PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research

CONSTANTINE 1 FRÈRES MENTOURI UNIVERSITY

FACULTY OF EXACT SCIENCES | DEPARTMENT OF MATHEMATICS

COURSE AND SOLVED EXERCISES

INTRODUCTION TO DESCRIPTIVE STATISTICS

Intended for First-Year Undergraduate Students

in Mathematics and Informatics

Prepared by Dr. Sarra LEULMI

2025/2026

IN LOVING MEMORY OF

MY BELOVED FATHER

PROFESSOR LEULMI SALAH

Table of Contents

I. **Introduction**..... 5

II. **Univariate Descriptive Statistics**..... 6

1. **Basic statistical terms and concepts**: 6

1.1. Population **6**

1.2. Sample **6**

1.3. Statistical unit (individual) **6**

1.4. Character (statistical variable) **6**

1.5. Modalities of a character **6**

1.5.1. **Definition** 6

1.5.2. **Types of characters** 7

Quantitative character 7

1.6. Absolute frequency and relative frequency **8**

1.6.1. **Absolute frequency** 8

1.6.2. **Relative frequency** 9

1.7. **Percentage** 9

1.8. Cumulative Absolute frequency and Cumulative relative frequency **9**

1.8.1. **Cumulative absolute frequency** 9

1.8.2. **Cumulative relative frequency** 9

1.9. Statistical series **10**

1.10. Statistical table **10**

1.11. Graphical representations **12**

1.11.1. **Qualitative variable case** 12

1.11.2. **Quantitative variable case** 14

2.1. Graphical representation of statistical series **14**

2.1.1. **Bar chart (Differential diagram)** 15

2.1.2. **Frequency polygon** 15

2.1.3. **Cumulative chart (integral diagram)** 15

2.2. Characteristics of central tendency **16**

2.2.1. **Mode** 16

2.2.2. Median **16**

2.2.3. **Quartiles** 17

2.2.4. **Quantile** ... 18
2.2.5. **Arithmetic mean** ... 18
2.2.6. **Harmonic mean** ... 19
2.2.7. **Geometric mean** ... 21
2.3. Characteristics of dispersion ... **22**
2.3.1. **Range** ... 22
2.3.2. **Interquartile Range** ... 22
2.3.3. **Variance** ... 23
2.3.4. **Standard deviation** ... 25
2.3.5. **Coefficient of variation** ... 25
3. Study of a continous quantitative variable ... 25
3.1. Statistical table ... **26**
3.2. Graphical representation ... **27**
3.2.1. **Differential chart (Histogram)** ... 27
1.1.1. Polygon ... 28
1.1.2. Integral diagram ... 28
3.3. Characteristics of central tendency ... **29**
3.3.1. **Modal class** ... 29
3.3.2. **Median and quartiles** ... 29
3.3.3. **Arithmetic mean** ... 30
3.3.4. **Harmonic mean** ... 31
3.3.5. **Geometric mean** ... 31
3.4. Characteristics of dispersion ... **32**
3.4.1. **Range** ... 32
3.4.2. **Interquartile range** ... 33
3.4.3. **Variance** ... 33
3.4.4. **Standard deviation** ... 34
3.4.5. **Coefficient of variation** ... 34
4. Series of Exercises 01 ... 35
III. Bivariate Descriptive Statistics ... 69
1. **Introduction** ... 69
2. **Basic statistical terms and concepts** ... 69
2.1. Statistical table ... **69**

2.2. Absolute frequency 69
2.3. Relative Frequency 70
3. **Marginal Distributions** 73
3.1. Marginal absolute frequency 73
3.2. Marginal relative frequency 73
4. **Marginal characteristics** 75
4.1. Marginal mean 75
4.2. Marginal variance 75
4.3. Marginal standard deviation 76
5. **Conditional distribution** 79
5.1. Conditional Relative Frequency 79
6. **Conditional characteristics** 80
6.1. Conditional mean 80
6.2. Conditional variance 80
6.3. Conditional standard deviation 81
1. **Covariance of two characters** 82
1.1. Definition 82
1.2. Properties of the covariance 84
2. **Correlation coefficient** 87
3. **Scatter plot** 91
4. **Regression Line et Mayer Line** 93
4.1. Regression Line 93
4.2. Mayer Line 93
5. **Linear fitting** 95
6. **Regression Curve** 97
7. **Regression Corridor** 97
8. **Functional adjustment** 98
8.1. Fitting the type $Y = c \times dX$ 98
8.2. Fitting the type $X = a \times bY$: 101
9. **Series of Exercises 02** 104
IV. Bibliography 130

I. Introduction

This course presents the fundamental concepts of descriptive statistics and probability theory, in accordance with the curriculum of the first-year undergraduate mathematics program at Constantine 1 University, Frères Mentouri. The course is structured into two main parts.

The first part is devoted to univariate descriptive statistics. It begins with the study of statistical series involving a single variable. We introduce the three types of statistical variables: qualitative, discrete quantitative and continuous quantitative and explain how to distinguish between them. Each type is then examined in detail. In particular, we present both numerical and graphical methods for representing statistical data. For quantitative variables, we also study measures of central tendency and dispersion, including the arithmetic mean, quantiles, mode, variance, and standard deviation. All concepts are illustrated through practical examples.

The second part of the course focuses on statistical series involving two variables. We first study marginal distributions and their main characteristics. Next, we examine conditional distributions, conditional means, and conditional variances. We then define covariance and the correlation coefficient, along with their properties. Finally, we introduce several types of data fitting, such as linear, exponential, and power fitting. As in the first part, numerous examples are provided to illustrate the concepts discussed.

Each part of the course is followed by a set of application exercises, with complete solutions presented at the end of the course.

Univariate Descriptive Statistics

II. Univariate Descriptive Statistics

1. **Basic statistical terms and concepts**:

We begin by introducing some fundamental concepts from descriptive statistics.

1.1. **Population:**

A population refers to the entire collection of individuals, items, or observations that are the focus of a statistical analysis, and it is denoted by the symbol Ω. The total number of units contained in the population is known as the population size, and it is denoted by the symbol *N*.

1.2. **Sample:**

A sample is a smaller group selected from the population, chosen in such a way that it reflects the characteristics of the entire population. Sampling is used when the population is too large to study exhaustively. For results to be reliable, the sample must represent the population accurately and must not favor or exclude any particular trait or subgroup.

1.3. **Statistical unit (individual)**:

Each element or member of a population is called a statistical unit or individual, and it is denoted by the symbol ω.

1.4. **Character (statistical variable):**

A character is a specific feature, attribute, or property of the statistical units that the study aims to examine. It is denoted by the symbols $C_i, Z_i, Y_i, X_i \ldots$
A X character is a function

$$X: \Omega \to C(\Omega)$$
$$\omega \to C(\omega)$$

1.5. **Modalities of a character:**

1.5.1. **Definition:**

The modalities of a character are the different possible values, categories, or states that the character can take.

In summary:
Population: A set (or collection)
Individual: An element of the population.
Characters (or statistical variable): An application.
Modalities of the characters: The images (or outcomes) of the application.

These definitions will be illustrated with examples.

1.5.2. **Types of characters:**

Characters (or variables) in statistics can be grouped into two main types**:** qualitative and quantitative.

Qualitative character:

A qualitative character describes attributes, categories, or qualities. Its modalities are expressed using words or symbols. These characters correspond to categorical variables. Qualitative characters may also be encoded numerically, but the numbers simply represent categories rather than quantities.

Remark:

How to know that a variable is true quantitative ask yourself: Do the mathematical relationships between these numbers have meaning?

Example: The Age

The average between 10 and 20 years old is 15, which is a meaningful value

qualitative variable with numerical modalities: when you see numbers used as labels ranks or codes rather than as measurable quantities you are dealing with numerical modalities.

Example: Bank account number, zip code, phone number, player's jersey number,

Blood group Gender.

Quantitative character:

A quantitative character measures numerical values or counts. It corresponds to numeric variables. Quantitative characters are divided into two subtypes:

Discrete quantitative character (or discrete statistical variable):

Discrete variable can take only distinct, separate, and countable values. Its set of possible values is finite or countably. The modalities are individual numbers.

Continuous quantitative character (or continuous statistical variable):

A continuous variable can take an infinite number of values within a given interval (range). Its set of possible values is infinite and uncountable. The modalities are intervals, often called class intervals.

Example 1: In an Algerian company, we want to determine the education level of 10 of its workers. We found the following table:

Education level	Number
Primary	04
Secondary	03
University	03
Total	10

Explanation:

Population: 10 workers.

Individual: Each worker.

Characters**:** Education level.

Type of characters: Qualitative variable.

Modalities of the character**:** Primary, Secondary, University.

$$C:\ \Omega \to C(\Omega)$$
$$\omega \to C(\omega) \in \{\text{Primary, Secondary, University.}\}$$

Example 2: In a class of 20 students, we want to study the number of books each student read during the holidays. The following table was obtained:

Number of books read (X)	0	1	2	3	4
Number of students	2	5	7	4	2

Explanation:
Population: The 20 students
Individual: Each student
Characters X: Number of books read
Type of characters: Quantitative discrete**.**
Modalities of the character: 0,1,2,3,4.

$$X:\ \Omega \to X(\Omega)$$
$$\omega \to X(\omega) \in \{0,1,2,3,4\}$$

Example 3: In Constantine city, the daily maximum temperature (°C) is recorded over 30 days. Since temperature can take any value within a range, it is a continuous variable. The following table was obtained:

Temperature (°C)	[15, 20[	[20, 25[	[25, 30[	[30, 35[	[35, 40[
Number of days (n_i)	2	5	10	8	5

Explanation:
Population: 30 days.
Individual: Each day.
Characters (X): Daily maximum temperature.
Type: Quantitative continuous.
Modalities (intervals): [15;20[, [20;25[, [25;30[, [30;35[, [35;40[.
Reason for intervals: Continuous variables take many possible values, so we group them into classes for easier analysis (like histograms or frequency tables).

$$Y:\ \Omega \to Y(\Omega)$$
$$\omega \to Y(\omega) \in \{[15;20[,\ [20;25[,\ [25;30[,\ [30;35[,\ [35;40[\}$$

1.6. **Absolute frequency and relative frequency**:

1.6.1. **Absolute frequency:**

Absolute frequency represents the number of occurrences of each modality. It is denoted by n_i where:

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

1.6.2. **Relative frequency**:

Relative frequency is the proportion of the character corresponding to this modality. It is denoted by f_i, where:

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

We can observe that:

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

This makes sense because:

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

1.7. Percentage:

A percentage is a way of expressing a number as a fraction of 100.

$$p_i = f_i \times 100.$$

Note that

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} (f_i \times 100) = 100 \sum_{i=1}^{k} f_i = 100 \times 1 = 100$$

1.8. Cumulative Absolute frequency and Cumulative relative frequency:

1.8.1. Cumulative absolute frequency:

This represents the number of individuals whose variable value is less than or equal to a given value.

It is denoted by the symbol N_i and written as

$$\mathrm{i} \in \{1, 2, \ldots, \mathrm{k}\}, \qquad N_i = card\{\omega : X(\omega) \leq X_i\}$$

and

$$N_i = \sum_{i/X_i \leq x} n_i = \sum_{j=1}^{i} n_j$$

Note that, the first value is n_1, and the last value is N (which represents the sum of all absolute frequencies) and we have

$$\mathrm{i} \in \{1, 2, \ldots, \mathrm{k}\}, \qquad N_i - N_{i-1} = n_i$$

1.8.2. Cumulative relative frequency:

These are obtained by dividing the cumulative frequencies by the total frequency. They are denoted by the symbol F_i and written as

$$F_i = \frac{N_i}{N} = \sum_{j=1}^{i} f_j$$

Where, The first value is f_1, and the last value is **1** (which represents the sum of all relative frequencies) and we have

$$i \in \{1, 2, \ldots, k\}, \quad F_i - F_{i-1} = f_i$$

1.9. Statistical series:

A statistical series is the set of pairs (C_i, n_i), where:
C_i represents a value (modality) of the characters,
n_i is the corresponding frequency**.**

1.10. Statistical table:

The purpose of a statistical table is to collect the raw data of a statistical study in a table, which mainly consists of two columns**:**
The left column shows the modalities of the characters,
The right column contains the absolute frequencies corresponding to these values.
As shown in the following table:

For **discrete quantitative** characters, we write the **values of the variable**, which should be arranged in **ascending order**.

For **continuous quantitative** characters, the column can be written in two ways:

- Record the class intervals inside the cells.

-Write the class boundaries on the lines.

Modalities of the variable	absolute frequencies
C_1	n_1
C_2	n_2
…	…
C_i	n_i
…	…
C_k	n_k
Sum	N

The frequencies corresponding to the values are placed in this column

Notes:

- The statistical table can be presented in row format (see Examples 2 and 3 above).
- Additional columns can be added to the statistical table (or row table) as the statistical work progresses, without changing its title as a statistical table.

Example 4: A survey was conducted on 50 families in a town to find out how many pets each family owns. The collected data are:

1 2 0 1 3 2 1 0 2 1

3 1 2 0 1 2 1 3 1 2
0 1 2 1 3 0 1 2 1 0
1 2 1 3 2 1 0 1 2 1
1 0 2 1 3 1 2 1 0 1

Task: Create a statistical table and calculate relative frequencies.
Step 1: Arrange the data in ascending order

0, 0, 0, 0, 0, 0, 0, 0, 0.
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,1,1,1,1,1,1.
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2.
3, 3, 3, 3, 3, 3.
Step 2: Create the statistical table.

Number of pets (x_i)	Absolute frequency (n_i)	Relative frequency ($f_i = n_i / 50$)
0	9	0.18
1	22	0.44
2	13	0.26
3	6	0.12
Sum	50	1

Example: We study the example 1 mentioned above.
The statistical table:

Education level	Number (n_i)	f_i
Primary	04	4/10=0.4
Secondary	03	3/10=0.3
University	03	3/10=0.3
Total	10	1

Example: We study the example 2 mentioned above.
The statistical table:

X_i	n_i	f_i	N_i	F_i
			0	0
0	2	0.10	2	0.10
1	5	0.25	7	0,35
2	7	0.35	14	0,70
3	4	0.20	18	0,90
4	2	0.10	20	1
Σ	20			
$\frac{\Sigma}{N}$	1			

1.11. **Graphical representations:**

A graphical representation is one of the methods used to describe data. In many practical situations, it is easier to extract information from a graph than by reading a statistical table.
There are many types of graphical representations, and they generally provide similar results, but they depend on the type of characters studied. For example:
For a qualitative characteristics variables: Bar charts and Pie charts.
For a quantitative characteristics variables (discrete or continuous): Differencial diagrams and Cumulative frequency charts (Integral diagrams).

1.11.1. **Qualitative variable case:**

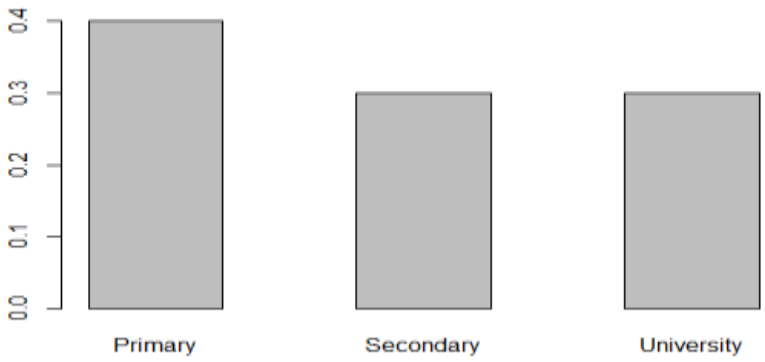
Example: (see Example1)

Statistical table:

Education level	Number (n_i)	f_i
Primary	04	4/10=0.4
Secondary	03	3/10=0.3
University	03	3/10=0.3
Total	10	1

Graphical representations:
In this case, a statistical series can be illustrated using two main types of charts**:** the bar chart and the pie chart**.**

Bar chart:
This type of chart is composed of bars that represent the different modalities of the variable. The height of each bar corresponds to either the absolute frequency or the relative frequency of that modality.

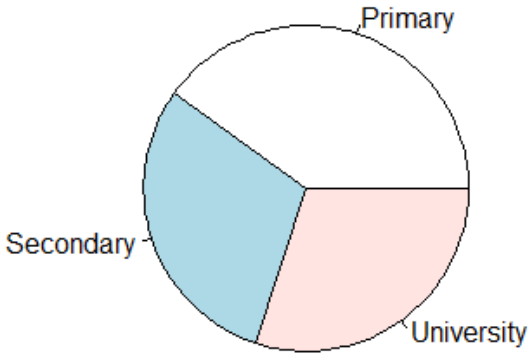


Pie chart:
A pie chart is a circular diagram divided into segments, with each segment representing one modality. The size of each segment is proportional to the relative frequency of the modality and is calculated using the formula

$$\theta_i = f_i \times 360^{\circ}.$$

Statistical table:

Education level	Number (n_i)	f_i	$\theta_i = f_i \times 360^{\circ}$
Primary	04	4/10=0.4	144°
Secondary	03	3/10=0.3	108°
University	03	3/10=0.3	108°
Total	10	1	360°



1.11.2. **Quantitative variable case:**

In this part, a quantitative characteristic (statistical variable), whether discrete or continuous, is studied in four steps:
Statistical table (previously defined).
Graphical representations.
Characteristics of central tendency
Characteristics of dispersion.

2. Study of a discrete quantitative variable:

Let X be a discrete quantitative variable, so,

$$X: \Omega \to X(\Omega) \subset \mathbb{R}$$

where the set of values of the variable

$$X(\Omega) = \{X_1, X_2, \dots, X_k\}$$

with

$$X_1 < X_2 < \cdots < X_k$$

2.1. Graphical representation of statistical series:

In this ease, the statistical series can be represented by two types of graphics: The differential diagram and the integral diagram.

2.1.1. Bar chart (Differential diagram):

A differential or bar chart is made up of vertical bars that represent the values of a variable ($\boldsymbol{X}$). The height of each bar corresponds to its frequency -either absolute ($\boldsymbol{n_i}$) or relative ($\boldsymbol{f_i}$)-.

2.1.2. Frequency polygon:

When the tops of these bars are connected with straight lines, the resulting broken line is called a frequency polygon.

2.1.3. Cumulative chart (integral diagram):

Constructing an integral diagram requires defining the cumulative distribution function.

Empirical cumulative distribution function (ECDF): It is a function defined by

$$F: \mathbb{R} \to \mathbb{R}$$
$$x \to F(x) = \sum_{i/X_i \leq x} f_i$$

Note that

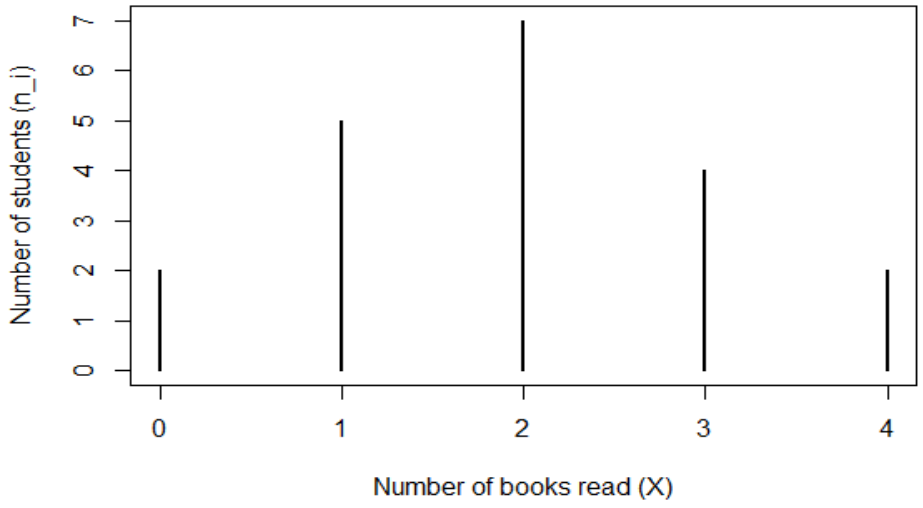
$$\forall x \in \mathbb{R}/\ 0 \leq F(x) \leq 1.$$

The function F is increasing and right-continuous.

