha) = \alpha$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	Accept H_0 if $T \leq t_\alpha$ with t_α such that $P(\lvert T\rvert > t_\alpha) = 2\alpha$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	Accept H_0 if $T \geq -t_\alpha$ with t_α such that $P(\lvert T\rvert > t_\alpha) = 2\alpha$

5.5 Exercices

Exercise 1

A researcher conducted a study on two samples of mice that he captured in two different locations. He obtained the following results:

Sample 01	$n_1 = 50$	$\bar{x}_1 = 51g$	$\sigma^2_{ech1} = 256$
Sample 02	$n_2 = 50$	$\bar{x}_2 = 45g$	$\sigma^2_{ech2} = 144$

Can these mice belong to the same population at a 95% confidence level?

Answer:

This involves comparing the 2 observed means for 2 independent samples. Let μ_1 and μ_2 be the true means in the populations from which the samples are drawn.

1. Null Hypothesis: There is no difference between $\mu_1 = \mu_2$

$$\begin{cases} H_0 & : \quad \mu_1 = \mu_2 \\ H_1 & : \quad \mu_1 \neq \mu_2 \end{cases}$$

2. Under the null hypothesis H_0: $n > 30$

$$\text{The test statistic} \quad Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{51 - 45}{\sqrt{\frac{256}{50} + \frac{144}{50}}} = 2.121$$

   which follows $\mathcal{N}(0, 1)$.

3. The threshold $\alpha = 0.05$, we have $z_\alpha = 1.96$. Since $Z > z_\alpha$, we reject H_0, which means there is a significant difference between the two means. Therefore, these mice belong to two different populations.

Exercise 2

To highlight the possible effect of a drug on heart rate, two groups of 100 subjects each are formed by random selection among patients treated with this drug:

In the first group, a placebo is administered; in the second group, the drug is administered. The estimated means and variances for each group are:

$\bar{x}_1 = 80 \quad \sigma_1^2 = 5$ For the heart rate Y_1 of the control group;

$\bar{x}_2 = 81 \quad \sigma_2^2 = 3$ For the heart rate Y_2 of the treated group.

Conduct a two-tailed test of H_0 ($\mu_1 = \mu_2$) against H_1 ($\mu_1 \neq \mu_2$) with a significance level of 1%.

Exercise 3

The pH (degree of acidity) was measured in two types of chemical solutions A and B. In solution A, six measurements were taken, with a mean pH of 7.52 and an estimated standard deviation of 0.024. In solution B, five measurements were taken, with a mean pH of 7.49 and an estimated standard deviation of 0.032.

Determine if, at a significance level of 0.05, the two solutions have different pH values.

One-Way ANOVA 6

6.1 Introduction

Analysis of variance (ANOVA) aims to study the influence of one or more factors on a quantitative variable. We will focus here on the case where the levels, or modalities, of the factors are fixed by the experimenter. This is referred to as a fixed model.

It involves comparing means for several groups (> 2). It compares the between-group variance (the variance between the different groups: the deviation of group means from the overall mean) to the within-group variance (the sum of fluctuations within each group). If there is no difference between the groups, these two variances are (approximately) equal. Otherwise, the between-group variance will necessarily be larger.

ANOVA can be summarized as a multiple comparison of means from different samples formed by the different modalities of the factors. Thus, the conditions for applying the parametric test for comparing means apply again.

The analysis of variance (ANOVA) can be viewed as a generalization of the Student's t-test.

Consider I independent populations. For $i = 1, 2, ..., I$, we have a sample $\{Y_{i1}, Y_{i2}, ..., Y_{in_i}\}$ of size n_i drawn from a distribution $\mathcal{N}(\mu_i, \sigma^2)$. Note that σ^2 does not vary from one sample to another. We assume equality of variances. Let

$N = n_1 + n_2 + ... + n_I$ be the total size of all samples. From these samples, we would like to test the hypotheses:

$$\begin{cases} H_0 & : \quad \mu_1 = \mu_2 = ... = \mu_I \\ H_1 & : \quad \text{there exists } i \neq i' \text{ such that } \mu_i \neq \mu_{i'} \end{cases}$$

For $i = 1, 2, ..., I$, let the mean in population i be

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik} \quad \text{which follows} \quad \mathcal{N}(\mu_i, \sigma^2/n_i)$$

The overall mean is defined by

$$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^{I} \sum_{k=1}^{n_i} Y_{ik} = \frac{1}{N} \sum_{i=1}^{I} n_i \bar{Y}_{i.} \quad \text{which follows} \quad \mathcal{N}(\mu, \sigma^2/N)$$

where μ is the mean of $\{\mu_1, \mu_2, ..., \mu_I\}$ weighted by $\{n_1, n_2, ..., n_I\}$.

$$\mu = \frac{1}{N} \sum_{i=1}^{I} n_i \mu_i$$

Under H_0, μ is the common value of the μ_i.

6.2 Decomposition of Squares

Consider the following sum

$$SST = \sum_{i=1}^{I} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{..})^2$$

This is the sum of squares of the deviations of the observations Y_{ik} from the overall mean $\bar{Y}_{..}$. It measures the total variability of the observations. Its name SST refers to Sum of Squares Total.

The decomposition of squares is a very important technique in ANOVA. It consists of decomposing SST into two independent terms: SSW or Sum of Squares Within (samples) and SSB or Sum of Squares Between (samples).