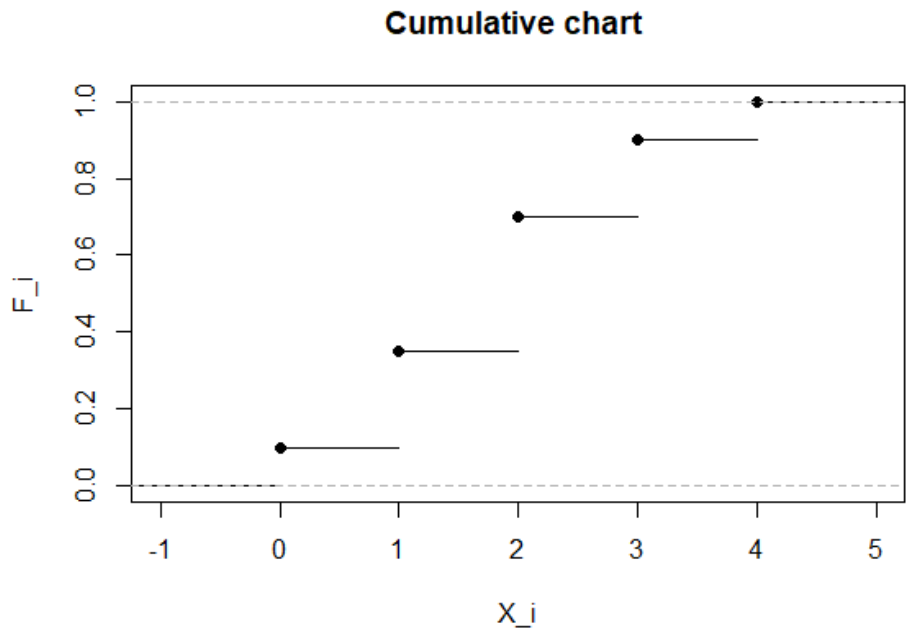
Integral diagram: The graph of the cumulative function is called the cumulative chart, and it appears in the form of a staircase. The length of the step at X_i is equal to f_i.

Example: We study the example 2 mentioned above.

Bar chart:



Integral diagram:



2.2. Characteristics of central tendency:

Measures of central tendency are statistical values that show how data cluster around a central value. They summarize and represent a dataset without replacing the actual data. The main measures include

- ✓ The mode.
- ✓ The median and quartiles.
- ✓ The quantiles
- ✓ The arithmetic mean.
- ✓ The harmonic mean.
- ✓ The geometric mean.

Some of these measures, such as the mode, median, and quartiles, and quantiles, describe the position of data, while others, including the arithmetic mean, not only indicate position but also serve to summarize the entire dataset as a whole.

2.2.1. Mode:

The mode is the value that appears most (obsolute or relative) frequently in a data set. It represents the observation with the highest frequency. It mode is not always unique. Graphically, it corresponds to the tallest bar in the bar chart. It is denoted by M_o.

Example: We study the example 2 mentioned above.

The highest frequency is 7, so the mode is M_o=2.

2.2.2. Median:

The median is the value of the variable X_i that divides the population into two equal parts in terms of the number of individuals. It is denoted by M and satisfies:

$$F(M) = \frac{1}{2}$$

If the total frequency N is odd, the median is the middle value.
If N is even, the median is the arithmetic mean of the two middle values.
In other words, the values of the variable for the first half of the population are less than the median, and for the second half, they are greater. To determine the median, we can use one of the following methods:

- Using the cumulative frequency column:
  Locate $\frac{N}{2}$. The median is the value X_i corresponding to the interval containing $\frac{N}{2}$ between the cumulative absolute frequencies N_{i-1} and N_i.

Example: We study the example 2 mentioned above.

$$\frac{N}{2} = \frac{20}{2} = 10 \rightarrow 7 \leq 10 \leq 14$$
$$\rightarrow M = 2.$$

- Using the cumulative absolute frequency column:
  Locate $\frac{1}{2}$. The median is the value X_i corresponding to the interval containing $\frac{1}{2}$ between the cumulative relative frequencies F_{i-1} and F_i.
  In our example

$$0.35 \leq F(M) = 0.5 \leq 0.7 \rightarrow M = 2.$$

- Using the cumulative curve (graphically):

Since $F(M) = \mathbf{\frac{1}{2}}$, the median corresponds to the point on the cumulative chart where the cumulative frequency reaches *0.5*.

2.2.3. **Quartiles:**

These are values of the variable X_i that divide the population into quarters:
First quartile Q_1: At this value, one-quarter of the population is accumulated.

Second quartile Q_2: At this value, half of the population is accumulated (the median ($Q_2 = M$).

Third quartile Q_3: At this value, three-quarters of the population is accumulated.

To determine these quartiles, we use the same methods used for finding the median:
Example: We study the example 2 mentioned above.
For Q_1

- Using the cumulative absolute frequency column: Determine the position $\frac{N}{4}$.

$$\frac{N}{4} = \frac{20}{4} = 5 \rightarrow 2 \leq 5 \leq 7$$
$$\rightarrow Q_1 = 1.$$

- Using the cumulative relative frequency column: Determine the position $\frac{1}{4}$.

$$0.10 \leq 0.25 \leq 0.70 \rightarrow Q_1 = 1.$$

- Using the cumulative frequency curve (graphically): $F(Q_1) = \frac{1}{4}$.

For Q_3

- Using the cumulative absolute frequency column: Determine the position $\frac{3N}{4}$.

$$\frac{3N}{4} = \frac{3 \times 20}{4} = 15 \rightarrow 14 \leq 15 \leq 18$$
$$\rightarrow Q_3 = 3.$$

- Using the cumulative relative frequency column: Determine the position $\frac{3}{4}$.

$$0.70 \leq 0.75 \leq 0.90 \rightarrow Q_3 = 3.$$

- Using the cumulative frequency curve (graphically): $F(Q_3) = \frac{3}{4}$.

2.2.4. **Quantile:**

Let $p \in [0,1]$, the quantile of order p (or p-th quantile) of X is the value x_p of X which divides the dataset in two parts such that p-proportion of the data are less than or equal to x_p and (1 -p)-proportion of the data are greater than x_p. In other words

$$x_p = inf\{x \in \mathbb{R}, \quad F(x) > p\}$$

where F is the ECDF of X.

Remark: As a particular cases, we have

For $p = 0.5$, $x_{0.5}$ is called the median of X, denoted by M.

For $p = 0.25$, $x_{0.25}$is called the first quartile of X, denoted by Q_1.

For $p = 0.75$, $x_{0.75}$ is called the third quartile of X, denoted bv Q_3.

2.2.5. **Arithmetic mean:**

Definition:

The arithmetic mean is considered one of the most important and widely used statistical measures in practical applications. It is generally defined as the sum of all values divided by their number

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

In general, when the data are grouped in a frequency table:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i X_i = \sum_{i=1}^{k} f_i X_i$$

Indeed,

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i X_i = \sum_{i=1}^{k} \frac{n_i}{N} X_i = \sum_{i=1}^{k} f_i X_i$$

Example: We study the example 2 mentioned above.

According to the statistical table, we have

X_i	n_i	f_i	N_i	F_i	n_iX_i
			0	0	
0	2	0.10	2	0.10	0
1	5	0.25	7	0,35	5
2	7	0.35	14	0,70	14
3	4	0.20	18	0,90	12
4	2	0.10	20	1	8
Σ	20				39
$\frac{\Sigma}{N}$	1				1.95

Thus,

$$\bar{X} = 1.95$$

Properties of the mean:

Let a and b be two real numbers, and X be a statistical variable. If we use a transformation $Y = aX + b$, we have

$$\bar{Y} = a\bar{X} + b$$

Indeed, we have for all $i \in \{1,\dots,k\}$; $Y_i = aX_i + b$, then

$$\begin{aligned}\bar{Y} &= \frac{1}{N}\sum_{i=1}^{k} n_i Y_i \\ &= \frac{1}{N}\sum_{i=1}^{k} n_i(aX_i + b) \\ &= a\frac{1}{N}\sum_{i=1}^{k} n_i X_i + b\frac{1}{N}\sum_{i=1}^{k} n_i \\ &= a\bar{X} + b\frac{1}{N}N \\ &= a\bar{X} + b.\end{aligned}$$

2.2.6. **Harmonic mean:**

The reciprocal of the arithmetic mean of reciprocals of the data. Useful when averaging rates or ratios.

$$\bar{X}_{HM} = \frac{N}{\sum_{k=1}^{N} \frac{1}{X_i}}$$

Or,

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{X_i}}$$

Note that it is important: If $X_i = 0$, the reciprocal $\frac{1}{X_i}$ is undefined. Usually, harmonic mean is only defined for strictly positive data.

Example: Suppose you have three values

$$X = \{2,\ 4,\ 8\}$$

Step 1: Take the reciprocals.

$$\frac{1}{2} = 0.5, \qquad \frac{1}{4} = 0.25, \qquad \frac{1}{8} = 0.125.$$

Step 2: Add the reciprocals.

$$0.5 + 0.25 + 0.125 = 0.875$$

Step 3: Apply the harmonic mean formula.

$$\bar{X}_{HM} = \frac{N}{\sum_{k=1}^{N} \frac{1}{X_i}} = \frac{3}{0.875} = 3.43.$$

Example: The number of phone calls received by a customer-service agent each hour was recorded, and the following data were obtained:

5, 2, 2, 4, 6, 3, 4, 2, 2, 5, 2, 5, 6, 3, 5, 2, 2, 4, 3, 3

X_i	n_i	$\frac{n_i}{X_i}$
2	7	$\frac{7}{2} = 3.5$
3	4	$\frac{4}{3} = 1.33$
4	3	$\frac{3}{4} = 0.75$
5	4	$\frac{4}{5} = 0.8$

6	2	$\frac{2}{6} = 0.333$
Sum	20	6.716

Now, we apply the harmonic mean formula.

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{X_i}} = \frac{20}{6.716} = 2.98$$

2.2.7. **Geometric mean:**

The n-th root of the product of n values. Often used for growth rates, percentages, or ratios.

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{N} X_i}$$

Or,

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} X_i^{n_i}}$$

Example:

Suppose you have three values:

$$X = \{2,\ 4,\ 8\}$$

Step 1: Multiply the values

$$2 \times 4 \times 8 = 64$$

Step 2: Take the cube root (because there are 3 numbers)

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{N} X_i} = \sqrt[3]{64} = 4$$

Example: The number of phone calls received by a customer-service agent each hour was recorded, and the following data were obtained:

5, 2, 2, 4, 6, 3, 4, 2, 2, 5, 2, 5, 6, 3, 5, 2, 2, 4, 3, 3

X_i	n_i	$X_i^{n_i}$
2	7	2^7 =128
3	4	3^4 =81
4	3	4^3 =64
5	4	5^4 =625
6	2	6^2=36
Sum	20	

We have,

$\prod_{i=1}^{k} X_i^{n_i}$ =128×81×64×625×36=14929920000

So, we apply the geometric mean formula.

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} X_i^{n_i}}$$

$$= \sqrt[20]{14929920000}$$

$$= 3.19.$$

Remark:

If the product is very large, we calculate

$\ln(\bar{X}_{GM}) = \frac{1}{N}\sum_{i=1}^{k} n_i \ln(X_i)$.

2.3. Characteristics of dispersion:

Dispersion parameters are statistical measures that indicate how spread out the observations are around a given value.

2.3.1. Range:

It is the difference between the largest value and the smallest value in the data.
$R = \max(X_i) - \min(X_i) = X_k - X_1$.
Example: We study the example 2 mentioned above, we have
$R = \max(X_i) - \min(X_i) = 4 - 1 = 3$.

2.3.2. Interquartile Range:

It is the difference between the third quartile Q_3 and the first quartile Q_1, and we write:

$$I_Q = Q_3 - Q_1.$$

Example: We study the example 2 mentioned above, we have

$$I_Q = Q_3 - Q_1 = 3 - 1 = 2.$$

2.3.3. **Variance:**

Definition: It is one of the most important and widely used measures of dispersion because it relies on all the values of the sample in its calculations, specifically on the deviations of the values from their arithmetic mean.

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,(X_i - \bar{X})^2$$

Or,

$$Var(X) = \sum_{i=1}^{k} f_i\,(X_i - \bar{X})^2$$

Indeed,

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,(X_i - \bar{X})^2$$

$$= \sum_{i=1}^{k} \frac{n_i}{N}(X_i - \bar{X})^2$$

$$= \sum_{i=1}^{k} f_i\,(X_i - \bar{X})^2$$

Note that

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,X_i^2 - \bar{X}^2$$

or

$$Var(X) = \sum_{i=1}^{k} f_i\,X_i^2 - \bar{X}^2$$

Indeed,

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,(X_i - \bar{X})^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} n_i\,(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$$

$$= \frac{1}{N}\sum_{i=1}^{k} n_i\,X_i^2 - 2\bar{X}\frac{1}{N}\sum_{i=1}^{k} n_i\,X_i + \bar{X}^2\frac{1}{N}\sum_{i=1}^{k} n_i$$

As,

$$\tfrac{1}{N}\textstyle\sum_{i=1}^{k} n_i = 1 \quad \text{and} \quad \bar{X} = \tfrac{1}{N}\textstyle\sum_{i=1}^{k} n_i\,X_i$$

Thus,

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,X_i^2 - 2\bar{X}^2 + \bar{X}^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2$$

Note: To calculate $n_i X_i^2$ and $f_i X_i^2$, we can choose one of the two methods:
$n_i X_i^2 = (n_i X_i) X_i$ and $f_i X_i^2 = (f_i X_i) X_i$ (the faster method).
We add a column to calculate X_i^2, then calculate $n_i X_i^2$ $(f_i X_i^2)$ in another column.

Properties of the variance:

Let a and b be two real numbers, and X be a statistical variable. If we use a transformation $Y = aX + b$, we have

$$Var(Y) = a^2 Var(X)$$

Indeed, we have for all $i \in \{1,\dots,k\}$; $Y_i = aX_i + b$, then

$$Var(Y) = \frac{1}{N}\sum_{i=1}^{k} n_i Y_i^2 - \bar{Y}^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} n_i (aX_i + b)^2 - (a\bar{X} + b)^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} n_i (a^2 X_i^2 + 2abX_i + b^2) - (a^2\bar{X}^2 + 2ab\bar{X} + b^2)$$

$$= a^2 \frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 + 2ab\frac{1}{N}\sum_{i=1}^{k} n_i X_i + b^2 \frac{1}{N}\sum_{i=1}^{k} n_i - a^2\bar{X}^2 - 2ab\bar{X} - b^2$$

$$= a^2\left(\frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2\right) + 2ab\bar{X} + b^2 - 2ab\bar{X} - b^2$$

$$= a^2 Var(X).$$

Example: We study the example 2 mentioned above, we have

X_i	n_i	f_i	N_i	F_i	$n_i X_i$	$n_i X_i^2$
			0	0		
0	2	0.10	2	0.10	0	0
1	5	0.25	7	0.35	5	5
2	7	0.35	14	0.70	14	28
3	4	0.20	18	0.90	12	36
4	2	0.10	20	1	8	32

Σ	20				39	101
$\frac{\Sigma}{N}$	1				1.95	5.05

Thus,

$$Var(X) = 5.05 - (1.95)^2 = 1.2475.$$

2.3.4. **Standard deviation:**

Defintion: Since the value of the variance is always positive, the standard deviation is the square root of the variance.

Example: We study the example 2 mentioned above, we have

$$\begin{aligned}\sigma_X &= \sqrt{Var(X)} \\ &= \sqrt{1.2475} \\ &= 1.116.\end{aligned}$$

Properties of the standard deviation:

Let a be a and b two real numbers, and X be a statistical variable. If we use a transformation $Y = aX + b$, we have

$$\sigma_Y = |a|\sigma_X$$

Indeed, we have for all $i \in \{1,\dots,k\}$; $Y_i = aX_i + b$, then

$$\begin{aligned}\sigma_Y &= \sqrt{Var(Y)} \\ &= \sqrt{a^2 Var(X)} \\ &= \sqrt{a^2}\sqrt{Var(X)} \\ &= |a|\sigma_X.\end{aligned}$$

2.3.5. **Coefficient of variation:**

It is a statistical measure of relative variability, calculated as the ratio of the standard deviation to the mean, usually expressed as a percentage:

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100\%$$

It tells you how much variability exists in relation to the average value. A higher CV means more relative variability, while a lower CV means the data points are more consistent.

Example: We study the example 2 mentioned above, we have

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100\% = \frac{1.116}{1.95} \times 100\% = 57.23\%$$

3. Study of a continous quantitative variable:

A continuous quantitative variable is a variable that can take an infinite number of possible values within a given range. The set of values for such a variable is infinite and uncountable. Its values are grouped into intervals called “class intervals”.

$$X: \Omega \to X(\Omega)$$

Where,

$$X(\Omega) = [e_0, e_k[$$

Since the values of the characteristic are given as intervals, we need to recall the width and center of each interval.

The **class width** $[e_{i-1}, e_i[$ is denoted by a_i and is the distance between its boundaries, and we have

$$a_i = e_i - e_{i-1}$$

The **class center** $[e_{i-1}, e_i[$ is denoted by c_i and is the midpoint of the interval, and we have

$$c_i = \frac{e_i + e_{i-1}}{2}$$

3.1. Statistical table:

We can write the statistical table in the following two ways:

First method

Ranges	n_i	f_i
$[e_0, e_1[$	n_1	
$[e_1, e_2[$	n_2	
$[e_2, e_3[$	n_3	
⋮	⋮	⋮
$[e_{k-1}, e_k[$	n_k	f_k

Second method, which is the better method.

e_i	n_i	f_i
e_0		
	n_1	f_1
e_1		
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
e_{k-1}		
	n_k	f_k
e_k		
Σ	N	1
$\frac{\Sigma}{N}$	1	

Example: We study the example 3 mentioned above, we have
Statistical table:

e_i	n_i	f_i	N_i	F_i	c_i	a_i
15			0	0		
	2	0.0667			17.5	5
20			2	0.0667		
	5	0.1667			22.5	5
25			7	0.2334		
	10				27.5	5
30		0.3333	17	0.5667		
	8				32.5	5
35		0.2667	25	0.8334		
	5				37.5	5
40		1		1		
Σ	30		30			
$\frac{\Sigma}{N}$	1					

3.2. **Graphical representation:**

3.2.1. **Differential chart (Histogram):**

On the horizontal axis, each class is represented by a rectangle whose width is a_i and whose height is $\frac{n_i}{a_i}$ or $\frac{f_i}{a_i}$.

Note:

If all classes have the same width, then the height of the rectangle is simply n_i or f_i.

We assume that the distribution within each class is uniform, and we treat the individuals as if the distances between them within each class are equal. Therefore, we can calculate the class centers using

$$c_i = \frac{e_i + e_{i+1}}{2}$$

The area of each rectangle is $a_i \times \frac{n_i}{a_i}$ (or $a_i \times \frac{f_i}{a_i}$), and thus the sum of the areas is:

$$\sum_i a_i \times \frac{n_i}{a_i} = \sum_i n_i = N$$

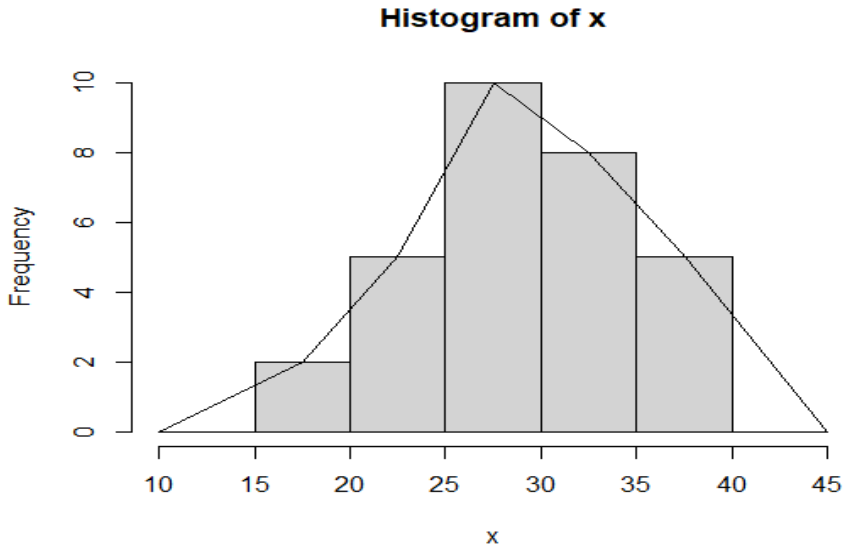
Or

$$\sum_i a_i \times \frac{f_i}{a_i} = \sum_i f_i = 1.$$

1.1.1. Polygon:

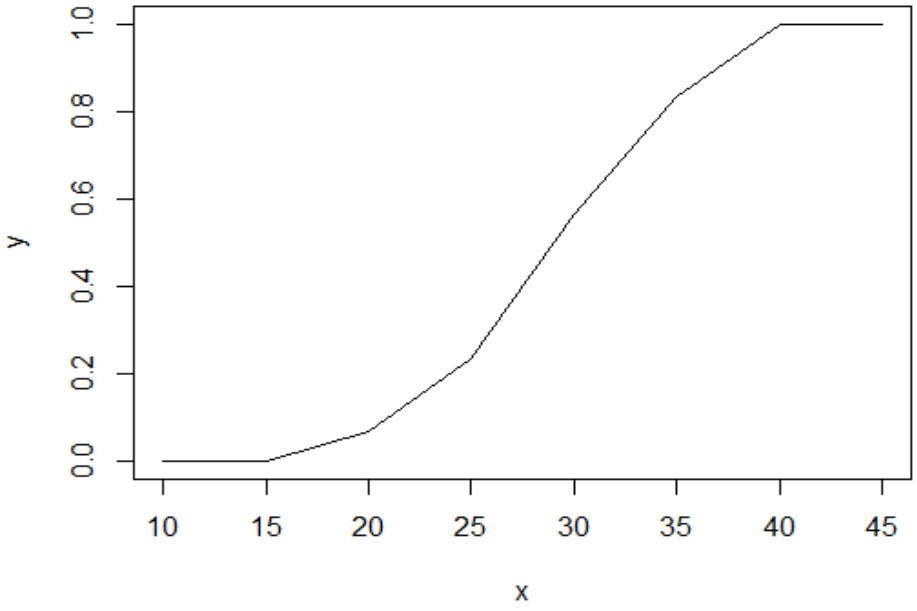
It is a broken line (polygon) made of straight segments connecting points whose coordinates are the center of class (as the x-values) and the cumulative frequencies F_i (as the y-values).

Example: We study the example 3 mentioned above, we have



1.1.2. Integral diagram:

The Integral diagram is the graph of the cumulative function.



3.3. Characteristics of central tendency:

3.3.1. Modal class:

The modal class is the class corresponding to the largest value of $\frac{n_i}{a_i}$ (or $\frac{f_i}{a_i}$). Graphically, the modal class can also be determined from the histogram as the tallest rectangle.
Note: If all classes have the same width, then modal class is the class corresponding to the largest value of n_i or f_i.
Example: The largest value of $n_i = 10$, so the modal class is [25, 30[.

3.3.2. Median and quartiles:

Linear Interpolation Method: This method is used to find the position of a point *Z* or its cumulative frequency *F(Z)*, where $Z \in [e_{i-1}, e_i[$, using Thales' theorem.

Median:

Since *F(M)=0.5*, we look for $M \in [e_{i-1}, e_i[$. Using Thales' theorem

$$\frac{M - e_{i-1}}{e_i - e_{i-1}} = \frac{F(M) - F(e_{i-1})}{F(e_i) - F(e_{i-1})}$$

Since
$F(e_i) - F(e_{i-1}) = f_i, \quad e_i - e_{i-1} = a_i$ and $F(M) = 0.5$
So,

$$M = e_{i-1} + a_i \frac{0.5 - F(e_{i-1})}{f_i}$$

Example: We study the example 3 mentioned above, we have

$F(M) = 0.5 \Rightarrow 0.2334 < 0.5 < 0.5667$

$$\Rightarrow M \in [25, 30[$$
$$\Rightarrow M = 25 + 5\frac{0.5 - 0.2334}{0.3333} = 28.999$$
$$\Rightarrow M \approx 29^{\circ}$$

First quartile Q_1:

$F(Q_1) = 0.25$, and we find Q_1 the same way

$$\frac{Q_1 - e_{i-1}}{e_i - e_{i-1}} = \frac{F(Q_1) - F(e_{i-1})}{F(e_i) - F(e_{i-1})}$$

Thus,

$$Q_1 = e_{i-1} + a_i \frac{0.25 - F(e_{i-1})}{f_i}$$

Example: We study the example 3 mentioned above, we have

$$F(20) = 0.2334 < F(Q_1) = 0.25 < F(30) = 0.5667$$

So,

$$Q_1 \in [20, 25[$$

Thus,

$$Q_1 = 20 + 5\frac{0.25 - 0.2334}{0.3333} = 20.2490$$

Then,

$$Q_1 \approx 20.25^{\circ}$$

Third quartile Q_3:

$F(Q_3) = 0.75$, and we find Q_3 the same way

$$\frac{Q_3 - e_{i-1}}{e_i - e_{i-1}} = \frac{F(Q_3) - F(e_{i-1})}{F(e_i) - F(e_{i-1})}$$

Thus,

$$Q_3 = e_{i-1} + a_i \frac{0.75 - F(e_{i-1})}{f_i}$$

Example: We study the example 3 mentioned above, we have

$$F(30) = 0.5667 < F(Q_3) = 0.75 < F(35) = 0.8334$$

So,

$$30 < Q_3 < 35$$

Thus,

$$Q_3 = 30 + 5\frac{0.75-0.5667}{0.2667} = 33.4375$$

Then,

$$Q_3 \approx 33.44^{\circ}$$

3.3.3. **Arithmetic mean:**

Assuming that the distribution within the classes is uniform, it is justified to take the center of class instead of all the values inside it. The arithmetic mean is then given by

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} n_i c_i = \sum_{i=1}^{n} f_i c_i$$

Example: We study the example 3 mentioned above, we have

Statistical table

e_i	n_i	f_i	N_i	F_i	c_i	a_i	$n_i c_i$
15			0	0			
	2	0.0667			17.5	5	35
20			2	0.0667			
	5	0.1667			22.5	5	112.5
25			7	0.2334			
	10				27.5	5	275
30		0.3333	17	0.5667			
	8				32.5	5	260
35		0.2667	25	0.8334			
	5				37.5	5	187.5
40		1	30	1			
Σ	30						870
$\frac{\Sigma}{N}$	1						29

So,

$$\bar{X} = 29$$

3.3.4. **Harmonic mean:**

The reciprocal of the arithmetic mean of reciprocals of the data. Useful when averaging rates or ratios.

$$\bar{X}_{HM} = \frac{N}{\sum_{k=1}^{N} \frac{1}{c_i}}$$

Or,

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{c_i}}$$

Example: The speeds (in km/h) of cars passing through a checkpoint were recorded over different time intervals. The number of cars and their average speeds in each interval are shown below:

$[e_{i-1}, e_i[$	n_i	c_i	$\frac{n_i}{c_i}$
[0,3[	4	2	$\frac{4}{2} = 2$
[3,5[	2	4	$\frac{2}{4} = 0.5$
[5,7[	1	6	$\frac{1}{6}$ =0.1667
[7,9[	1	7	$\frac{1}{7}$ =0.1429
[9,11[	2	8	$\frac{2}{8} = 0.25$
Sum	10		3.0596

Now, we apply the harmonic mean formula.

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{c_i}} = \frac{10}{3.0596} = 3.267.$$

3.3.5. **Geometric mean:**

The n-th root of the product of n values. Often used for growth rates, percentages, or ratios.

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{N} c_i}$$

Or,

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} c_i^{n_i}}$$

Example: The speeds (in km/h) of cars passing through a checkpoint were recorded over different time intervals. The number of cars and their average speeds in each interval are shown below:

$[e_{i-1}, e_i[$	n_i	c_i	$c_i^{n_i}$
[0,3[	4	2	$2^4 = 16$
[3,5[	*2*	4	$4^2 = 16$
[5,7[	*1*	6	$6^1 = 6$
[7,9[	1	7	$7^1 = 7$
[9,11[	2	8	$8^2 = 64$
Sum	10		

We have,

$$\prod_{i=1}^{k} c_i^{n_i} = 16\times16\times6\times7\times64 = 688128.$$

So, we apply the geometric mean formula.

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} c_i^{n_i}}$$

$$= \sqrt[10]{688128}$$

$$= 3.84$$

Remark:

If the product can be huge, we calculate

$$\ln(\bar{X}_{GM}) = \frac{1}{N}\sum_{i=1}^{k} n_i \ln(c_i).$$

3.4. Characteristics of dispersion:

Dispersion parameters are statistical measures that indicate how spread out the observations are around a given value.

3.4.1. Range:

The range is defined by

$$R = e_k - e_0$$

Example: We study the example 3 mentioned above, we have

$$R = e_k - e_0 = 40 - 15 = 25$$

3.4.2. **Interquartile range:**

It is the difference between the third quartile Q_3 and the first quartile Q_1, and we write:

$$I_Q = Q_3 - Q_1$$

Example: We study the example 3 mentioned above, we have

$$I_Q = Q_3 - Q_1 = 33.44 - 25.25 = 8.19^{\circ}.$$

3.4.3. **Variance:**

It is one of the most important and widely used measures of dispersion because it relies on all the values of the sample in its calculations, specifically on the deviations of the values from their arithmetic mean.

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,(c_i - \bar{X})^2$$

or

$$Var(X) = \sum_{i=1}^{k} f_i\,(c_i - \bar{X})^2$$

Note that

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\,c_i^2 - \bar{X}^2$$

Or,

$$Var(X) = \sum_{i=1}^{k} f_i\,c_i^2 - \bar{X}^2$$

Example: We study the example 3 mentioned above, we have

e_i	n_i	f_i	N_i	F_i	c_i	a_i	$n_i c_i$	$n_i c_i^2$
15			0	0				
20	2	0.0667	2	0.0667	17.5	5	35	612.50
25	5	0.1667	7	0.2334	22.5	5	112.5	2 531.25
30	10	0.3333	17	0.5667	27.5	5	275	7 562.50
35	8	0.2667	25	0.8334	32.5	5	260	8 450.00
40	5	1	30	1	32.5	5	187.5	7 031.25
Σ	30						870	26 187.50
$\frac{\Sigma}{N}$	1						29	872.9166

Thus,

$$Var(X) = 872.9166 - (29)^2 = 31.9166.$$

3.4.4. **Standard deviation:**

The standard deviation is given by

$$\sigma_X = \sqrt{Var(X)}$$

Example: We study the example 3 mentioned above, we have

$$\sigma_X = \sqrt{Var(X)} = \sqrt{31.9166} = 5.648.$$

3.4.5. **Coefficient of variation:**

It is a statistical measure of relative variability, calculated as the ratio of the standard deviation to the mean, usually expressed as a percentage:

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100\%$$

Example: We study the example 2 mentioned above, we have

$$\begin{aligned} \text{CV} &= \frac{\sigma_X}{\bar{X}} \times 100\% \\ &= \frac{5.648}{29} \times 100\% \\ &= 19.48\,\% \end{aligned}$$

4. Series of Exercises 01:

Exercise 01: For every example, specify the population, the statistical unit, the studied character and its type.
Number of students in a class, Height of a person, Eye color, Age of a tree (in years), Type of car (SUV, sedan, truck), Temperature of a room, Bank account number , Number of pets owned, phone number, Number of children in a family.

Solution 01:

1. The study of the number of students in a class:

Population: All students in the school.
Statistical unit: Each class.
Studied character: Number of students in each class.
Type: Quantitative-Discrete.

2. The study of the height of a person:

Population: All people in the group or community being studied.
Statistical unit: Each person.
Studied character: Height of a person.
Type: Quantitative-Continuous.

3. The study of the eye color of a person:

Population: All people observed.
Statistical unit: Each person.
Studied character: Eye color.
Type: Qualitative.

4. The study of the age of a tree in years:

Population: All trees in a forest or area.
Statistical unit: Each tree.
Studied character: Age of the tree (in years).
Type: Quantitative-Discrete.

5. The study of the type of car (SUV, sedan, or truck):

Population: All cars in a parking lot or city.
Statistical unit: Each car.
Studied character: Type of car.
Type: Qualitative.

6. The study of the temperature of a room:

Population: All rooms in a building or house.
Statistical unit: Each room.
Studied character: Temperature of the room.
Type: Quantitative-Continuous.

7. The study of the bank account number of customers:

Population: All customers of a bank.
Statistical unit: Each customer.
Studied character: Bank account number.
Type: Qualitative.

8. The study of the number of pets owned by a person:

Population: All people in a community.
Statistical unit: Each person.
Studied character: Number of pets owned.
Type: Quantitative-Discrete.

9. The study of the phone number in each house of a neighborhood:

Population: All houses in the neighborhood.
Statistical unit: Each house.
Studied character: Phone number of the house.
Type: Qualitative.

10. The study of the number of children in a family:

Population: All families in the community.
Statistical unit: Each family.
Studied character: Number of children per family.
Type: Quantitative – Discrete.

Exercise 02:

A survey asked 50 households which type of pet they own.. The results of this survey are given in the following table.

Type of Pet	Number of Households
Cat	14
Bird	6
Dog	22
Fish	5
No Pet	3

1. Determine the population, the studied character, its type and its modalities,
2. Construct the statistical table with absolute and relative frequencies.
3. Draw the appropriate graphical representations.

Solution 02:

1. Identify the qualitative variable:

Population: 50 households.
Qualitative variable: Type of pet owned.
Type: Qualitative nominal.
Modalities: Dog, Cat, Bird, Fish, No Pet.

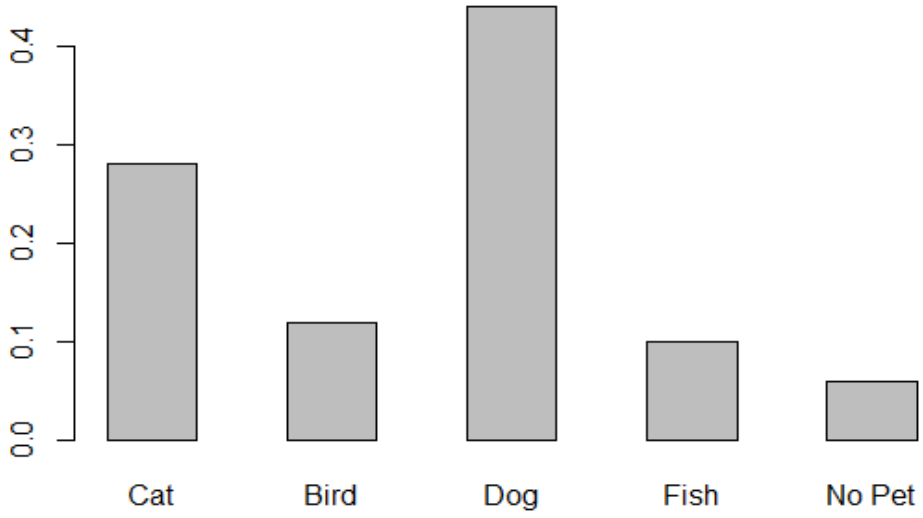
2. Statistical table with absolute and relative frequencies:

Type of Pet	Number of Households (n_i)	f_i	$\theta_i = f_i \times 360^\circ$
Cat	14	0.28	100.8°
Bird	6	0.12	43.2°
Dog	22	0.44	158.4°
Fish	5	0.10	36°
No Pet	3	0.06	21.6°
Σ	50	1	360°

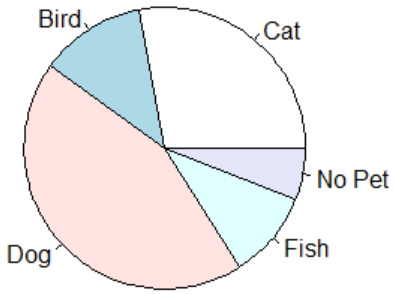
3. Draw the appropriate graphical representations:

In the qualitative ease, the appropriate graphieal representations are the bar chart and the pie chart.

Bar chart:



Pie chart:



Exercise 03:

A survey was done among 100 students about their favorite school subject.

Subjects	Number of steudents
Mathematics	28
Science	22
English	18
History	12
Physical Ed.	20

1. Determine the population, the studied character, its type and its modalities,
2. Construct the statistical table with absolute and relative frequencies.
3. Draw the appropriate graphical representations.

Solution 03:

1. Identify the qualitative variable:

Population: All 100 students surveyed.
Studied character (variable): The favorite school subject.
Type of variable: Qualitative nominal**,** because the answers are categories (names).
Modalities: (categories) Mathematics, Science, English, History, Physical Education.

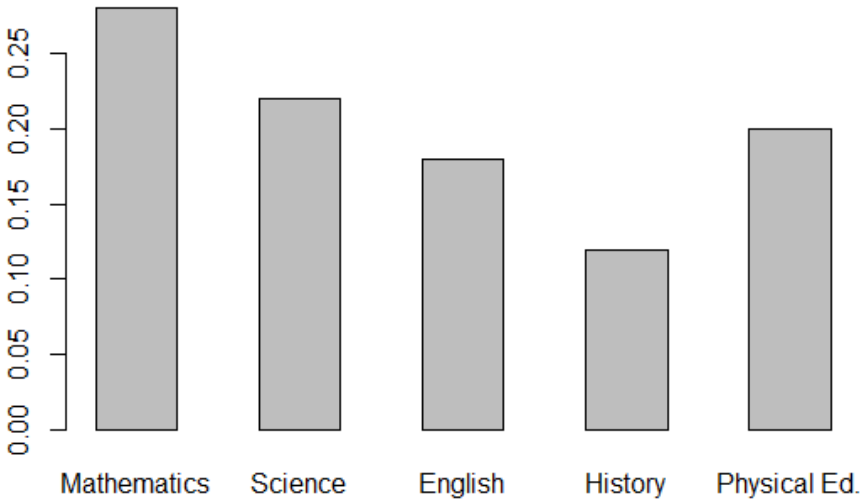
2. Statistical table with absolute and relative frequencies:

Subjects	n_i	f_i	$\theta_i = f_i \times 360^\circ$
Mathematics	28	0.28	100.8°
Science	22	0.22	79.2°
English	18	0.18	64.8°
History	12	0.12	43.2°
Physical Ed.	20	0.20	72°
Σ	100	1	360°

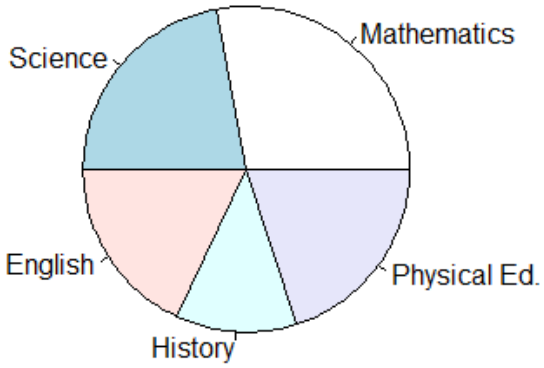
3. Draw the appropriate graphical representations:

In the qualitative ease, the appropriate graphieal representations are the bar chart and the pie chart.