The SSW is defined by:

$$SSW = \sum_{i=1}^{I} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{i.})^2 = \sum_{i=1}^{I} (n_i - 1) S_i^2,$$

$$\text{where} \quad S_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_{i.})^2$$

is the variance in population i. The SSW term is thus a sum of terms that each measure the variability within a sample. A crucial assumption of the ANOVA model is the homogeneity of variances. This term is then related to the measurement of this common variability. This will be confirmed by calculations later.

The SSB is defined by:

$$SSB = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2$$

It measures the deviation between the mean in population i and the overall mean. If these deviations are substantial, we tend to favor the alternative hypothesis, that is, the treatment means are different.

Proposition 6.2.1 *The following decomposition is always true:* $SST = SSW + SSB$

6.3 Mean Comparison Test, One-Way ANOVA

The principle of the test statistic is to compare the magnitudes of SSA and SSR, and it is given by:

$$F = \frac{SSB/(I-1)}{SSW/(N-I)}$$

where $(I-1)$ is the number of degrees of freedom of SSB, that is, the number of independent terms in this sum. $(N-I)$ is the number of degrees of freedom of SSW.

1. **Under the null hypothesis H_0, the statistic F follows a Fisher distribution $\mathcal{F}_{I-1,N-I}$**

2. **Risk choice**. We choose the risk α of making an error by rejecting H_0.

3. The rejection of the null hypothesis H_0 at the level α occurs when this quantity is too high, that is, when:

$$F_{obs} = \frac{SSB/(I-1)}{SSW/(N-I)} > \mathcal{F}_{I-1,N-I,\alpha} \quad \text{or} \quad P(F > F_{obs}) \leq \alpha$$

6.4 ANOVA Table

The ANOVA table existed long before computers. It allows summarizing all the calculations necessary to perform the test $H_0 : \mu_1 = \mu_2 = ... = \mu_I$ versus H_1 : there exists $i \neq i'$ such that $\mu_i \neq \mu_{i'}$ when the variance is unknown. This situation is very common in practice. Today, most statistical analysis software directly provides the elements of this table. In the case of one-way ANOVA, this table is presented as follows:

Factor	Sum of Squares	Degrees of freedom	Squares Means	F Statistics	Decision
Between Treatments	SSB	$I-1$	$MSB = \frac{SSB}{I-1}$	$F_{obs} = \frac{MSB}{MSW}$	H_0 or H_1
Within Treatments	SSW	$N-I$	$MSW = \frac{SSW}{N-I}$	***	***
Total	SST	$N-1$	***	***	***

6.5 Tukey-Kramer HSD Test

If we reject the null hypothesis H_0, it means that at least one population differs from the others (on average). In this case, we will seek to compare the means two by two. If there are I populations, this amounts to conducting $I(I-1)/2$ tests. We can do this using the Tukey-Kramer HSD (Honestly Significant Difference) test.

a) Based on these previous samples, we would like to test the multiple hypotheses (one for each pair (i, j)):

$$\begin{cases} H_0 & : & \mu_i = \mu_j \\ H_1 & : & \mu_i \neq \mu_j \end{cases}$$

b) Test statistic for each test

$$T = \frac{\bar{X}_i - \bar{X}_j}{SE\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

Where SE is the residual standard deviation estimated from multiple samples:

$$SE = \sqrt{MSW}$$

We see that this statistic resembles that of the Student's t-test, but the variance is calculated using all observations and not just those from the samples from populations i and j.

Example: A researcher conducted an experiment to compare three types of fertilizers for tomatoes. The following table summarizes the data:

Type of Fertilizer	Sample Size	Mean	Standard Deviation
A	21	6.50	1.25
B	21	4.75	1.05
C	21	5.10	1.40

We have $n_1 = n_2 = n_3 = 21$. Thus, $N = 63$. The overall mean is equal to:

$$\bar{Y}_{..} = \frac{1}{N}\sum_{i=1}^{I} n_i \bar{Y}_{i.} = \frac{(21 * 6.50) + (21 * 4.75) + (21 * 5.10)}{63} = 5.45$$

Next, we calculate SSB and SSW:

$$SSB = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2 = 21(6.50 - 5.45)^2 + 21(4.75 - 5.45)^2 + 21(5.10 - 5.45)^2$$

$$= 36.015$$

$$SSW = \sum_{i=1}^{I}(n_i - 1)S_i^2 = 20(1.25)^2 + 20(1.05)^2 + 20(1.40)^2 = 92.5$$

We obtain the following ANOVA table

Factor	Sum of Squares	Degrees of freedom	Squares Means	F Statistics	Decision
Between Treatments	36.015	2	18.0075	11.6805	H_1
Within Treatments	92.500	60	1.5417	***	***
Total	128.515	62	***	***	***

Example 2: We want to compare the average heights, expressed in meters, of trees from three types of beech forests. We are actually looking to see whether there are significant average differences in tree heights among the three types of forests. We assume that the normality and equality of variances hypotheses are satisfied. The data are provided in the following table:

Type 1	**Type 2**	**Type 3**
23.4	22.5	18.9
24.4	22.9	21.1
24.6	23.7	21.2
24.9	24.0	22.1
25.0	24.4	22.5
26.2	24.5	23.6
26.3	25.3	24.5
26.8	26.0	24.6
26.8	26.2	26.2
26.9	26.4	26.7
27.0	26.7	
27.6	26.9	
27.7	27.4	
	28.5	

There are three modalities, the three types of beech forests. The values related to the 37 locations where height measurements were taken yield the respective means:

$$\bar{Y}_{1.} = 25.97 \qquad \bar{Y}_{2.} = 25.39 \qquad \bar{Y}_{3.} = 23.14 \qquad \text{and} \qquad \bar{Y}_{..} = 24.98$$

Applied to the first observation of the first sample ($x_{1,1} = 23.4$), the observed model of analysis of variance is written as:

$$(23.4 - 24.98) = (25.97 - 24.98) + (23.4 - 25.97) \quad \text{or} \quad -1.58 = 0.99 - 2.57$$

The negative effect of 1.58 m between this particular observation and the overall mean arises from the fact that the location in question belongs to a certain type of forest whose mean is 0.99 m higher than the overall mean, and that this location has a height that is 2.57 m lower than the mean of all observations related to this same type of forest.