Bar chart:



Pie chart:



Exercise 04: A coach surveyed his team and found the following distribution for weekly exercise hours. The mean is known to be 3.4 hours. Find the missing frequency a.

Hours exercised (x)	1	2	3	4	5	6
Number of players	3	5	a	7	4	2

Solution 04:
We have the mean

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} n_i x_i = 3.4.$$

Or,

$$\sum_{i=1}^{n} n_i x_i = 1 \times 3 + 2 \times 5 + 3 \times \mathrm{a} + 4 \times 7 + 5 \times 4 + 6 \times 2 = 73 + 3\mathrm{a}$$

and

$$N = 3 + 5 + \mathrm{a} + 7 + 4 + 2 = 21 + \mathrm{a}.$$

Thus,

$$\frac{1}{N}\sum_{i=1}^{n} n_i x_i = 3.4 \Rightarrow \frac{73 + 3\mathrm{a}}{21 + \mathrm{a}} = 3.4$$

$$\Rightarrow 73 + 3\mathrm{a} = 3.4(21 + \mathrm{a})$$

$$\Rightarrow 1.6 = 0.4\mathrm{a}$$

$$\Rightarrow \mathrm{a} = \frac{1.6}{0.4} = 4$$

Exercise 05:

A survey was conducted among 30 employees to find out how many cups of coffee they drink per day. The results are summarized in the following table:

X_i	0	1	2	3	4	5
f_i	0.1	0.2	0.3	0.25	0.05	0.1

1. Determine the population, its size, the character studied and its type.
2. Construct the appropriate statistical table.
3. Represent the data with the appropriate graphs.
4. Determine the percentage of employees who drink 2 cups, and the percentage of employees who drink less than 3 cups**.**
5. Calculate the characteristics of central tendency.
6. Calculate the characteristics of dispersion**.**
7. Caculate $\bar{Y}$ and σ_X using the transformation $Y = 4X - 5$.

Solution 05:

1. Determine the population and its size:

Population: The 30 employees surveyed.
Population size (N): 30.

Determine the character studied and its type:

Character studied: Number of cups of coffee consumed per day.
Type: Quantitative discrete (because the number of cups is countable and takes integer values).

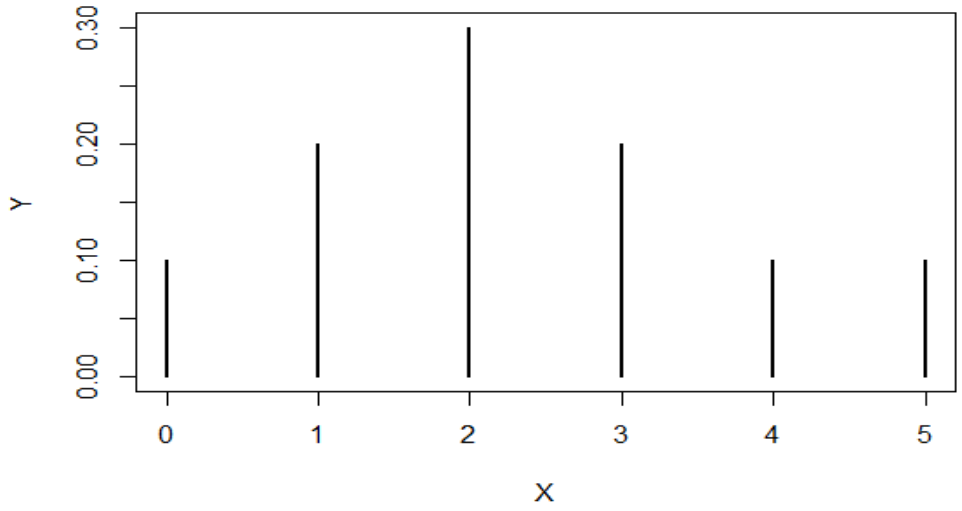
2. Construct the statistical table:

X_i	f_i	n_i	N_i	F_i	n_iX_i	$n_iX_i^2$
			0	0		
0	0.10	3	3	0.10	0	0
1	0.20	6	9	0.30	6	6
2	0.30	9	18	0.60	18	36
3	0.20	6	24	0.80	18	54
4	0.10	3	27	0.90	12	48
5	0.10	3	30	1	15	75

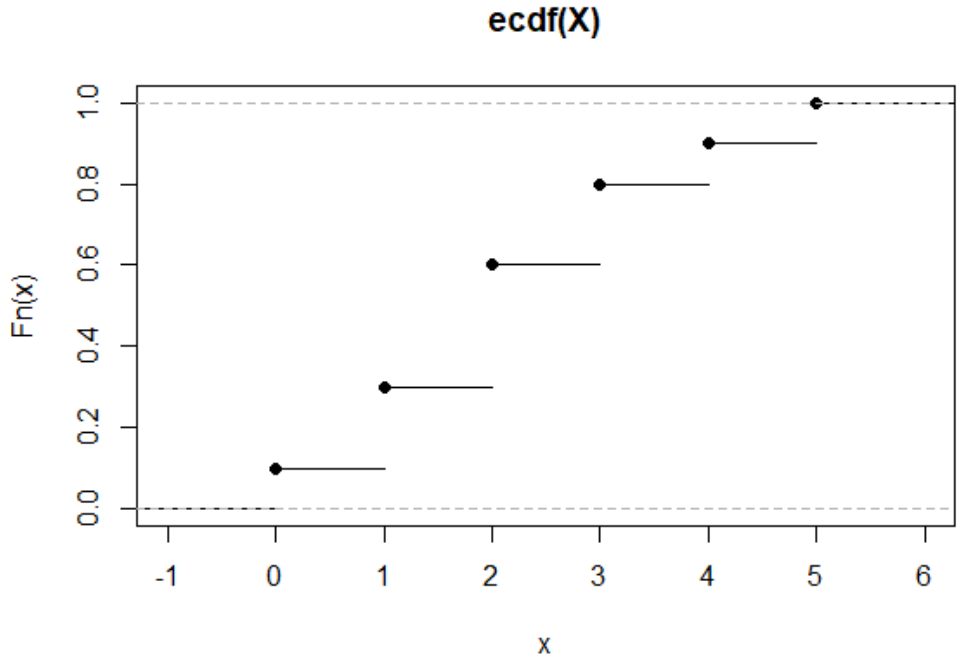
Σ	1	30			69	219
$\frac{\Sigma}{N}$		1			2.3	7.3

3. Represent the data with the appropriate graphs:

Bar chart:



Cumulative frequency curve:



4. The percentages:

The percentage of employees who drink 2 cups

$$\text{Percentage} = f_3 \times 100 = 0.3 \times 100 = 30\%.$$

The percentage of employees who drink less than 3 cups.

$$\text{Percentage} = (f_1 + f_2 + f_3) \times 100 = (0.10 + 0.20 + 0.30) \times 100 = 0.60 \times 100 = 60\%.$$

5. Characteristics of central tendency:

Mode: The highest frequency is 9, so the mode is M_o=2.

Median:

- Using the cumulative frequency column: Determine the position $\frac{N}{2}$.

$$\frac{N}{2} = \frac{30}{2} = 15 \Rightarrow 9 \leq 15 \leq 28$$
$$\Rightarrow M = 2.$$

- Using the cumulative absolute frequency column: Determine the position $\frac{1}{2}$.

$$0.30 \leq F(M) = 0.5 \leq 0.6 \rightarrow M = 2.$$

Quartiles: These are values of the variable X_i that divide the population into quarters:

First quartile Q_1:

- Using the cumulative absolute frequency column: Determine the position $\frac{N}{4}$.

$$\frac{N}{4} = \frac{30}{4} = 7.5 \Rightarrow 3 \leq 7.5 \leq 9$$
$$\Rightarrow Q_1 = 1.$$

- Using the cumulative relative frequency column: Determine the position $\frac{1}{4}$.

$$0.10 \leq 0.25 \leq 0.30 \Rightarrow Q_1 = 1.$$

Third quartile Q_3

- Using the cumulative absolute frequency column: Determine the position $\frac{3N}{4}$.

$$\frac{3N}{4} = \frac{3 \times 30}{4} = 22.5 \Rightarrow 18 \leq 22.5 \leq 24$$
$$\Rightarrow Q_3 = 3.$$

- Using the cumulative relative frequency column: Determine the position $\frac{3}{4}$.

$0.60 \leq 0.75 \leq 0.80 \Rightarrow Q_3 = 3.$

Arithmetic mean:

From the statistical table, we have

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i X_i = \sum_{i=1}^{k} f_i X_i = 2.3$$

6. Characteristics of dispersion:

Range:

$$R = \max(X_i) - \min(X_i) = 5 - 0 = 5$$

Interquartile Range:

$$I_Q = Q_3 - Q_1 = 3 - 1 = 2.$$

Variance:

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2 = \sum_{i=1}^{k} f_i X_i^2 - \bar{X}^2$$
$$= 7.3 - (2.3)^2$$
$$= 7.3 - 5.29$$
$$= 2.01.$$

Standard deviation:

$$\sigma_X = \sqrt{Var(X)}$$
$$= \sqrt{2.01}$$
$$= 1.417$$

Coefficient of variation:

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100 = \frac{1.417}{2.3} \times 100 = 61.61\%$$

7. Caculate $\bar{Y}$ and σ_X using the transformation $Y = 4X - 5$:

Since, $Y = 4X - 5$. So,

$$\bar{Y} = 4\bar{X} - 5 = 4 \times 2.3 - 5 = 4.2$$

and

$$\sigma_Y = 3\sigma_X = 4 \times 1.417 = 5.668$$

Exercise 06:

We want to study the fruit production of a certain variety of strawberries. 150 strawberry bushes were selected in a field. The number of fruits on each bush was counted. The following graph was obtained:



1. What does this chart represent? What is the type of the studied characteristic?
2. Construct the statistical table.
3. Calculate the characteristics of central tendency.
4. Calculate the characteristics of dispersion.
5. Represent the data using another appropriate graph.

Solution 06:

1. The chart represents the number of fruits produced by each of 150 strawberry bushes in a field.
   Type of the studied characteristic: The characteristic is "number of fruits per bush", so, this is a quantitative discrete variable.

2. Statistical table:

X_i	n_i	f_i	N_i	F_i	n_iX_i	$n_iX_i^2$	$\frac{n_i}{X_i}$	$n_i \ln X_i$
			0	0				
8	10	0.0625	10	0.0625	80	640	1.25	20.794
10	20	0.125	30	0.1875	200	2000	2	46.052
16	40	0.25	70	0.4375	640	10240	2.5	110.904
20	20	0.125	90	0.5625	400	8000	1	59.914
32	30	0.1875	120	0.75	960	30720	0.9375	103.971
42	40	0.25	160	1	1680	70560	0.9524	149.508
Σ	160	1			3960	121160	8.6399	491.143
$\frac{\Sigma}{N}$	1				24.75	757.25		

3. Characteristics of central tendency:

Mode: The highest frequency is 40, so the modes are M_o=16 and M_o=42.

Median:

- Using the cumulative frequency column: Determine the position $\frac{N}{2}$.

$$\frac{N}{2} = \frac{160}{2} = 80 \Rightarrow 70 \leq 80 \leq 90$$
$$\Rightarrow M = 20.$$

- Using the cumulative absolute frequency column: Determine the position $\frac{1}{2}$.

$$0.4375 \leq F(M) = 0.5 \leq 0.5625 \Rightarrow M = 20.$$

Quartiles: These are values of the variable X_i that divide the population into quarters:

First quartile Q_1:

- Using the cumulative absolute frequency column: Determine the position $\frac{N}{4}$.

$$\frac{N}{4} = \frac{160}{4} = 40 \Rightarrow 30 \leq 40 \leq 70$$
$$\Rightarrow Q_1 = 16.$$

- Using the cumulative relative frequency column: Determine the position $\frac{1}{4}$.

$$0.1875 \leq 0.25 \leq 0.4375 \Rightarrow Q_1 = 16.$$

Third quartile Q_3:

- Using the cumulative absolute frequency column: Determine the position $\frac{3N}{4}$.

$$\frac{3N}{4} = \frac{3 \times 160}{4} = 120$$

120 appears exactly in the column N_i of the statistical table, $Q_3 \in [32,42[$. So, we take the smallest value, Thus

$$Q_3 = 32.$$

- Using the cumulative relative frequency column: Determine the position $\frac{3}{4}$.

0.75 appears exactly in the column F_i of the statistical table, $Q_3 \in [32,42[$. So, we take the smallest value, Thus

$$Q_3 = 32.$$

Arithmetic mean:
From the statistical table, we have

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i X_i = \sum_{i=1}^{k} f_i X_i = 24.75.$$

Harmonic mean:
From the statistical table, we have

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{X_i}} = \frac{160}{8.6399} = 18.52.$$

Geometric mean:
From the statistical table, we have

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} X_i^{n_i}}$$

Since the product can be huge, we calculate

$$\ln(\bar{X}_{GM}) = \frac{1}{N}\sum_{i=1}^{k} n_i \ln(X_i)$$

So,

$$= \frac{491.143}{160}$$
$$= 3.0696$$

Then,

$$\bar{X}_{GM} = e^{3.0696} = 21.52$$

4. Characteristics of dispersion:

Range:

$$R = \max(X_i) - \min(X_i) = 42 - 8 = 34$$

Interquartile Range:

$$I_Q = Q_3 - Q_1 = 32 - 16 = 16.$$

Variance:

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 - \bar{X}^2 = \sum_{i=1}^{k} f_i X_i^2 - \bar{X}^2$$
$$= 757.25 - (24.75)^2$$
$$= 757.25 - 612.5625$$
$$= 144.6875$$

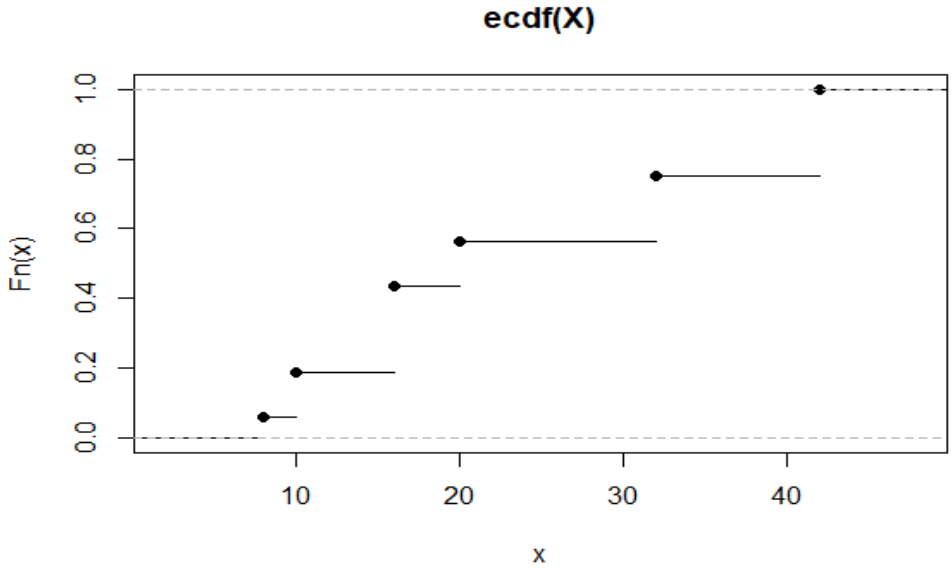
Standard deviation:

$$\sigma_X = \sqrt{Var(X)}$$
$$= \sqrt{144.6875}$$
$$= 12.03$$

Coefficient of variation:

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100 = \frac{12.03}{24.75} \times 100 = 48.6\%$$

5. Represent the data using another appropriate graph:



Exercise 07:

Suppose the following data represent the number of hours spent studying per week by 80 students:

82, 95, 88, 73, 91, 102, 85, 97, 79, 84, 87, 93, 90, 89, 92, 86, 94, 101, 78, 80, 105, 77, 99, 96, 83, 100, 88, 84, 110, 98, 103, 81, 107, 85, 102, 95, 99, 87, 100, 91, 96, 89, 105, 94, 97, 108, 93, 86, 111, 100, 115, 90, 104, 92, 97, 101, 85, 109, 98, 87, 105, 92, 100, 103, 88, 106, 95, 101, 99, 108, 102, 107, 94, 96, 99, 103, 101, 97, 100, 92, 105, 98

1. Identify the studied variable and specify its type (quantitative/qualitative).
2. Construct the statistical table.
3. Determine the characteristics of central tendency.
4. Calculate the characteristics of dispersion.
5. Represent the Data with the appropriate graphs.

Solution 07:

1. The studied variable is quantitative.

2. Statistical table:

To construct a frequency table, we follow these steps:
Step 1: Organize the data in ascending order: 73 to 115.
Step 2: Count the frequency of each value.
We will group the data into a table with the following columns:
Step 3: Count the occurrences: We will group the data into a table with the below columns.
73 → 1
77 → 1
78 → 1
79 → 1
80 → 1
81 → 1
82 → 1
83 → 1
84 → 2
85 → 3
86 → 2
87 → 3
88 → 3
89 → 2
90 → 2
91 → 2
92 → 4
93 → 2
94 → 3
95 → 3
96 → 3
97 → 4
98 → 3
99 → 4
100 → 4

101 → 3
102 → 3
103 → 3
104 → 1
105 → 4
106 → 1
107 → 2
108 → 2
109 → 1
110 → 1
111 → 1
115 → 1

Note: We have a large dataset (80 students) with many different values. Listing every single distinct value in a frequency table can become very long and hard to read. For example, in our raw frequency table, there were over 40 different study-hour values, which is a lot to process at a glance. By grouping the data into intervals (like 70–79, 80–89, etc.), the table becomes shorter and easier to interpret. This allows us to quickly see which ranges are most common without getting lost in small details.

How chosing the Number of classes?

When grouping data, we need to divide the nember of data into intervals (classes). If we choose too few classes, the table is too coarse and loses detail. If we choose too many classes, the table becomes long and hard to read.

So we aim for a reasonable number of classes, enough to show patterns, but not too many.

A method of grouping data is **Sturges' formula**:

$$k = 1 + log_2(N)$$

Determine class width: Once k is chosen, calculate class width**:**

$\text{Class width} = \frac{Range}{k}$,

with,

$$Range = \text{Max} - \text{Min}$$

In our exercise $N = 80$, so

$$k = 1 + \log_2(N) = 1 + \log_2(80) = 7.32 \rightarrow \text{round to } 7 \text{ or } 8 \text{ classes}$$

We choose $k = 7$.

In other hand

$$Range = \text{Max} - \text{Min} = 115 - 73 = 42$$

So, we calculate class width

$$\text{Class width} = \frac{Range}{k} = \frac{42}{7} = 6$$

Check the data carefully:

[73,79[: 73,77,78,79 → 4
[79,85[: 80,81,82,83,84,84,85,85,85 → 9
[85,91[: 86,86,87,87,87,88,88,88,89,89,90,90,91 → 13
[91,97[: 91,92,92,92,92,93,93,94,94,94,95,95,95,96,96,96,97,97,97,97 → 20
[97,103[: 98,98,98,99,99,99,100,100,100,100,101,101,101,102,102,102,103,103,103 → 19

[103,109[: 104,105,105,105,105,106,107,107,108,108,109 → 11
[109,116[: 110,111,115 → 4
Thus, the statistical table is given as

e_i	n_i	f_i	N_i	F_i	c_i	a_i	$\frac{f_i}{a_i}$	$n_i c_i$	$n_i c_i^2$	$\frac{n_i}{c_i}$	$n_i \ln(c_i)$
73			0	0							
79	4	0.05	4	0.05	76	6	0.0083	304	23104	0.0526	17.3228
85	9	0.1125	13	0.1625	82	6	0.0187	738	60516	0.1097	39.6603
91	13	0.1625	26	0.325	88	6	0.0270	1144	100672	0.1477	58.2049
97	20	0.25	46	0.575	94	6	0.0416	1880	176720	0.2127	90.866
103	19	0.2375	65	0.8125	100	6	0.0395	1900	190000	0.19	87.4988
109	11	0.1375	76	0.95	106	6	0.0229	1166	123596	0.1037	51.2974
116	4	0.05	80		112.5	7	0.0071	450	50625	0.0355	18.8964
Σ	80	1		1				7582	725233	0.8519	363.746
$\frac{\Sigma}{N}$	1							94.775	9065.4125		

3. Characteristics of central tendency:

Modal class:

Modal class: [91, 97[, because the highest $\frac{f_i}{a_i}$ is 0.0416.

Median (M): We have

$\Rightarrow$ 0.325< 0.5 <*0.575*

$$\Rightarrow M \in [91, 97[$$

So,

$$M = e_{i-1} + a_i \frac{0.50 - F(e_{i-1})}{f_i}$$

$$= 91 + 6\frac{0.5-0.325}{0.25} = 95.2.$$

Thus,

$$M = 95.2$$

Quartiles:

First quartile Q_1**:** We have

$$F(Q_1) = 0.25 \Rightarrow F(85) = 0.1625 < 0.25 < F(91) = 0.325$$
$$\Rightarrow 85 < Q_1 < 91$$

So,

$$Q_1 = e_{i-1} + a_i \frac{0.25 - F(e_{i-1})}{f_i}$$
$$= 85 + 6\frac{0.25 - 0.1625}{0.1625} = 88.231$$

Thus,

$$Q_1 = 88.231$$

Third quartile Q_3**:** We have

$$F(Q_3) = 0.75 \Rightarrow F(97) = 0.575 < 0.75 < F(103) = 0.8125$$
$$\Rightarrow 97 < Q_3 < 103$$

So,

$$Q_3 = e_{i-1} + a_i \frac{0.75 - F(e_{i-1})}{f_i}$$
$$= 97 + 6\frac{0.75 - 0.575}{0.2375} = 101.42$$

Thus,

$$Q_3 = 107.737$$

Arithmetic mean:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} n_i c_i = \sum_{i=1}^{n} f_i c_i$$

From the statistical table, we get

$$\bar{X} = 94.775$$

Harmonic mean:

From the statistical table, we get

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{c_i}} = \frac{80}{0.8519} = 93.91.$$

Geometric mean:

From the statistical table, we get

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} c_i^{n_i}}$$

Since the product can be huge, we calculate

$$\ln(\bar{X}_{GM}) = \frac{1}{N}\sum_{i=1}^{k} n_i \ln(c_i)$$

So,

$$= \frac{363.746}{80}$$

$$= 4.5468$$

Then,

$$\bar{X}_{GM} = e^{4.5468} = 94.33$$

4. Characteristics of dispersion:

Range:

$$R = e_k - e_0 = 116 - 73 = 43$$

Interquartile range:

$$I_Q = Q_3 - Q_1 = 101.42 - 88.231 = 13.189.$$

Variance:

Note that

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\, c_i^2 - \bar{X}^2$$

Thus,

$$Var(X) = 9065.4125 - (94.775)^2$$
$$= 9065.4125 - 8902.3006$$
$$\approx 83.1119.$$

Standard deviation:

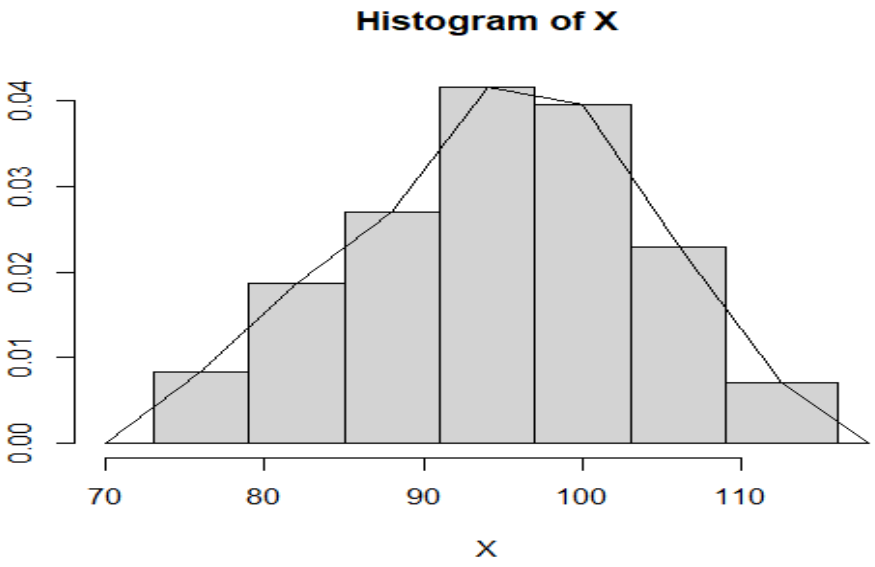
$$\sigma_X = \sqrt{Var(X)}$$
$$= \sqrt{83.1119}$$
$$= 9.1166.$$

Coefficient of variation:

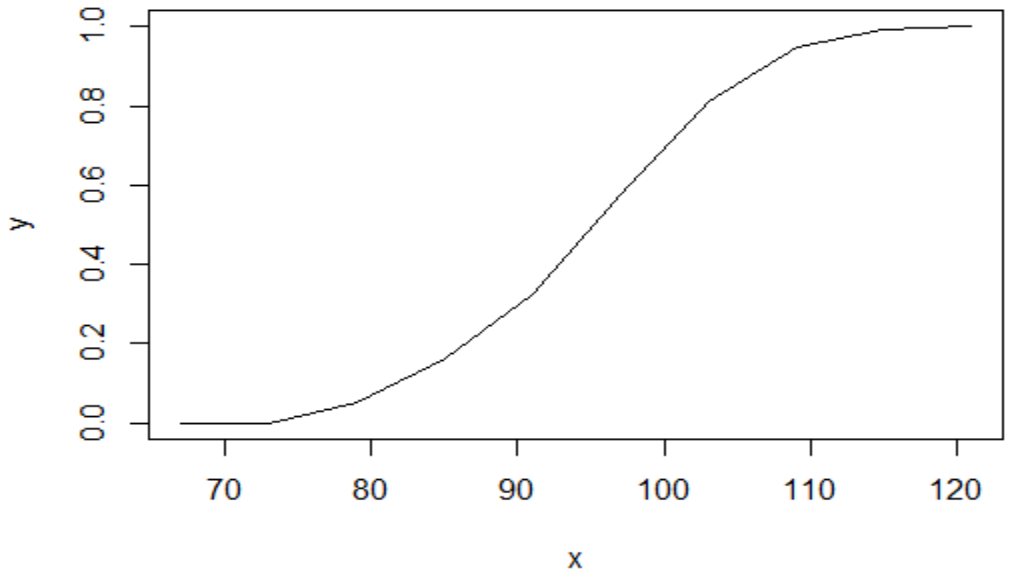
$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100 = \frac{9.102}{94.775} \times 100 = 9.6192\ \%$$

5. Represent the data with the appropriate graphs:

Bar chart:



Intergral diagram:



Exercise 08:
The following table presents the distribution of children's weights (in kilograms) across different intervals.

Class Interval (kg)	[0,2 [	[2,4[	[4,6[	[6,8[	[8,12[	[12,20[

Frequency	90	25	30	11	44	200

1. What is the type of the studied characteristic?
2. Create the statistical table.
3. Draw the bar and cumulative charts.
4. Calculate the quartiles and then determinite them graphically.
5. Determine R such that *10%* of the individuals are within the range $[R, Q_1[$.

Solution 08:

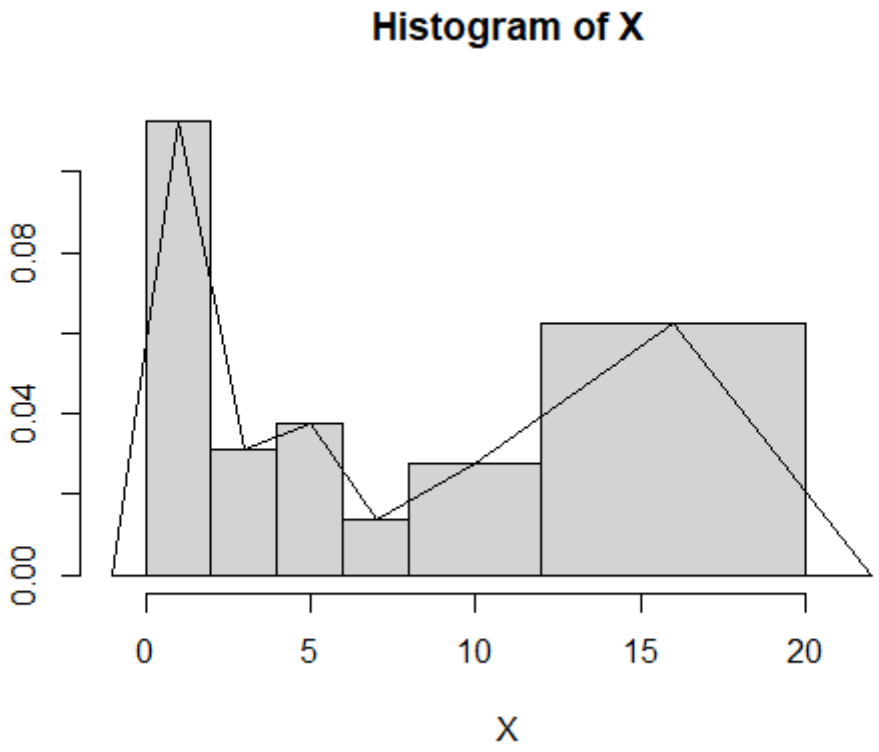
1. The type of the studied characteristic

The characteristic is weight of children, measured in kilograms, grouped in intervals.
This is a quantitative continuous variable because weight can take any value within an interval.

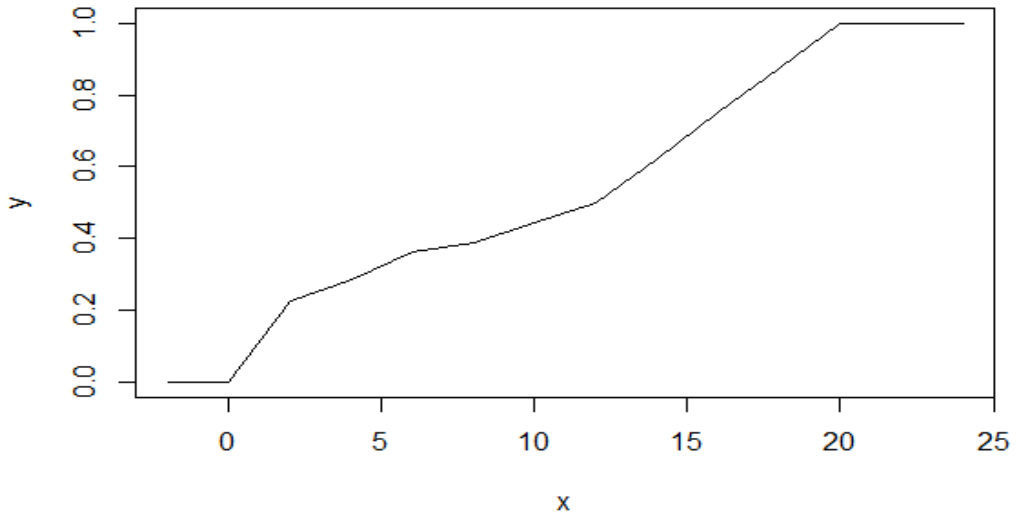
2. The statistical table:

e_i	n_i	f_i	N_i	F_i	a_i	$\frac{f_i}{a_i}$
~~0~~			~~0~~	~~0~~		
~~2~~	90	0.225	~~90~~	~~0.225~~	2	0.1125
~~4~~	25	0.0625	~~115~~	~~0.2875~~	2	0.03125
~~6~~	30	0.075	~~145~~	~~0.3625~~	2	0.0375
~~8~~	11	0.0275	~~156~~	~~0.39~~	2	0.01375
~~12~~	44	0.11	~~200~~	~~0.50~~	4	0.0275
~~20~~	200	0.50	~~400~~	~~1~~	8	0.0625
Σ	400	1				

3. Draw the bar chart:



Draw the cumulative chart:



4. The quartiles:

Median (M):

Since $F(M) = 0.5$ appears exactly in the column F_i of the statistical table, the value of the first quartile corresponds to the upper bound of the class where this cumulative frequency is reached. Therefore:

$$M = 12.$$

First qartile Q_1:

Since,

$$F(2) = 0.225 < F(Q_1) = 0.25 < F(4) = 0.2875$$

So,

$$2 < Q_1 < 4$$

Thus,

$$Q_1 = e_{i-1} + a_i \frac{0.25 - F(e_{i-1})}{f_i}$$
$$= 2 + 2\frac{0.25 - 0.225}{0.0625}$$
$$=2.8$$

Then,

$$Q_1 = 2.8$$

Third quartile Q_3:

Since,

$$F(12) = 0{,}50 < F(Q_3) = 0.75 < F(20) = 1$$

So,

$$12 < Q_3 < 20$$

Thus,

$$Q_3 = e_{i-1} + a_i \frac{0.75 - F(e_{i-1})}{f_i}$$
$$= 12 + 8\frac{0.75 - 0.50}{0.5}$$
$$=16$$

Then,

$$Q_3 = 16$$

Determine R such that *10%* of the individuals are within the range $[R, Q_1[$:

We have

$$\frac{R - e_{i-1}}{e_i - e_{i-1}} = \frac{F(R) - F(e_{i-1})}{F(e_i) - F(e_{i-1})}$$

Since

$F(e_i) - F(e_{i-1}) = f_i, \quad e_i - e_{i-1} = a_i$

So,

$$\frac{R - e_{i-1}}{a_i} = \frac{F(R) - F(e_{i-1})}{f_i}$$

Thus,

$$R = e_{i-1} + a_i \frac{F(R) - F(e_{i-1})}{f_i}$$

Firstly, we calculate $F(R) = ?$

Percentage of individuals within the range $[R, Q_1[$ is

$Percentage = \big(F(Q_1) - F(R)\big) \times 100 \Rightarrow 10 = \big(0.25 - F(R)\big) \times 100$

$$\Rightarrow F(R) = 0.25 - \frac{10}{100}$$

$$\Rightarrow F(R) = 0.25 - 0.10$$

$$\Rightarrow F(R) = 0.15$$

Note that,

$0 < F(R) = 0.15 < 0.225$

So,

$0 < R < 2$

Thus

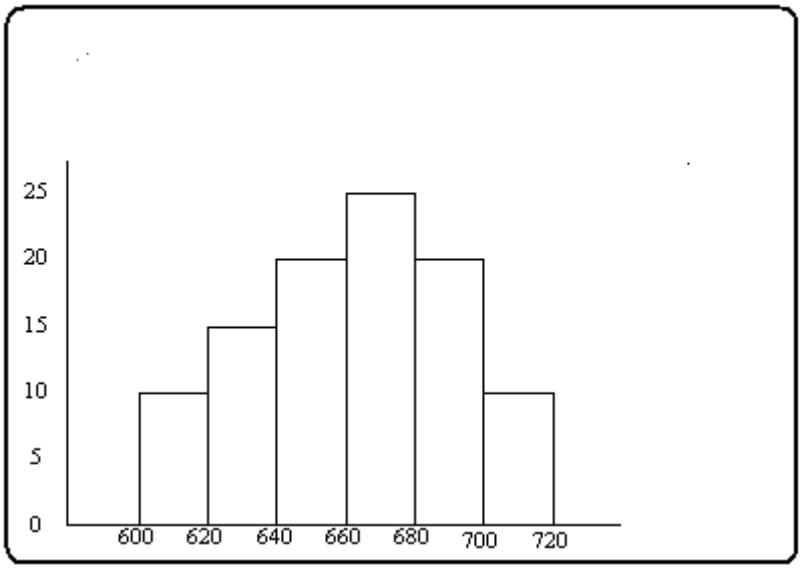
$$R = e_{i-1} + a_i \frac{F(R) - F(e_{i-1})}{f_i}$$

$$= 0 + 2\,\frac{0.15-0}{0.225}$$

$$= 1.333$$

Exercise 09:

The following chart representing the weights of a chicken.



1. What does this chart represent? What is the type of the studied characteristic?
2. Determine the modal class.
3. Calculate the characteristics of central tendency.
4. Calculate the characteristics of dispersion.
5. Draw the cumulative curve and determine the quartiles graphically.
6. Calculate the percentage of individuals within the range $[\bar{X}\text{-}\sigma_X\ , \bar{X} + \sigma_X[$.

7. Calculate the percentage of individuals within the range $[M\ , \bar{X} + \sigma_X[$.
8. Caculate $\bar{Y}$ and σ_X using the transformation $Y = -3X + 5$.

Solution 09:

1. The chart representing the bar chart weights of a chicken.

Type of the studied characteristic is continuous quantitative**.**

2. Modal class:

Modal class: [660, 680[, because the highest is $n_i = 25$ and all $a_i = 20$.