A similar calculation can be performed for each of the other 36 trees, and by summing the squares of the deviations obtained, we arrive at the three sums of squares of deviations:

$$SSB = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2 = 13(0.99)^2 + 14(0.41)^2 + 10(1.84)^2 = 48.88$$

$$SSW = \sum_{i=1}^{I} \sum_{k=1}^{n_i} \left(Y_{ik} - \bar{Y}_{i.}\right)^2 = \sum_{i=1}^{I} (n_i - 1) S_i^2 = 116.65$$

This approach is not the one usually followed, as it is often done using software, but it is useful from a didactic perspective to understand the mechanism of analysis of variance.

The ANOVA table below presents the sum of squares of deviations obtained in this way:

Factor	Sum of Squares	Degrees of freedom	Squares Means	F Statistics	Decision
Between Treatments	48.88	2	24.44	7.12	H_1
Within Treatments	116.65	34	3.431	***	***
Total	165.53	36	4.598	***	***

From the previous table, we obtain:

$$F_{obs} = 24.4/3.431 = 7.12 > \mathcal{F}_{I-1,N-I,\alpha} \quad \text{and} \quad P(F > 7.12) = 0.0026 \leq \alpha$$

with 2 and 34 degrees of freedom. The hypothesis of equality of average heights of trees in the three types of beech forests must therefore be rejected, even at the 1% level: the observed differences between the three types of beech forests are highly significant.

The confidence limits for the differences are, for a confidence level of 95%, and for the first two types of forests:

$$25.97 - 25.39 \pm 2.032\sqrt{3.431\left(\frac{1}{13}+\frac{1}{14}\right)} = 0.58 \pm 1.45 = -0.87 \text{ and } 2.03m;$$

for the first and third types of forests:

$$25.97 - 23.14 \pm 2.032\sqrt{3.431\left(\frac{1}{13}+\frac{1}{10}\right)} = 2.83 \pm 1.58 = 1.25 \text{ and } 4.41m;$$

and for the last two types of forests:

$$25.39 - 23.14 \pm 2.032\sqrt{3.431\left(\frac{1}{14}+\frac{1}{10}\right)} = 2.25 \pm 1.56 = 0.69 \text{ and } 3.81m;$$

The fact that the first confidence interval contains zero indicates that there is no significant difference between the first two types of trees, which was already the conclusion we reached in an example from the previous chapter.

6.6 Exercises

Exercise 1: We consider five treatments T1,....,T5 against cold sores, one of which is a placebo (treatment T1). These treatments were randomly administered to thirty patients (six patients per treatment group). The delay, expressed in days, between the appearance of cold sores and complete healing was recorded for each of the thirty patients, detailed below:

T1	T2	T3	T4	T5
5	4	6	7	9
8	6	4	4	3
7	6	4	6	5
7	3	5	6	7
10	5	4	3	7
8	6	3	5	6

Compare the means of the healing delays observed in the five independent samples (treatment groups).

Exercise 2: Fifteen calves were randomly divided into three groups, with the calves in each group receiving a specific diet. The weight gains observed over the same period, expressed in kg, are presented below, with one data point missing:

Diet 1	:	42.1	37.7	45.1	43.1	
Diet 2	:	45.2	54.2	38.1	48.3	55.1
Diet 3	:	48.3	44.1	56.9	42.2	54.0

Can we consider that the observed differences in means between the diets of the three groups are significant?

If so, estimate these mean differences and determine their 95% confidence limits.

Correlation and linear regression 7

7.1 Bivariate Statistical Series

In this chapter, we explore relationships between random variables measured simultaneously. Many experiments aim to uncover connections between such variables. For example:

- Dosage of a drug versus recovery time
- Amount of fertilizer versus plant growth
- Height and weight measurements

We present two main approaches for analyzing these relationships:

1. **Correlation**: This quantifies the strength and direction of a relationship between variables but does not imply causation or dependence of one variable on the other.

1. **Regression**: This models the relationship by establishing how one variable depends on another, providing a predictive framework.

Our data take the form of pairs of observations and present in the form of a table

$$(x_1, y_1), (x_2, y_2), (x_13, y_3), ..., (x_n, y_n).$$

We analyze relationships between variables collected together, meaning (X, Y) forms a bivariate random variable. The observations on each pair are assumed to be independent. These data typically arise from a random sample of n individuals drawn from a population.

Definition 7.1.1 *A **bivariate statistical series** (or **double statistical series**) is a statistical series where two variables are studied simultaneously.*

Remark 7.1.1

1. *In this chapter, we will only study bivariate statistical series in which both variables are quantitative.*
   If, for each of the n individuals in the population, we denote x_i and y_i as the values taken by the two variables, the statistical series can then be presented in the form of a table.

Variable X	x_1	x_2	x_3	.....	x_n
Variable Y	y_1	y_2	y_3	.....	y_n

2. *If one of the two variables represents a measure of time, the series is called a **time series**.*

Definition 7.1.2 *In an orthogonal coordinate system, the set of points M_i with coordinates x_i, y_i forms the **scatter plot** associated with the bivariate statistical series.*

Mean Point

Definition 7.1.3 *The mean point of a scatter plot is a point G with coordinates $\bar{x}, \bar{y}$, where:*
$\bar{x}$ represents the mean of the x_i:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

$\bar{y}$ represents the mean of the y_i:

$$\bar{y} = \frac{y_1 + y_2 + y_3 + ... + y_n}{n} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

The first step to analyze such data is always to draw a scatter diagram.