Statistical table:

e_i	n_i	f_i	N_i	F_i	c_i	a_i	$n_i c_i$	$n_i c_i^2$	$\frac{n_i}{c_i}$	$n_i\, ln(c_i)$
600			0	0						
	10	0.10			610	20	6100	3721000	0.0163	64.134
620			10	0.10						
	15	0.15			630	20	9450	5953500	0.0238	96.685
640			25	0.25						
	20	0.20			650	20	13000	8450000	0.0307	129.54
660			45	0.45						
	25	0.25			670	20	16750	11222500	0.0373	162.682
680			70	0.70						
	20	0.20			690	20	13800	9522000	0.0289	130.734
700			90	0.90						
	10	0.10			710	20	7100	5041000	0.0140	65.653
720			100	1						
Σ	100	1					66200	43910000	0.1510	649.428
$\frac{\Sigma}{N}$	1						662	439100		

3. Characteristics of central tendency:

Median (M):

We have

$F(M) = 0.5 \Rightarrow 0.45 < 0.5 < 0.70$

$$\Rightarrow M \in [660, 680[$$
$$\Rightarrow M = e_{i-1} + a_i \frac{0.50 - F(e_{i-1})}{f_i}$$
$$\Rightarrow M = 660 + 20\frac{0.5 - 0.45}{0.25}$$
$$\Rightarrow M = 664$$

First qartile Q_1**:**

Since $F(Q_1) = 0.25$ appears exactly in the column F_i of the statistical table, the value of the first quartile corresponds to the upper bound of the class where this cumulative frequency is reached. Therefore:

$$Q_1 = 640.$$

Third quartile Q_3**:**

We have

$$F(680) = 0.70 < F(Q_3) = 0.75 < F(700) = 0.90$$

So,

$$680 < Q_3 < 700$$

Thus,

$$Q_3 = e_{i-1} + a_i \frac{0.75 - F(e_{i-1})}{f_i}$$
$$= 680 + 20\frac{0.75 - 0.70}{0.20}$$
$$=685$$

Then,

$$Q_3 = 685\ g$$

Arithmetic mean:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} n_i c_i = \sum_{i=1}^{n} f_i c_i$$

From the statistical table, we get

$$\bar{X} = 662$$

Harmonic mean:

From the statistical table, we get

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{c_i}} = \frac{100}{0.1510} = 662.25.$$

Geometric mean:

From the statistical table, we get

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} c_i^{n_i}}$$

Since the product can be huge, we calculate

$$\ln(\bar{X}_{GM}) = \frac{1}{N}\sum_{i=1}^{k} n_i \ln(c_i)$$

So,

$$= \frac{649.428}{100}$$

$$= 6.49428$$

Then,

$$\bar{X}_{GM} = e^{6.49428} = 661.348$$

4. Characteristics of dispersion:

Range:

$$R = e_k - e_0 = 720 - 600 = 120$$

Interquartile range:

$$I_Q = Q_3 - Q_1 = 685 - 640 = 45.$$

Variance:

Note that

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\, c_i^2 - \bar{X}^2$$

Thus,

$$Var(X) = 439100.00 - (662)^2$$
$$= 439100 - 438244$$
$$= 856.$$

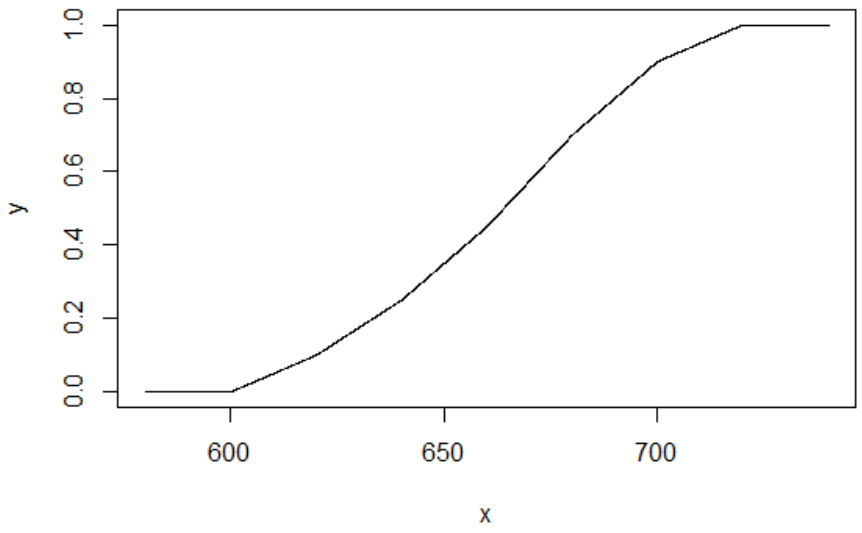
Standard deviation:

$$\sigma_X = \sqrt{Var(X)}$$
$$= \sqrt{856}$$
$$= 29.26.$$

Coefficient of variation:

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100 = \frac{29.26}{662} \times 100 = 4.42\%$$

5. Draw the cumulative curve and determine the quartiles graphically:



6. Percentage of individuals within the range $[\bar{X}\text{-}\sigma_X\ , \bar{X}+\sigma_X[$:

We have,

$$\bar{X}-\sigma_X = 662 - 29.26 = 632.74$$
$$\bar{X}+\sigma_X = 662 + 29.26 = 691.26$$

So,

$$[\bar{X}\text{-}\sigma_X\ , \bar{X}+\sigma_X[= [632.74\ ,\ 691.26[.$$

Calculate $F(\bar{X}\text{-}\sigma_X) = F(632.74\ )$

We have

$$\frac{Z-e_{i-1}}{e_i-e_{i-1}} = \frac{F(Z)-F(e_{i-1})}{F(e_i)-F(e_{i-1})}$$

Since

$F(e_i)-F(e_{i-1}) = f_i, \quad e_i-e_{i-1} = a_i$

So,

$$\frac{Z-e_{i-1}}{a_i} = \frac{F(Z)-F(e_{i-1})}{f_i}$$
$$F(Z) = F(e_{i-1}) + f_i\frac{Z-e_{i-1}}{a_i}$$

Note that

$$620 < \bar{X}\text{-}\sigma_X = 632.74 < 640$$

So,

$$F(620) = 0.10 < F(632.74\ ) < F(640) = 0.25$$

and

$$f_i = 0.15, \qquad a_i = 20$$

Thus,

$$F(632.74) = F(e_{i-1}) + f_i \frac{Z - e_{i-1}}{a_i}$$
$$= 0.10 + 0.15 \frac{632.74 - 620}{20}$$
$$= 0.1956$$

Calculate $F(\bar{X}+\sigma_X) = F(691.26)$
Note that

$$680 < \bar{X}+\sigma_X = 691.26 < 700$$

So,

$$F(680) = 0.70 < F(691.26) < F(700) = 0.90$$

and

$$f_i = 0.20, \qquad a_i = 20$$

Thus,

$$F(691.26) = F(e_{i-1}) + f_i \frac{Z - e_{i-1}}{a_i}$$
$$= 0.70 + 0.20 \frac{691.26 - 680}{20}$$
$$= 0.8126$$

Percentage of individuals within the range $[\bar{X}-\sigma_X , \bar{X} + \sigma_X[$ is

$$Percentage = \big(F(691.26) - F(632.74)\big) \times 100$$
$$= (0.813 - 0.196) \times 100$$
$$= 0.617 \times 100$$
$$= 61.7\%$$

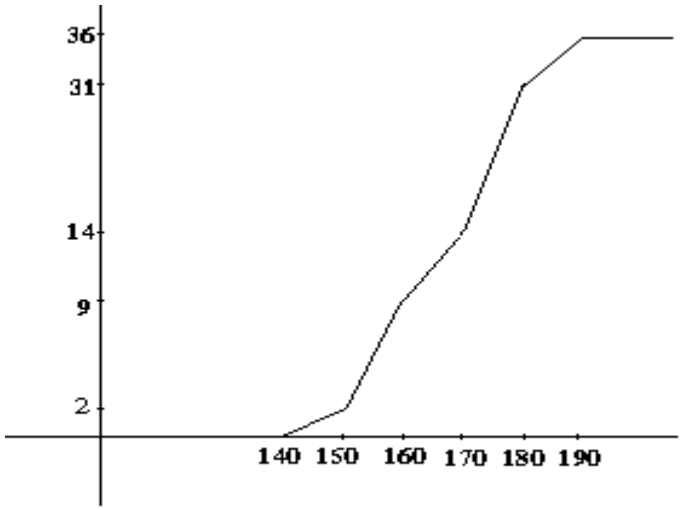
7. Percentage of individuals within the range $[M , \bar{X} + \sigma_X[$: is

$$Percentage = \big(F(691.26) - F(M)\big) \times 100$$
$$= (0.813 - 0.5) \times 100$$
$$= 0.313 \times 100$$
$$= 31.3\%$$

8. Caculate $\bar{Y}$ and σ_Y using the transformation $Y = -3X + 5$:
   Since, $Y = -3X + 5$. So,
   $\bar{Y} = -3\bar{X} + 5 == -3 \times 662 + 5 = -1981$
   and

$$\sigma_Y = 3\sigma_X = 3 \times 29.26 = 87.78$$

Exercise 10:
The following graph represents the height of university students.



1. What is the type of the studied variable?
2. Construct the statistical table.
3. Calculate the measures of central tendency.
4. Calculate the measures of dispersion.
5. Represent the data with the appropriate graph.
6. Calculate the percentage of individuals within the range $[M\ , \bar{X} + \sigma_X[$.

Solution 10:

1. Type of the studied characteristic is continuous quantitative**.**

2. Statistical table:

e_i	N_i	n_i	f_i	F_i	c_i	a_i	$n_i c_i$	$n_i c_i^2$	$\frac{n_i}{c_i}$	$n_i\, ln(c_i)$
140	0			0						
150	2	2	0.0556	0.0556	145	10	290	42050	0.01379	9.9534
160	9	7	0.1944	0.2500	155	10	1085	168175	0.04516	35.3039
170	14	5	0.1389	0.3889	165	10	825	136125	0.03030	25.5297
180	31	17	0.4722	0.8611	175	10	2975	520625	0.09714	87.8013
190	36	5	0.1389	1	185	10	925	171125	0.02702	26.1017
Σ		36	1				6100	1038100	0.21341	184.6900
$\frac{\Sigma}{N}$	1						169.44	28836.11		

3. Characteristics of central tendency:

Modal class:

Modal class: [170, 180[, because the highest is $n_i = 17$ and all $a_i = 10$.

Median (M):

We have

$F(M) = 0.5 \Rightarrow$*0.3889*$< 0.5 <$*0.8611*

$$\Rightarrow M \in [170, 180[$$

$$\Rightarrow M = e_{i-1} + a_i \frac{0.50 - F(e_{i-1})}{f_i}$$

$$\Rightarrow M = 170 + 10 \frac{0.5 - 0.3889}{0.4722}$$

$$\Rightarrow M = 172.35$$

First qartile Q_1**:**

Since $F(Q_1) = 0.25$ appears exactly in the column F_i of the statistical table, the value of the first quartile corresponds to the upper bound of the class where this cumulative frequency is reached. Therefore:

$$Q_1 = 160.$$

Third quartile Q_3**:** We have

$F(170) = 0.3889 < F(Q_3) = 0{,}75 < F(180) =$*0.8611*

So,

$$170 < Q_3 < 180$$

Thus,

$$Q_3 = e_{i-1} + a_i \frac{0.75 - F(e_{i-1})}{f_i}$$
$$= 170 + 10 \frac{0.75 - 0.3889}{0.4722}$$
$$= 177.65$$

Then,

$$Q_3 = 177.65$$

Arithmetic mean:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{n} n_i c_i = \sum_{i=1}^{n} f_i c_i$$

From the statistical table, we get

$$\bar{X} = 169.44$$

Harmonic mean:

From the statistical table, we get

$$\bar{X}_{HM} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{c_i}} = \frac{36}{0.21341} = 168.69.$$

Geometric mean:

From the statistical table, we get

$$\bar{X}_{GM} = \sqrt[N]{\prod_{i=1}^{k} c_i^{n_i}}$$

Since the product can be huge, we calculate

$$\ln(\bar{X}_{GM}) = \frac{1}{N} \sum_{i=1}^{k} n_i \ln(c_i)$$

So,

$$\ln(\bar{X}_{GM}) = \frac{184.69}{36}$$
$$= 5.1302$$

Then,

$$\bar{X}_{GM} = e^{5.1302} = 169.05$$

4. Characteristics of dispersion:

Range:

$$R = e_k - e_0 = 190 - 140 = 50$$

Interquartile range:

$$I_Q = Q_3 - Q_1 = 177.65 - 160 = 17.65.$$

Variance:
Note that

$$Var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i\, c_i^2 - \bar{X}^2$$

Thus,

$$\begin{aligned} Var(X) &= 28836.11 - (169.44)^2 \\ &= 28836.11 - 28709.9136 \\ &\approx 126.1964. \end{aligned}$$

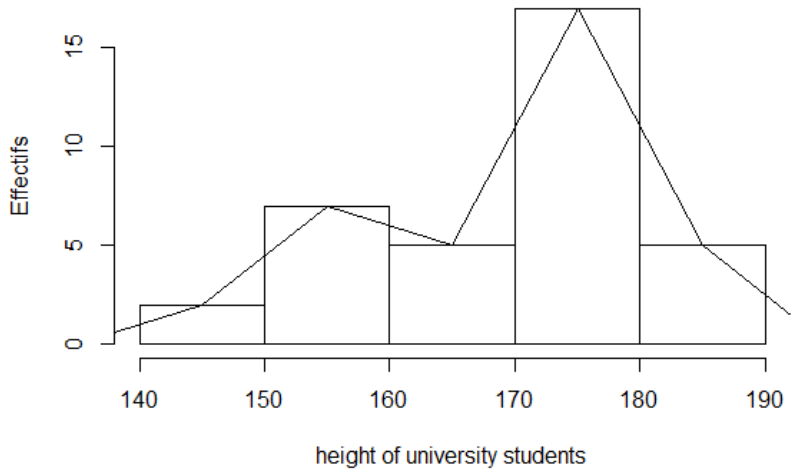
Standard deviation:

$$\begin{aligned} \sigma_X &= \sqrt{Var(X)} \\ &= \sqrt{126.1964} \\ &= 11.23. \end{aligned}$$

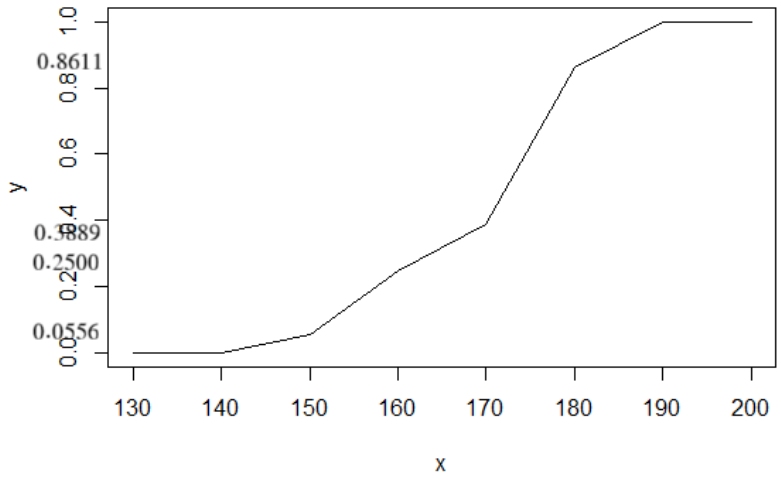
Coefficient of variation:

$$\text{CV} = \frac{\sigma_X}{\bar{X}} \times 100 = \frac{11.23}{169.44} \times 100 = 6.63\%$$

5. Draw the bar chart:



Integral diagram



6. Percentage of individuals within the range $[M\ , \bar{X} + \sigma_X[$:

Since,

$$\bar{X} - \sigma_X = 169.44 + 11.23 = 180.67$$

So,

$$[M,\ \bar{X} + \sigma_X[= [172.35\ ,\ 180.67[.$$

Calculate $F(\bar{X}+\sigma_X) = F(180.67)$

Note that

$$180 < \bar{X}+\sigma_X = 180.67 < 190$$

So,

$$F(180) = 0.8611 < F(180.67\ ) < F(190) = 1$$

and

$$f_i = 0.1389, \qquad a_i = 10$$

Thus,

$$F(180.67) = F(e_{i-1}) + f_i \frac{Z - e_{i-1}}{a_i}$$

$$= 0.8611 + 0.1389 \frac{180.67-180}{10}$$

$$= 0.8704$$

Therefore, the percentage of individuals within the range $[M\ , \bar{X} + \sigma_X[$ is

$$\text{Percentage} = \big(F(180.67\ ) - F(M)\big) \times 100$$

$$= (0.8704 - 0.5) \times 100$$

$$= 0.3704 \times 100$$

$$= 37.04\%$$

Bivariate Descriptive Statistics

III. Bivariate Descriptive Statistics

1. Introduction:

We say that a descriptive statistic is bivariate if the characteristic appears in the form of a pair $Z(X,Y)$

$$Z: \Omega \to \mathbb{R}^2$$
$$\omega \to Z(\omega) = \big(X(\omega), Y(\omega)\big)$$

Examples:

- See if families with more children tend to have more pets.
  - ✓ X Number of children in a family.
  - ✓ Y Number of pets in the family.

  Note that, both can take countable values (integers).

- Study how blood pressure varies with age.
  - ✓ X Age of a person (years).
  - ✓ Y Systolic blood pressure (mmHg).

  Note that, both variables can take any value in a range.

- Number of practice sessions in relation to sprint time.
  - ✓ X Number of practice sessions per week (discrete).
  - ✓ Y Sprint time in seconds (continuous).

  Note that, one variable is continuous (Y), the other is discrete (X).

Conclusion: From these examples, we can identify three cases:

- ✓ X and Y can be both continuous.
- ✓ X and Y can be both discrete.
- ✓ One variable can be continuous and the other discrete.

Objective: The goal of this chapter is to determine the relationship between X and Y, if one exists.

2. Basic statistical terms and concepts:

2.1. Statistical table:

Let the population be Ω with $card(\Omega) = N$, and let the statistical series be $\big(X_i, Y_j\big)_{\{i=1,\dots,k, j=1,\dots,p\}}$

- If X is discrete, then the values of the characteristic are $X(\Omega) = \{X_1, X_2, \dots, X_k\}$.
- If X is continuous, then $(X_i)_{1\le i\le k}$ represent the class centers.
- If Y is discrete, then the values of the characteristic are $Y(\Omega) = \{Y_1, Y_2, \dots, Y_p\}$.
- If Y is continuous, then $(Y_j)_{1\le j\le p}$ represent the class centers.

2.2. Absolute frequency:

Absolute frequency represents the number of the bivariate outcomes $\big(X_i, Y_j\big)_{\{i=1,\dots,k, j=1,\dots,p\}}$ occur, and it is denoted by n_{ij}. Note that,

$$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = N$$

2.3. **Relative Frequency:**

It is denoted by n_{ij}, where

$$f_{ij} = \frac{n_{ij}}{N}$$

and N is the total number of observations. Note that,

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

Example 1: In the study of whether families with more children tend to have more pets, we obtained the following table:

Family	1	2	3	4	5
Number of children (X)	2	0	3	1	2
Number of pets (Y)	3	2	1	1	2

Population: The all families in the study, with $N = 5$.

Characteristic (or variable):

- ✓ X Number of children in a family. Its type is quantitative (discrete), where

  $X(\Omega) = \{0,1,2,3\}$.

- ✓ Y Number of pets in a family Its type is quantitative (discrete), where,

  $Y(\Omega) = \{0,1,2,3\}$.

Through this example, we were able to calculate the absolute frequencies n_{ij}.

X \ Y	0	1	2	3
0	0	0	1	0

1	0	1	0	0
2	0	0	1	1
3	0	1	0	0

Through this example, we were able to calculate the relative frequencies $f_{ij} = \frac{n_{ij}}{N}$.

X \ Y	0	1	2	3
0	0	0	$\frac{1}{5} = 0.2$	0
1	0	$\frac{1}{5} = 0.2$	0	0
2	0	0	$\frac{1}{5} = 0.2$	$\frac{1}{5} = 0.2$
3	0	$\frac{1}{5} = 0.2$	0	0

Example 2: A researcher wants to study how systolic blood pressure (Y) varies with age (X). The variables are defined as

X \ Y	[80,100[	[100,120[	[120,140[	[140,1600[
[20,30[	3	8	5	4
[30,40[	1	7	12	10
[40,50[	2	4	5	4

Population: All the people whose age and systolic blood pressure were considered in the study, with $N = 65$.

Characteristic (or variable):

- ✓ X Number of children in a family. Its type is quantitative (*continuous*), where

  $X(\Omega) = \{[20,30[, [30,40[, [40,50[\}$.

- ✓ Y Number of pets in a family Its type is quantitative (*continuous*), where,

  $Y(\Omega) = \{[80,100[, [100,120[, [120,140[, [140,160[\}$.

Through this example, we were able to calculate the relative frequencies $f_{ij} = \frac{n_{ij}}{N}$, as well as the class midpoints (class centers).

$[20,30[\rightarrow X_1 = \frac{20+30}{2} = 25$ $[80,100[\rightarrow Y_1 = \frac{80+100}{2} = 90$ $[30,40[\rightarrow X_2 = \frac{30+40}{2} = 35$ $[100,120[\rightarrow Y_2 = \frac{100+120}{2} = 110$

$[40,50[\rightarrow X_3 = \frac{40+50}{2} = 45$ $[120,140[\rightarrow Y_3 = \frac{120+140}{2} = 130$

$[140,160[\rightarrow Y_4 = \frac{140+160}{2} = 150$

X \ Y	$Y_1 = 90$	$Y_2 = 110$	$Y_3 = 130$	$Y_4 = 150$
$X_1 = 25$	$\frac{3}{65} = 0.0462$	$\frac{8}{65} = 0.1231$	$\frac{5}{65} = 0.0769$	$\frac{4}{65} = 0.0615$
$X_2 = 35$	$\frac{1}{65} = 0.0154$	$\frac{7}{65} = 0.1077$	$\frac{12}{65} = 0.1846$	$\frac{10}{65} = 0.1538$
$X_3 = 45$	$\frac{2}{65} = 0.0308$	$\frac{4}{65} = 0.0615$	$\frac{5}{65} = 0.0769$	$\frac{4}{65} = 0.0615$

Example 3: The followig table shows the number of practice sessions per week (X), which is a discrete variable, in relation to sprint time in seconds (Y), which is a continuous variable.

X \ Y	[10,12[	[12,14[	[14,16[	[16,18[
1	6	7	5	4
2	8	9	4	4
3	9	10	5	3
4	7	9	6	4

Population: The population consists of individuals whose sprint time and number of practice sessions per week are being measured, with $N = 100$.

Characteristic (or variable):

- ✓ X Number of practice sessions per week. Its type is quantitative (*discrete*), where

  $X(\Omega) = \{1, 2, 3, 4\}$.

- ✓ Y Sprint time. Its type is quantitative (*continuous*), where,

  $Y(\Omega) = \{[10,12[, [12,14[, [14,16[, [16,18[\}$.

Through this example, we were able to calculate the relative frequencies $f_{ij} = \frac{n_{ij}}{N}$, as well as the class midpoints (class centers).

$[10, 12[\rightarrow Y_1 = \frac{10+12}{2} = 11$ $\quad$ $[14, 16[\rightarrow Y_3 = \frac{14+16}{2} = 15$

$[12, 14[\rightarrow Y_2 = \frac{12+14}{2} = 13$ $\quad$ $[16, 18[\rightarrow Y_4 = \frac{16+18}{2} = 17$

X \ Y	$Y_1 = 11$	$Y_2 = 13$	$Y_3 = 15$	$Y_4 = 17$
1	0.06	0.07	0.05	0.04
2	0.08	0.09	0.04	0.04
3	0.09	0.10	0.05	0.03
4	0.07	0.09	0.06	0.04

3. Marginal Distributions:

3.1. Marginal absolute frequency:

The marginal frequency of X is denoted by $n_{i.}$, where

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

The marginal frequency of Y is denoted by $n_{.j}$, where

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

3.2. Marginal relative frequency:

The marginal relative frequency of X is denoted by $n_{i.}$, where

$$f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{N}$$

The marginal relative frequency of Y is denoted by $n_{.j}$, where

$$f_{.j} = \sum_{j=1}^{p} f_{ij} = \frac{n_{.j}}{N}$$

Note that,

$$\sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j} = \sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = N$$

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

Reading the table:

Reading a statistical (contingency) table means understanding how to interpret the numbers inside the table as well as the marginal frequencies.

- Reading the frequencies inside the table:

  Each value n_{ij} represents the number of individuals (the frequency) that

  belong simultaneously to: category i of variable X, and category j of variable Y.

- Reading the marginal frequencies:

  The marginal frequency for X tells you how many individuals fall in category i of X, regardless of Y.

  The marginal frequency for Y tells you how many individuals fall in category i of Y, regardless of X.

Example: We study the example 1 mentioned above.

X \ Y	0	1	2	3	$n_{i.}$
0	0	0	1	0	1
1	0	1	0	0	1
2	0	0	1	1	2
3	0	1	0	0	1
$n_{.j}$	0	2	2	1	$N = 5$

n_{12}=0 There are 0 families with 0 children and 1 pet.
$n_{31} = 0$ There are 0 families with 2 children and 0 pets.
$n_{1.} = 1$ There is 1 family with 0 children, regardless of how many pets they have.
$n_{.4} = 1$ There is 1 family with 3 pets, regardless of how many children they have.

Example: We study the example 2 mentioned above.

Y / X	[80,100[$Y_1 = 90$	[100,120[$Y_2 = 110$	[120,140[$Y_3 = 130$	[140,160[$Y_4 = 150$	$n_{i.}$
[20,30[$X_1 = 25$	3	8	5	4	20
[30,40[$X_2 = 35$	1	7	12	10	30
[40,50[$X_3 = 45$	2	4	5	4	15

$n_{.j}$	6	19	22	18	$N = 65$

n_{12}=5 individuals aged [20,30[with systolic blood pressure between 120 and 140.

n_{31}=2 individuals aged [40,50[with systolic blood pressure between 80 and 100.

$n_{1.} = 20$ This represents the total number of individuals aged [20,30[, across all blood pressure categories

$n_{.4} = 18$ This means there are 18 individuals with blood pressure [140,160[mmHg across all age groups.

Example: We study the example 3 mentioned above.

X \ Y	[10,12[$Y_1 = 11$	[12,14[$Y_2 = 13$	[14,16[$Y_3 = 15$	[16,18[$Y_4 = 17$	$n_{i.}$
1	6	7	5	4	22
2	8	9	4	4	25
3	9	10	5	3	27
4	7	9	6	4	26
$n_{.j}$	30	35	20	15	$N = 100$

n_{12}=7 There are 7 individuals whose sprint time is between 12 and 14 seconds and who practice 1 session per week.

$n_{31} = 9$ There are 9 individuals whose sprint time is between 10 and 12 seconds and who practice 3 sessions per week.

$n_{1.} = 22$ There are 22 individuals who practice 4 sessions per week, regardless of their sprint time.

$n_{.4} = 15$ There are 15 individuals whose sprint time is between 16 and 18 seconds**,** regardless of the number of practice sessions per week.

4. Marginal characteristics:

4.1. Marginal mean:

The marginal mean of X is defined by

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i$$

The marginal mean of Y is defined by

$$\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{p} n_{.j}Y_j$$

4.2. Marginal variance:

The marginal variance of X is given by

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i^2 - \bar{X}_M{}^2$$

The marginal variance of Y is given by

$$Var_M(Y) = \frac{1}{N}\sum_{j=1}^{p} n_{.j}Y_j^2 - \bar{Y}_M{}^2$$

4.3. **Marginal standard deviation:**

The marginal standard deviation of X is defined by

$$\sigma_M(X) = \sqrt{Var_M(X)}$$

The marginal standard deviation of Y is given by

$$\sigma_M(Y) = \sqrt{Var_M(Y)}$$

Remark:

If one of the variables X and Y (or the two of them) is continuous, we replace the values X_i and/or Y_j, bv the centres of the class intervals.

Example: We study the example 1 mentioned above.

Y / X	0	1	2	3	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$
0	0	0	1	0	1	0	0
1	0	1	0	0	1	1	1
2	0	0	1	1	2	4	8
3	0	1	0	0	1	3	9
$n_{.j}$	0	2	2	1	$N = 5$	$\sum_{i=1}^{k} n_{i.}X_i = 8$	$\sum_{i=1}^{k} n_{i.}X_i^2 = 18$
$n_{.j}Y_j$	0	2	4	3	$\sum_{j=1}^{p} n_{.j}Y_j = 9$		
$n_{.j}Y_j^2$	0	2	8	9	$\sum_{j=1}^{p} n_{.j}Y_j^2 = 19$		

Marginal characteristics:

The marginal mean of X is

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{4} n_{i.}X_i = \frac{8}{5} = 1.6$$

The marginal mean of Y is

$$\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{4} n_{.j}Y_j = \frac{9}{5} = 1.8$$

The marginal variance of X is

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i^2 - \bar{X}_M{}^2 = \frac{18}{5} - (1.6)^2 = 1.04$$

The marginal variance of Y is given by

$$Var_M(Y) = \frac{1}{N}\sum_{j=1}^{p} n_{.j}Y_j^2 - \bar{Y}_M{}^2 = \frac{19}{5} - (1.8)^2 = 0.56$$

The marginal standard deviation of X is

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{1.04} = 1.02$$

The marginal standard deviation of X is

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{0.56} = 0.748$$

Example: We study the example 2 mentioned above.

Y / X	90	110	130	150	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$
25	3	8	5	4	20	500	12500
35	1	7	12	10	30	1050	36750
45	2	4	5	4	15	675	30375
$n_{.j}$	6	19	22	18	$N = 65$	$\sum_{i=1}^{k} n_{i.}X_i$ =2225	$\sum_{i=1}^{k} n_{i.}X_i^2$ =79625
$n_{.j}Y_j$	540	2090	2860	2700	$\sum_{j=1}^{p} n_{.j}Y_j = 8190$		
$n_{.j}Y_j^2$	48600	229900	371800	405000	$\sum_{j=1}^{p} n_{.j}Y_j^2$ =1055300		

Marginal characteristics:

The marginal mean of X is

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{4} n_{i.}X_i = \frac{2225}{65} = 34.23$$

The marginal mean of Y is

$$\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{4} n_{.j}Y_j = \frac{8190}{65} = 126$$

The marginal variance of X is

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i^2 - \bar{X}_M{}^2 = \frac{79625}{65} - (34.23)^2 = 53.31.$$

The marginal variance of Y is given by

$$Var_M(Y) = \frac{1}{N}\sum_{j=1}^{\boldsymbol{p}} n_{.j}Y_j^2 - \bar{Y}_M{}^2 = \frac{1055300}{65} - (126)^2 = 359.38$$

The marginal standard deviation of X is

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{53.31} = 7.30$$

The marginal standard deviation of X is

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{359.38} = 18.95$$

Example: We study the example 3 mentioned above.

Y / X	11	13	15	17	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$
1	6	7	5	4	22	22	22
2	8	9	4	4	25	50	100
3	9	10	5	3	27	81	243
4	7	9	6	4	26	104	416
$n_{.j}$	30	35	20	15	$N = 100$	$\sum_{i=1}^{\boldsymbol{k}} n_{i.}X_i = 257$	$\sum_{i=1}^{\boldsymbol{k}} n_{i.}X_i^2 = 781$
$n_{.j}Y_j$	330	455	300	255	$\sum_{j=1}^{\boldsymbol{p}} n_{.j}Y_j = 1340$		
$n_{.j}Y_j^2$	3630	5915	4500	4335	$\sum_{j=1}^{p} n_{.j}Y_j^2$ =18380		

Marginal characteristics:

The marginal mean of X is

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{\boldsymbol{4}} n_{i.}X_i = \frac{257}{100} = 2.57$$

The marginal mean of Y is

$$\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{4} n_{.j}Y_j = \frac{1340}{100} = 13.4$$

The marginal variance of X is

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i^2 - \bar{X}_M{}^2 = \frac{781}{100} - (2.57)^2 = 1.2051$$

The marginal variance of Y is given by

$$Var_M(Y) = \frac{1}{N}\sum_{j=1}^{p} n_{.j}Y_j^2 - \bar{Y}_M{}^2 = \frac{18380}{100} - (13.4)^2 = 4.24$$

The marginal standard deviation of X is

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{1.2051} = 1.098$$

The marginal standard deviation of X is

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{4.24} = 2.06$$

5. Conditional distribution:

The conditional distribution of X is the distribution of X that corresponds to a fixed value Y_j of Y.

5.1. Conditional Relative Frequency:

The i-th conditional relative frequency of X given $Y = Y_j$ represents the proportion of individuals whose X value is X_i among all individuals for which $Y = Y_j$. It can be computed as:

$$f_{i/Y=Y_j} = \frac{n_{ij}}{n_{.j}}$$

The j-th conditional relative frequency of Y given $X = X_i$ represents the proportion of individuals whose Y value is Y_j among all individuals for which $X = X_i$. It can be computed as:

$$f_{j/X=X_i} = \frac{n_{ij}}{n_{i.}}$$

Note that,

$\sum_{i=1}^{k} f_{i/Y=Y_j} = 1$ and $\sum_{j=1}^{p} f_{j/X=X_i} = 1$

Example: We study the example 3 mentioned above.

- Conditional distribution of X given $Y \in [12,14[$

We fix the sprint-time class $Y = [12,14[$
Column total:

$$n_{.2} = 35$$

Conditional relative frequencies

$$f_{i/Y=[12,14[} = \frac{n_{i2}}{n_{.2}}$$

X_i	n_{i2}	$f_{i/Y=[12,14[}$
1	7	$7/35=0.20$
2	9	$9/35\approx0.257$
3	10	$10/35\approx0.286$
4	9	$9/35\approx0.257$
$n_{.j}$	35	1

- Conditional distribution of Y given $X = 3$

From the table, when $X = 3$ (3 practice sessions per week), the observed frequencies are

Y_j	n_{3j}	$f_{j/X=3}$
$[10, 12[$	9	$9/27=0.333$
$[12, 14[$	10	$10/27\approx0.370$
$[14, 16[$	5	$5/27\approx0.185$
$[16, 18[$	3	$3/27\approx0.111$
$n_{.j}$	27	1

6. **Conditional characteristics:**

6.1. **Conditional mean:**

The conditional mean of X given $Y = Y_j$ is defined by

$$\bar{X}_{/Y=Y_j} = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} X_i$$

The conditional mean of Y given $X = X_i$ is defined by

$$\bar{Y}_{/X=X_i} = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} Y_j$$

6.2. **Conditional variance:**

The conditional variance of X given $Y = Y_j$ is defined by

$$Var(X/Y = Y_j) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} X_i^2 - \bar{X}_{Y=Y_j}{}^2$$

The conditional variance of Y given $X = X_i$ is defined by

$$Var(Y/X = X_i) = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} Y_j^2 - \bar{Y}_{/X=X_i}{}^2$$

6.3. **Conditional standard deviation:**

The conditional standard deviation of X given $Y = Y_j$ is defined by

$$\sigma(X/Y = Y_j) = \sqrt{Var(X/Y = Y_j)}$$

The conditional standard deviation of Y given $X = X_i$ is defined by

$$\sigma(Y/X = X_i) = \sqrt{Var(Y/X = X_i)}$$

Remark:

If one of the variables X and Y (or the two of them) is continuous, we replace the values X_i and/or Y_j, by the centers of the class intervals.

Example: We study the example 3 mentioned above.

- Conditional distribution of X given $Y \in [12,14[$

X_i	n_{i2}	$f_{i/Y=[12,14[}$	$n_{i2}X_i$	$n_{i2}X_i^2$
1	7	0.20	7	7
2	9	0.257	18	36
3	10	0.286	30	90
4	9	0.257	36	144
$n_{.j}$	35	1	91	277

The conditional mean of X given $Y = [12,14[$ is defined by

$$\bar{X}_{/Y=[12,14[} = \frac{1}{n_{.2}} \sum_{i=1}^{4} n_{i2}X_i = \frac{91}{35} = 2.6$$

The conditional variance of X given $Y = [12,14[$ is defined by

$$Var(X/Y = [12,14[) = \frac{1}{n_{.2}} \sum_{i=1}^{4} n_{i2}X_i^2 - \left(\bar{X}_{/Y=[12,14[}\right)^2 = \frac{277}{35} - (2.6)^2 = 1.154$$

The conditional standard deviation of X given $Y = [12,14[$ is defined by

$$\sigma(X/Y = [12,14[) = \sqrt{Var(X/Y = [12,14[)} = \sqrt{1.154} \approx 1.07$$

- Conditional distribution of Y given $X = 3$

From the table, when $X = 3$ (3 practice sessions per week), the observed frequencies are

Y_j	n_{3j}	$f_{j/X=3}$	$n_{3j}Y_j$	$n_{3j}Y_j^2$
$[10, 12[$	9	0.333	99	1089
$[12, 14[$	10	0.370	130	1690
$[14, 16[$	5	0.185	75	1125
$[16, 18[$	3	0.111	51	867
$n_{.j}$	27	1	355	4771

The conditional mean of Y given $X = 3$ is defined by

$$\bar{Y}_{/X=3} = \frac{1}{n_{3.}} \sum_{j=1}^{4} n_{3j} Y_j = \frac{355}{27} \approx 13.15$$

The conditional variance of Y given $X = 3$ is defined by

$$Var(Y/X = 3) = \frac{1}{n_{3.}} \sum_{j=1}^{4} n_{3j} Y_j^2 - \bar{Y}_{/X=3}{}^2 = \frac{4771}{27} - 13.15^2 \approx 3.78$$

The conditional standard deviation of Y given $X = 3$ is defined by

$$\sigma(Y/X = 3) = \sqrt{Var(Y/X = 3)} = \sqrt{3.78} \approx 1.94$$

Until now, we have focused on variables considered separately. We have examined their measures of central tendency and observed how they are dispersed. But what about their relationships with one another? How can we measure the strength of the connection between them?

In the following sections, we present some answers to these questions.

1. Covariance of two characters:

1.1. Definition:

The covariance of two characters X and Y is defined by $(X_i - \bar{X}_M)$

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} \, (X_i - \bar{X}_M)\big(Y_j - \bar{Y}_M\big)$$

Or,

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} X_i Y_j - \bar{X}_M \bar{Y}_M$$

Indeed,

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} (X_i - \bar{X}_M)\big(Y_j - \bar{Y}_M\big)$$

$$= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} \left(n_{ij} X_i Y_j - n_{ij} X_i \bar{Y}_M - n_{ij} \bar{X}_M \mathrm{Y} + n_{ij} \bar{X}_M \bar{Y}_M\right)$$

$$= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij} X_i Y_j - \bar{Y}_M \left(\frac{1}{N} \sum_{i=1}^{k} X_i \sum_{j=1}^{p} n_{ij} \right) - \bar{X}_M \left(\frac{1}{N} \sum_{j=1}^{p} Y_j \sum_{i=1}^{k} n_{ij} \right) + \bar{Y}_M \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}$$

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}X_iY_j - \bar{Y}_M\left(\frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i\right) - \bar{X}_M\left(\frac{1}{N}\sum_{j=1}^{p} n_{.j}Y_j\right) + \frac{\bar{X}_M\bar{Y}_M}{N} \times N$$

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}X_iY_j - 2\bar{X}_M\bar{Y}_M + \bar{X}_M\bar{Y}_M$$

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}X_iY_j - 2\bar{X}_M\bar{Y}_M + \bar{X}_M\bar{Y}_M$$

Example: We study the example 3 mentioned above.

X \ Y	11	13	15	17	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$	$n_{ij}X_iY_j$
1	66 6	91 7	75 5	68 4	22	22	22	300
2	176 8	234 9	120 4	136 4	25	50	100	666
3	297 9	390 10	225 5	153 3	27	81	243	1065
4	308 7	468 9	360 6	272 4	26	104	416	1408
$n_{.j}$	30	35	20	15	$N = 100$	$\sum_{i=1}^{k} n_{i.}X_i = 257$	$\sum_{i=1}^{k} n_{i.}X_i^2 = 781$	
$n_{.j}Y_j$	330	455	300	255	$\sum_{j=1}^{p} n_{.j}Y_j = 1340$			
$n_{.j}Y_j^2$	3630	5915	4500	4335	$\sum_{j=1}^{p} n_{.j}Y_j^2$=18380			

$n_{ij}X_iY_j$	847	1183	780	629				$\sum_{j=1}^{p} n_{ij}\,X_iY_j = 3439$

Calculate $Cov(X,Y)$:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\,X_iY_j\; - \bar{X}_M\bar{Y}_M = \frac{3439}{100} - (2.57 \times 13.4)$$

So,

$$Cov(X,Y) = \frac{3439}{100} - 34.438 = -0.048$$

1.2. **Properties of the covariance:**

Now, we give some properties of the covariance.