Example 7.1.1 *Consider the following data for the number of bacteria, in millions, and their ages, in days.*

Age (x)	1	2	3	4	5	6	7	8
No. of bacteria (y)	34	106	135	181	192	231	268	300

For this data set, we are interested in the following questions.

- Is there any relationship between the number of bacteria and and their ages?
- How would you describe this relationship?

Such questions can be more effectively addressed by examining a **scatter plot** of the data. Looking at the scatter plot, we see that

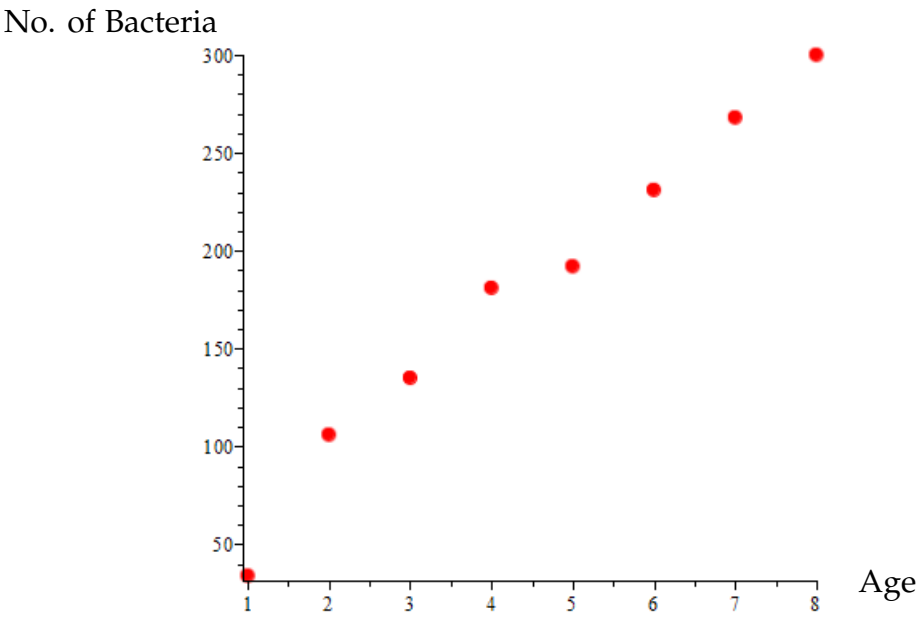


Figure 7.1 – *Scatter Plot of Age vs. Number of Bacteria.*

- The mean point is

$$\bar{x} = \frac{1+2+3+4+5+6+7+8}{8} = 4.5$$

$$\bar{y} = \frac{34+106+135+181+192+231+268+300}{8} = 180.875.$$

So the mean point $G(\bar{x} = 4.5; \bar{y} = 180.875)$.

- as number of bacteria increases, their age also increase – i.e. there is a **positive** relationship between 'Number of Bacteria' and their 'age'.
- It looks like we could draw a straight line through the data – i.e. there is a **linear** relationship.
- There won't be too much scatter around this line, and so this linear relationship is **strong**.
- So the number of bacteria and their age have a **strong, positive, linear** relationship.

7.1.1 Fitting a Bivariate Statistical Series

Fitting y as a function of x for a scatter plot involves finding a function f such that the curve with the equation $y = f(x)$ passes as "closely as possible" to the points in the scatter plot.

Remark 7.1.2 *In the remainder of this chapter, we will focus on linear fits, meaning cases where the statistical series can be approximated by a linear function (although this is not always the case).*

7.2 Fitting using the Least Squares Method

Principle of the Method:

Fitting y as a function of x for a scatter plot using the **least squares method** involves finding the function f within the chosen model that minimizes the sum of the squared differences between the observed values y_i and the predicted values $f(x_i)$ given by the model.
The function f must therefore minimize the following expression:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} M_i P_i^2.$$

Graphical Interpretation: (see Figure 7.2).

This amounts to minimizing the sum of the squares of the vertical distances

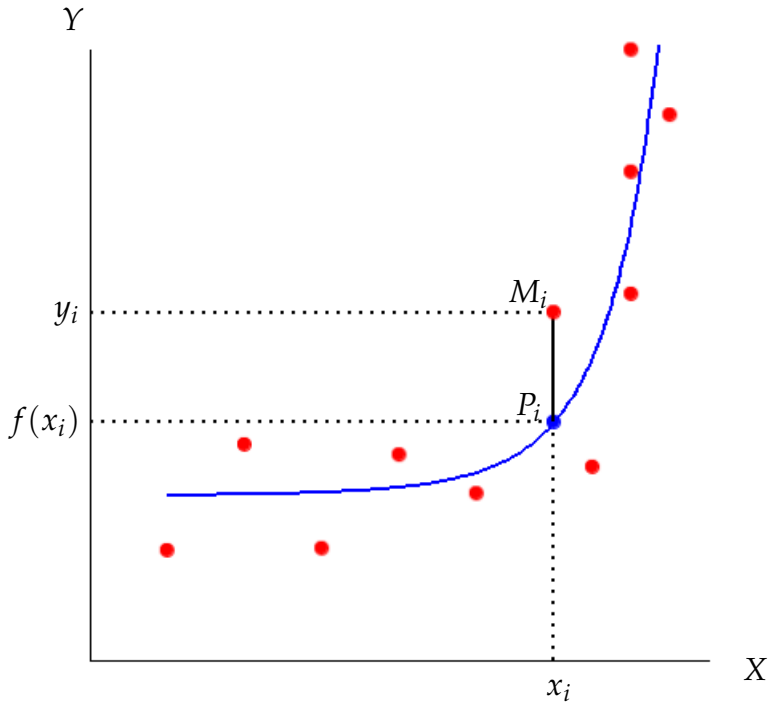


Figure 7.2 – *Least Squares Method.*

between the curve and the points in the scatter plot:

$$(M_1P_1)^2 + (M_2P_2)^2 + (M_3P_3)^2 + ... + (M_nP_n)^2.$$

Remark 7.2.1

1. *For a given value x_0 of the variable x; the function f allows us to* ***predict*** *the corresponding result for the variable y. We assume that $y_0 = f(x_0)$.*
2. *If x_0 lies within the range $[x_1, x_n]$, the process is called* ***interpolation****.*
3. *If x_0 lies outside the observation range of x the process is called* ***extrapolation****.*

7.3 Linear Fit using the Least Squares Method

The linear fit using the least squares method involves finding the best-fitting line for a scatter plot by minimizing the sum of the squared vertical distances between the observed points and the predicted points on the line.

Given a set of points (x_i, y_i) for $i = 1, 2, \ldots, n$, the goal is to find a linear function $y = ax + b$ such that the sum of squared residuals:

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

is minimized.

7.3.1 Steps to Determine the Line $y = ax + b$:

Definition 7.3.1 *The **covariance** of a bivariate statistical series (X, Y), where the variables X and Y are quantitative, is the number denoted as $cov(x, y)$ or σ_{xy}, defined by:*

$$\sigma_{xy} = cov(x, y) = \frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Remark 7.3.1

1. *The covariance can take positive, negative, or zero values.*
2. *When $x_i = y_i$ for all $i = 1, ..., n$, the covariance is equal to the variance.*

Theorem 7.3.1 *The covariance can also be written as:*

$$\sigma_{xy} = cov(x, y) = \left(\frac{1}{N} \sum_{i=1}^{n} x_i y_i \right) - \bar{x}\bar{y}$$

Proof.

$$\begin{aligned} cov(x,y) &= \frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{N}\sum_{i=1}^{n}(x_iy_i - y_i\bar{x} - x_i\bar{y} + \bar{x}\bar{y}) \\ &= \frac{1}{N}\sum_{i=1}^{n}x_iy_i - \frac{1}{N}\sum_{i=1}^{n}y_i\bar{x} - \frac{1}{N}\sum_{i=1}^{n}x_i\bar{y} + \frac{1}{N}\sum_{i=1}^{n}\bar{x}\bar{y} \\ &= \frac{1}{N}\sum_{i=1}^{n}x_iy_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} \\ &= \frac{1}{N}\sum_{i=1}^{n}x_iy_i - \bar{x}\bar{y} \end{aligned}$$

□

Reminder

The variance of the variable x is:

$$V(x) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 = cov(x,x)$$

The variance of the variable y is:

$$V(y) = \frac{1}{N}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \left(\frac{1}{N}\sum_{i=1}^{n}y_i^2\right) - \bar{y}^2 = cov(y,y)$$

It is used to calculate the **standard deviation**: $\sigma(x) = \sqrt{V(x)}$ and $\sigma(y) = \sqrt{V(y)}$.

Theorem 7.3.2 *In a linear fit using the **least squares method**, the line of best fit $y = ax + b$ minimizes the sum of the squared vertical distances between the observed points (x_i, y_i) and the corresponding points on the line. The coefficients a (slope) and b (intercept) are given by:*

1. ***Slope** a:*

$$a = \frac{cov(x,y)}{V(x)}.$$

2. ***Intercept*** *b: y passing the mean point* $G(\bar{x}, \bar{y})$*, so:*

$$b = \bar{y} - a\bar{x}.$$

Remark 7.3.2

1. *These two data points are sufficient to determine an equation of this line (see example).*

2. *This line is also called* ***the regression line of*** y ***on*** x*.*

Example 7.3.1 *We take the previous example again, to calculate the variance and the covariance, we can use the table*

Age (x)	**No. of bacteria** (y)	$x_i y_i$	x_i^2	y_i^2
1	34	34	1	1156
2	106	212	4	11236
3	135	405	9	18225
4	181	724	16	32761
5	192	960	25	36864
6	231	1386	36	53361
7	268	1876	49	71824
8	300	2400	64	90000
$\sum_{i=1}^{8} x_i = 36$	$\sum_{i=1}^{8} y_i = 1447$	$\sum x_i y_i = 7997$	$\sum x_i^2 = 204$	$\sum y_i^2 = 315427$

Table 7.1 – *Calculation of covariance and variance.*

We have already seen that, in this case, $\bar{x} = 4.5$ *and* $\bar{y} = 180.875$.
The covariance is then obtained by the following calculation:

$$\sigma_{xy} = cov(x, y) = \left(\frac{1}{N} \sum_{i=1}^{n} x_i y_i \right) - \bar{x}\bar{y} = \frac{7997}{8} - 4.5 \times 180.875 = 185.69$$

The variance of the variable x is:

$$V(x) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \frac{204}{8} - (4.5)^2 = 5.25$$

The variance of the variable y is:

$$V(y) = \left(\frac{1}{N}\sum_{i=1}^{n} y_i^2\right) - \bar{y}^2 = \frac{315427}{8} - (180.875)^2 = 6712.61$$

The slope of the regression line is therefore:

$$a = \frac{cov(x,y)}{V(x)} = \frac{185.69}{5.25} = 35.37$$

The regression line therefore has an equation of the form $y = 35.37x + b$. Furthermore, it passes through $G(4.5; 180.875)$ therefore:

$$\begin{aligned} 35.37 \times 4.5 + b &= 180.875 \\ 35.37 + b &= 180.875 \\ b &= 180.875 - 35.37 \\ b &= 145.50 \end{aligned}$$

The equation of the regression line is therefore:

$$y = 35.37x + 145.50$$

7.4 Linear correlation coefficient (Pearson's Coefficient)

Definition 7.4.1 *To determine the linear correlation coefficient, also known as Pearson's coefficient, Karl Pearson developed a method that uses the covariance between two variables within a dataset. This widely applied approach is represented by the symbol r. When analyzing the relationship between two variables x and y, Pearson's formula*

measures the degree of correlation as follows:

$$r = \frac{cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{\sigma_{xy}}{\sigma_x.\sigma_y}$$

Interpretation:

1. The correlation coefficient r is a value between -1 and 1 that measures the relationship between two variables x and y. The closer this coefficient is to the extremes -1 or 1, the stronger the linear correlation between the variables. A value close to 1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation.

2. If $r > 0$, the values of y tend to increase as the values of x increase, indicating a positive correlation.

3. If $r > 0$, the values of y tend to decrease as the values of x increase, indicating a negative correlation.

4. If $r = 0$, the variations in x and y are independent, meaning there is no correlation between x and y.

5. If $r = -1$ or $r = +1$, there is a perfect positive or negative correlation between x and y, respectively, meaning that all points (x_i, y_i) lie exactly on the regression line.

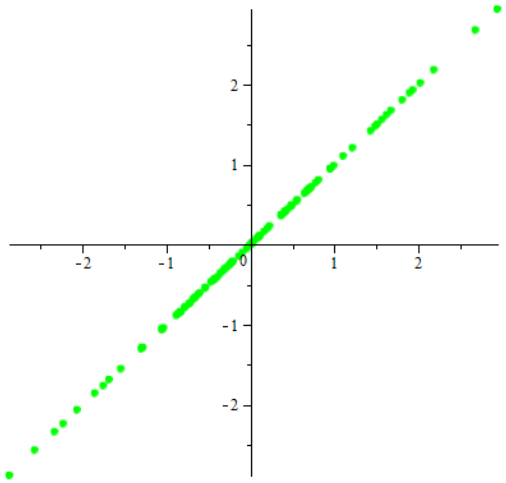
Example 7.4.1 *Calculate the linear correlation coefficient in Example 7.1.1.*
We have: $cov(x,y) = 185.69$, $\sigma_x = \sqrt{5.25} = 2.29$ *and* $\sigma_y = \sqrt{6712.61} = 81.93$.
Therefore

$$r = \frac{cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{\sigma_{xy}}{\sigma_x.\sigma_y} = \frac{185.69}{2.69 \times 81.93} = 0.84,$$

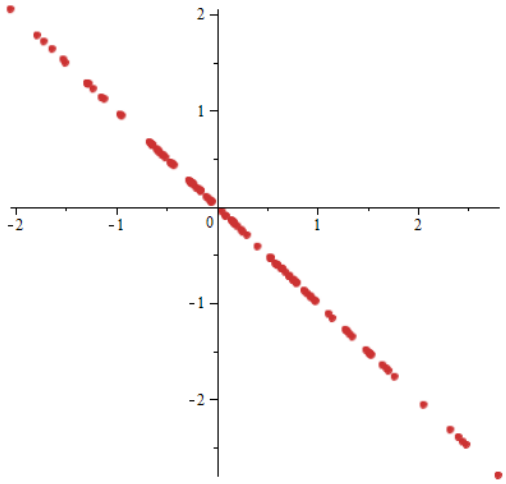
then there is a strong positive correlation between x and y.

7.5 Regression Line of x on y

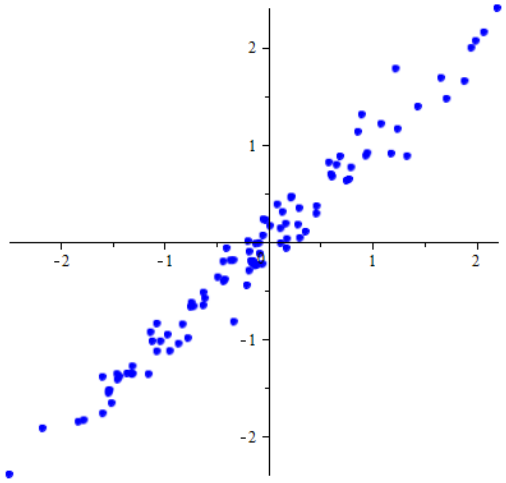
The regression line of x on y is the line that minimizes the sum of the squared horizontal distances between the observed points (x_i, y_i) and the corresponding



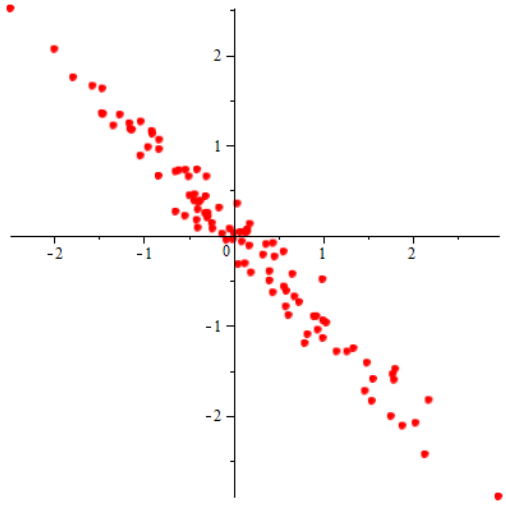
(a) *perfect positive correlation* $r = 1$



(b) *perfect negative correlation* $r = -1$



(c) *Strong Positive Correlation* $r = 0.9$



(d) *Strong Negative Correlation* $r = -0.9$

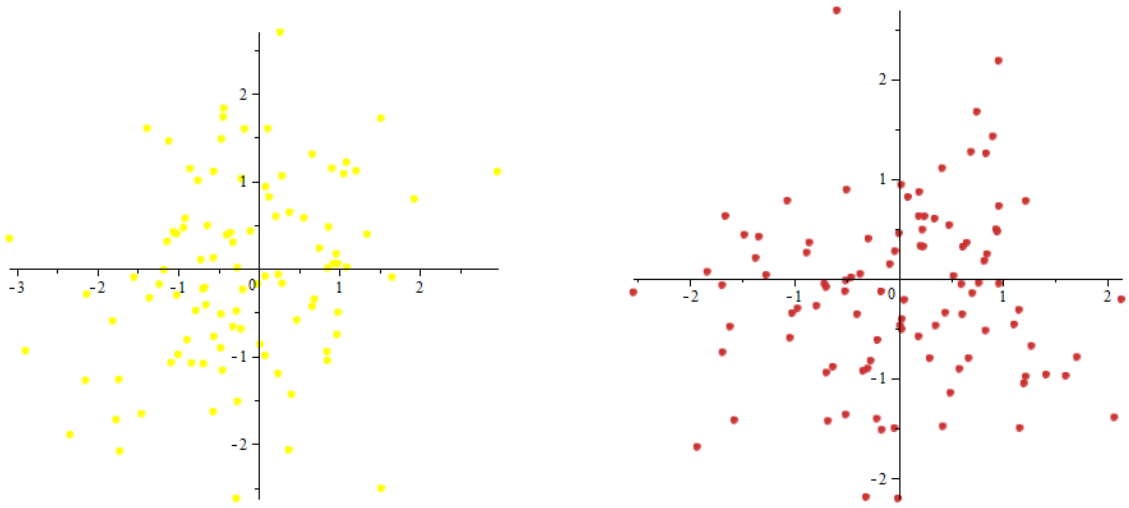
Figure 7.3 – *Scatter Plots with Different Correlation Coefficients*

points on the line. This is the best-fitting line that predicts x as a function of y, written as $x = a'y + b'$.

7.5.1 Steps to Determine the Line $x = a'y + b'$:

To find the regression line of x on y, we follow similar steps as for the regression line of y on x, but focusing on minimizing the horizontal distances instead.

Theorem 7.5.1 *In a linear fit using the* ***least squares method*** *for predicting x from y,*



(a) *Weak Positive Correlation* $r = 0.03$ (b) *Weak Negative Correlation* $r = -0.03$

Figure 7.4 – *Scatter Plots with Weak Correlation Coefficients*

the line of best fit $x = a'y + b'$ *minimizes the sum of the squared horizontal distances. The coefficients* a' *(slope) and* b' *(intercept) are given by:*

1. ***Slope*** a'*:*

$$a' = \frac{cov(x,y)}{V(y)}$$

where $cov(x,y)$ *is the covariance between* x *and* y*, and* $V(y)$ *is the variance of* y*.*

2. ***Intercept*** b'*: The line passes through the mean point* $G(\bar{x},\bar{y})$*, so:*

$$b' = \bar{x} - a'\bar{y}$$

Remark 7.5.1

1. *These two coefficients,* a' *and* b'*, are sufficient to determine the equation of the regression line of* x *on* y*.*

2. *This line is also known as the* ***regression line of*** x ***on*** y *and is used when we want to predict* x *based on values of* y*.*

7.5.2 Calculation of Covariance and Variance (Recap)

Definition 7.5.1 *The **covariance** between two variables X and Y in a dataset with n observations (x_i, y_i) is defined as:*

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

The **variance** of y is:

$$V(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \left(\frac{1}{n} \sum_{i=1}^{n} y_i^2 \right) - \bar{y}^2$$

7.5.3 Example of the Regression Line of x on y

Consider a dataset with points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Let's calculate the regression line of x on y using the formulas above.

Step 1: Calculate the Mean Values.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Step 2: Calculate the Covariance and Variance.

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad V(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Step 3: Determine the Slope a' and Intercept b'.

$$a' = \frac{\text{cov}(x, y)}{V(y)}, \quad b' = \bar{x} - a'\bar{y}$$

The resulting line $x = a'y + b'$ provides the best linear approximation of x in terms of y, minimizing the horizontal distances between the data points and the line.

7.5.4 Interpretation of the Regression Line of x on y

The regression line of x on y is typically used in situations where y is treated as the independent variable, and we aim to predict x based on y. It is especially useful in fields where one variable is naturally dependent on the other, allowing for more accurate predictions in cases where y is known, and we want to estimate x.

7.6 Relationship Between the Regression Lines of y on x and x on y

The regression lines of y on x and x on y are distinct lines with their own slopes and intercepts. These lines are the best linear fits that minimize, respectively, the vertical and horizontal distances between the observed data points and the predicted values.

Given two variables X and Y, the regression line of y on x has the equation:

$$y = ax + b$$

where

$$a = \frac{\text{cov}(x, y)}{V(x)}, \quad b = \bar{y} - a\bar{x}.$$

The regression line of x on y has the equation:

$$x = a'y + b'$$

where

$$a' = \frac{\text{cov}(x, y)}{V(y)}, \quad b' = \bar{x} - a'\bar{y}.$$

7.6.1 Relation Between the Slopes a and a'

The slopes a and a' of the regression lines of y on x and x on y are related by the following identity:

$$a \cdot a' = \frac{\text{cov}(x, y)^2}{V(x)V(y)} = r^2$$

where r is the **correlation coefficient** between X and Y, defined as:

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

with $\sigma_x = \sqrt{V(x)}$ and $\sigma_y = \sqrt{V(y)}$.

This implies that:

$$a \cdot a' = r^2.$$

Remark 7.6.1 *The correlation coefficient r measures the strength and direction of the linear relationship between X and Y. When $r = \pm 1$, the variables X and Y are perfectly linearly related, meaning both regression lines coincide. However, when $|r| < 1$, the two regression lines are distinct.*

7.6.2 Interpretation of the Relationship

The product of the slopes of the two regression lines, $a \cdot a' = r^2$, shows that:

- When $r = 1$ or $r = -1$ (perfect positive or negative correlation), $a \cdot a' = 1$, and the two regression lines coincide. This means that X and Y are perfectly linearly related.
- When $|r| < 1$, $a \cdot a' < 1$, indicating that the two lines do not coincide. Each line represents the best prediction of one variable in terms of the other, minimizing different types of distances (vertical for y on x and horizontal for x on y).

7.6.3 Example

Suppose we have a set of data points (x_i, y_i). By calculating the covariances and variances, we can determine the slopes a and a' for the regression lines. Then, using the correlation coefficient r, we can verify that $a \cdot a' = r^2$.

For instance, if $r = 0.8$, then the product of the slopes will be $0.8^2 = 0.64$. This confirms that the two lines are distinct but related through the correlation coefficient.

Theorem 7.6.1 *The closer r is to ± 1, the closer the two regression lines are to each other.*

When $r = 0$, the regression lines are perpendicular, indicating no linear relationship between X and Y.

This relationship between the two regression lines highlights the importance of the correlation coefficient in understanding the linear association between two variables.

Exercises on Two-Variable Statistics

Exercise 1

During a drought period, a farmer records the amount of water, expressed in m^3, used by their farm starting from day 1, and provides the following data:

X (number of days)	1	3	5	8	10
Y (volume used in m^3)	2.25	4.3	8	17.5	27

1. Calculate the linear correlation coefficient between X and Y. What can be deduced?
2. Provide the equation of the regression line of Y in terms of X.
3. Estimate the volume of water used on the 13th day of drought.

Exercise 2

In the context of research on the duration of the growing season in mountainous areas, meteorological stations are installed at different altitudes, measuring the average temperature (variable Y in degrees Celsius) and altitude (variable X in meters) of 10 stations. The data provided is as follows:

X (altitude in meters)	Data
Y (average temperature in °C)	Data

$\sum_{i=1}^{8} x_i = 19690, \quad \sum_{i=1}^{8} y_i = 20.3, \quad \sum x_i y_i = 17671,$
$\sum x_i^2 = 42925500, \quad \sum y_i^2 = 162.41.$

1. Calculate the linear correlation coefficient between X and Y. What can be deduced?

2. Provide the equation of the regression line of Y in terms of X.

3. What average temperature do you predict at 1100 m? At 2300 m?

Exercise 3

We study air pollution in 41 American cities, with the variable Y representing the volume of SO_2 in the air in micrograms per m^3, and the average annual temperature X in degrees Fahrenheit. The data provided is as follows:

X (average temperature in °F)	Data
Y (SO_2 volume in $\mu g/m^3$)	Data

$$\sum_{i=1}^{8} x_i = 2286, \quad \sum_{i=1}^{8} y_i = 1232, \quad \sum x_i y_i = 74598, \quad \sum x_i^2 = 129549, \quad \sum y_i^2 = 59050.$$

1. Calculate the linear correlation coefficient between X and Y. What can be deduced?

2. Provide the equation of the regression line of Y in terms of X.

3. What value of Y do you predict for a city with an average temperature of 60°F?

Exercise 4

In a sample of 60 individuals within the same age group, the variables height X in meters and shoe size Y were studied. The descriptive statistics obtained are as follows:

Means	$\bar{x} = 1.712$ and $\bar{y} = 41.71$
Standard deviations	$\sigma_x = 0.073$ and $\sigma_y = 1.04$
Covariance	$\sigma_{xy} = 0.066$

1. Calculate the linear correlation coefficient. What can be deduced?

2. Determine the regression line of Y in terms of X.

3. What shoe size can be estimated for an individual measuring 1.80 m?

Bibliography

[1] Ayache, A., & Hamonier, J. *Cours de Statistique Descriptive*, Université de Lille.

[2] Baccini, A. (2010). *Statistique Descriptive Élémentaire*, Institut de Mathématiques de Toulouse.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Springer.

[4] Lebart, L., Piron, M., & Morineau, A. (2006). *Statistique exploratoire multidimensionnelle*, Dunod.

[5] Marc M. Triola Mario F. Triola (2018). *Biostatistics for the Biological and Health Sciences.*, Jason Allen Roy.

[6] Montgomery, D. C. (2020). *Design and Analysis of Experiments.*, John Wiley, Sons.

[7] Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics.*, McGraw-Hill.

[8] Saporta, G. (2011). *Probabilités, analyse des données et statistique.*, Éditions Technip.