Let X and Y be two discrete statistical variables taking respectively the values $X_1 < X_2 < \cdots < X_k$ and $Y_1 < Y_2 < \cdots < Y_p$ and let $a, b, c, d \in \mathbb{R}$ be some constants. We have

I. $Cov(X,Y) = Cov(Y,X)$.
II. $Cov(X,X) = Var_M(X)$.
III. $Var(X+Y) = Var_M(X) + Var_M(Y) + 2Cov(X,Y)$.
IV. $Cov(aX+b, cY+d) = a\,Cov(X,Y)$.
V. $|Cov(X,Y)| \leq \sigma_M(X)\sigma_M(Y)$.

Proof:

I. We have

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{P} n_{ij}X_iY_j - \bar{X}_M\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{P}\sum_{i=1}^{k} n_{ij}Y_jX_i - \bar{Y}_M\bar{X}_M = Cov(Y,X).$$

II. We have

$$Cov(X,X) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}X_iX_j - \bar{X}_M\bar{X}_M$$

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{P} n_{ij}X_iX_j - (\bar{X}_M)^2$$

Since X can not take two different values at the same time, we have

$$n_{ij} = \begin{cases} n_i & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j. \end{cases}$$

Then,

$$Cov(X,X) = \frac{1}{N}\sum_{i=1}^{k} n_i X_i^2 - (\bar{X}_M)^2 = Var_M(X).$$

III. We have

$$\begin{aligned} Var(X+Y) &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\left(X_i + Y_j - \overline{X+Y}_M\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\left(X_i - \bar{X}_M + Y_j - \bar{Y}_M\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}(X_i - \bar{X}_M)^2 + \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\left(Y_j - \bar{Y}_M\right)^2 + \frac{2}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}(X - \bar{X}_M)\left(Y_j - \bar{Y}_M\right) \\ &= \frac{1}{N}\sum_{i=1}^{k}(X_i - \bar{X}_M)^2\sum_{j=1}^{p} n_{ij} + \frac{1}{N}\sum_{j=1}^{p}\left(Y_j - \bar{Y}_M\right)^2\sum_{i=1}^{k} n_{ij} + 2\,Cov(X,Y) \\ &= \frac{1}{N}\sum_{i=1}^{k} n_{i.}(X_i - \bar{X}_M)^2 + \frac{1}{N}\sum_{j=1}^{l} n_{.j}\left(Y_j - \bar{Y}_M\right)^2 + 2\,Cov(X,Y) \\ &= Var_M(X) + Var_M(Y) + 2\,Cov(X,Y) \end{aligned}$$

IV. We have

$$\begin{aligned} Cov(aX+b, cY+d) &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\left(aX_i + b - \overline{aX+b}\right)\left(cY_j + d - \overline{cY+d}\right) \\ &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}(aX_i + b - a\bar{X}_M - b)\left(cY_j + d - c\bar{Y}_M - d\right) \\ &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} acn_{ij}(X_i - \bar{X}_M)\left(Y_j - \bar{Y}_M\right) \\ &= \frac{ac}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}(X_i - \bar{X}_M)\left(Y_j - \bar{Y}_M\right) \\ &= a\,Cov(X,Y) \end{aligned}$$

V. To demonstrate this property, we will first rely on the Cauchy–Schwarz inequality, which we shall now establish.

For any real numbers $a_1, a_2, \dots, a_p$ and $b_1, b_2, \dots, b_p$, the following inequality holds

$$\left|\sum_{i=1}^{p} a_i b_i\right| \leq \sqrt{\sum_{i=1}^{p} a_i^2 \sum_{i=1}^{p} b_i^2}$$

Let for $t \in \mathbb{R}$

$$P(t) = \sum_{i=1}^{p} (a_i t + b_i)^2 = \sum_{i=1}^{p} (a_i^2 t^2 + 2a_i b_i t + b_i^2) = t^2 \sum_{i=1}^{p} a_i^2 + 2t \sum_{i=1}^{p} a_i b_i + \sum_{i=1}^{p} b_i^2.$$

Since, by definition, the polynomial $P(t)$ is positive for all $t \in \mathbb{R}$, its discriminant Δ must be negative. Therefore,

$$\Delta = \left(2\sum_{i=1}^{p} a_i b_i\right)^2 - 4\left(\sum_{i=1}^{p} a_i^2\right)\left(\sum_{i=1}^{p} b_i^2\right) = 4\left(\sum_{i=1}^{p} a_i b_i\right)^2 - 4\left(\sum_{i=1}^{p} a_i^2\right)\left(\sum_{i=1}^{p} b_i^2\right) \leq 0$$

$$\Longrightarrow 4\left(\sum_{i=1}^{p} a_i b_i\right)^2 \leq 4\left(\sum_{i=1}^{p} a_i^2\right)\left(\sum_{i=1}^{p} b_i^2\right)$$

$$\Longrightarrow \left(\sum_{i=1}^{p} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{p} a_i^2\right)\left(\sum_{i=1}^{p} b_i^2\right)$$

$$\Longrightarrow \sqrt{\left(\sum_{i=1}^{p} a_i b_i\right)^2} \leq \sqrt{\left(\sum_{i=1}^{p} a_i^2\right)\left(\sum_{i=1}^{p} b_i^2\right)}$$

$$\Longrightarrow \left|\sum_{i=1}^{p} a_i b_i\right| \leq \sqrt{\left(\sum_{i=1}^{p} a_i^2\right)\left(\sum_{i=1}^{p} b_i^2\right)}$$

Now, we can prove property V. We have

$$|Cov(X,Y)| = \left|\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}(X_i - \bar{X}_M)(Y_j - \bar{Y}_M)\right|$$

$$= \left|\sum_{i=1}^{k} \sum_{j=1}^{p} \sqrt{\frac{n_{ij}}{N}}(X_i - \bar{X}_M)\sqrt{\frac{n_{ij}}{N}}(Y_j - \bar{Y}_M)\right|$$

Now, the *Cauchy-Schwarz inequality* gives

$$|Cov(X,Y)| \leq \sqrt{\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}}{N}(X_i-\bar{X}_M)^2\right)\left(\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}}{N}\left(Y_j-\bar{Y}_M\right)^2\right)}$$

$$= \sqrt{\left(\frac{1}{N}\sum_{i=1}^{k}(X_i-\bar{X}_M)^2\sum_{j=1}^{p}n_{ij}\right)\left(\frac{1}{N}\sum_{j=1}^{p}\left(Y_j-\bar{Y}_M\right)^2\sum_{i=1}^{k}n_{ij}\right)}$$

$$= \sqrt{\left(\frac{1}{N}\sum_{i=1}^{k}n_{i.}(X_i-\bar{X}_M)^2\right)\left(\frac{1}{N}\sum_{j=1}^{p}n_{.j}\left(Y_j-\bar{Y}_M\right)^2\right)}$$

$$= \sqrt{Var_M(X)Var_M(Y)}$$

$$= \sigma_M(X)\sigma_M(Y).$$

2. **Correlation coefficient:**

The correlation coefficient covariance of the two statistical variables X and Y is given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)}, \quad \sigma_M(X) \neq 0, \sigma_M(Y) \neq 0$$

This is the property that expresses the strength of the linear relationship between the two variables X and Y.

Remark:

- The covariance $Cov(X,Y)$ may be negative, positive, or zero.
- The sign of $\rho(X,Y)$ is the same as the sign of $Cov(X,Y)$.
- The correlation satisfies: $-1 \leq \rho(X,Y) \leq 1$.
- If $|Cov(X,Y)| = \sigma_M(Y)\sigma_M(X)$, then, $\rho(X,Y) = 1$. In this case, the points lie exactly on a straight line.
- If $\rho \cong \pm 1$, then there is a strong linear relationship between X and Y.
- If $|\rho| \cong 0$, this means that there is no linear relationship between X and Y, but it does not mean that there is no relationship of another kind.

Example: We study the example 3 mentioned above.

We have,

$$Cov(X,Y) = -0.048$$

Thus,

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)} = \frac{-0.048}{1.098 \times 2.06}$$

So,

$$\rho(X,Y) = \frac{-0.048}{2.26188} = -0.0212$$

Important note:

It is possible to construct statistical tables in the following form:

X	X_1	X_2	...	X_n
Y	Y_1	Y_2	...	Y_n

Note that

$$\begin{cases} n_{ij} = 1, i = j \\ n_{ij} = 0, i \neq j \end{cases}$$

The marginal mean of X is

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

The marginal mean of Y is

$$\bar{Y}_M = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

The marginal variance of X is

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}_M{}^2$$

The marginal variance of Y is given by

$$Var_M(Y) = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}_M{}^2$$

The marginal standard deviation of X is

$$\sigma_M(X) = \sqrt{Var_M(X)}$$

The marginal standard deviation of X is

$$\sigma_M(Y) = \sqrt{Var_M(Y)}$$

The covariance is

$$\begin{cases} n_{ij} = 1, i = j \\ n_{ij} = 0, i \neq j \end{cases} \Leftrightarrow\ Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k} X_i\,Y_i - \bar{X}_M\bar{Y}_M$$

The correlation coefficient is

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)}$$

Example 4:
Consider the following table.

X	0.95	1.2	1.3	1.4	1.5	1.6	1.7	1.8
Y	0.2	0.4	0.5	0.7	0.6	0.9	0.7	1

From the above, the statistical table is as follows:

Y / X	0.2	0.4	0.5	0.7	0.6	0.9	0.7	1	$n_{i.}$
0.95	1	0	0	0	0	0	0	0	1
1.2	0	1	0	0	0	0	0	0	1
1.3	0	0	1	0	0	0	0	0	1
1.4	0	0	0	1	0	0	0	0	1
1.5	0	0	0	0	1	0	0	0	1
1.6	0	0	0	0	0	1	0	0	1
1.7	0	0	0	0	0	0	1	0	1
1.8	0	0	0	0	0	0		1	1
$n_{.j}$	1	1	1	1	1	1	1	1	N=8

Note that

$$\begin{cases} n_{ij} = 1, i = j \\ n_{ij} = 0, i \neq j \end{cases}$$

So,

$$n_{i.} = n_{.j} = 1, \quad \forall\, i = \overline{1,5}$$

In this case, it is preferable that the previous table be in the following form:

	X_i	Y_i	X_i^2	Y_i^2	X_iY_i
	0.95	*0.2*	0.902	*0.04*	*0.19*
	1.2	*0.4*	1.44	*0.16*	0.48
	1.3	*0.5*	*1.69*	0.25	0.65
	1.4	*0.7*	1.96	0.49	0.98
	1.5	*0.6*	2.25	0.36	0.9
	1.6	*0.9*	2.56	0.81	1.44
	1.7	*0.7*	*2.89*	0.49	1.19
	1.8	1	3.24	1	1.8
Σ	11.45	*5*	16.932	3.6	7.63
$\Sigma/_N$	1.431	0.625	2.116	0.45	0.954

Calculate $\bar{X}_M$:

From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{N} X_i = 1.431$$

Calculate $\bar{Y}_M$:

From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{i=1}^{N} Y_i = 0.625$$

Calculate $Var_M(X)$:

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = 2.116 - (1.431)^2 = 2.116 - 2.0477 = 0.0683$$

Calculate $Var_M(Y)$:

$$Var_M(Y) = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}_M{}^2$$

From the table, we have

$$Var_M(Y) = 0.45 - (0.625)^2 = 0.45 - 0.3906 = 0.0594$$

Calculate $\sigma_M(X)$:

$$\sigma_M(X) = \sqrt{Var_M(X)} = 0.2613$$

Calculate $\sigma_M(Y)$:

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = 0.2437$$

Calculate $Cov(X,Y)$:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}_M\bar{Y}_M = 0.954 - (1.431 \times 0.625)$$

So,

$$Cov(X,Y) = 0.954 - 0.8943 = 0.0597$$

Calculate $\rho(X,Y)$:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)} = \frac{0.0597}{0.2613 \times 0.2437}$$

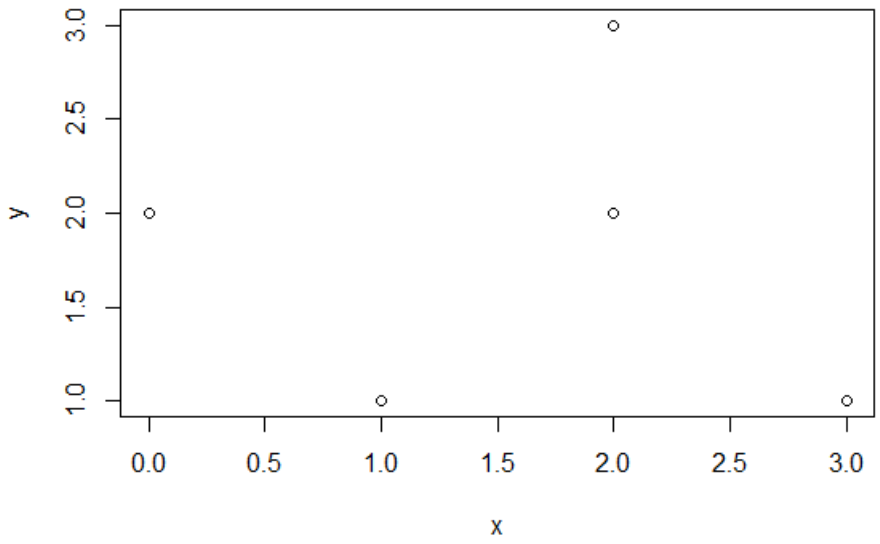
So,

$$\rho(X,Y) = \frac{0.0597}{0.06366} = 0.938$$

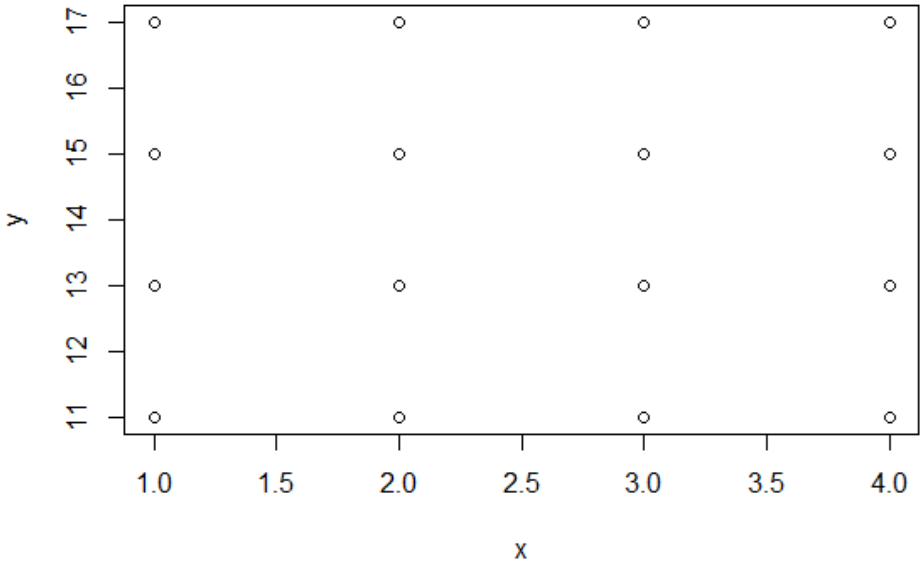
3. **Scatter plot:**

A scatter plot is a graph that shows the relationship between two variables, X and Y, by plotting points where the X value is the horizontal position and the Y value is the vertical position, where $n_{ij} \neq 0$.

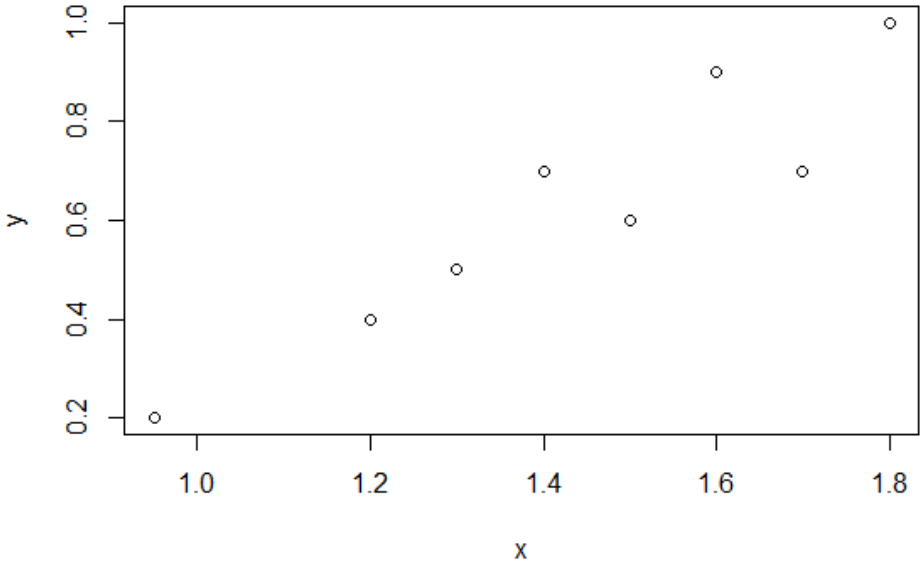
Example: We study the example 1 mentioned above.



Example: We study the example 3 mentioned above.



Example: We study the example 4 mentioned above.



4. Regression Line et Mayer Line:

4.1. Regression Line:

A regression line is a straight line that best fits a set of data points in a scatter plot. It shows the relationship between a dependent variable Y and an independent variable X. The most common is the **least squares regression line**:
$Y = aX + b$,

where

b = slope (rate of change of Y with respect to X).

a = intercept (value of Y when $X = 0$).

4.2. Mayer Line:

The Mayer line is a simplified method of drawing a regression line by splitting the data into two equal halves, finding the mean of X and Y in each half, and connecting the two points with a straight line. It's less precise than least squares but easier for manual calculations.

Example : We study the example 4 mentioned above.

Step 1: Split data into two halves

- First half: (0.95, 0.2), (1.2, 0.4), (1.3, 0.5), (1.4, 0.7) → mean:

$$x_1 = \frac{0.95 + 1.2 + 1.3 + 1.4}{4} = 1.2125$$

and

$$y_1 = \frac{0.2 + 0.4 + 0.5 + 0.7}{4} = 0.45$$

- Second half: (1.5, 0.6), (1.6, 0.9), (1.7, 0.7), (1.8, 1) → mean:

$$x_2 = \frac{1.5 + 1.6 + 1.7 + 1.8}{4} = 1.65$$

and

$$y_2 = \frac{0.6 + 0.9 + 0.7 + 1.0}{4} = 0.8$$

Step 2: Equation of the line through these points

- Slope

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{0.8 - 0.45}{1.65 - 1.2125} = \frac{0.35}{0.4375} = 0.8$$

- Intercept

$$a = y_1 - bx_1 = 0.45 - 0.8 \times 1.2125 = -0.52$$

Or,

$$a = y_2 - bx_2 = 0.8 - 0.8 \times 1.65 = -0.52$$

Step 3: Mayer Line Equation

$$Y = a + bX = 0.8X - 0.52.$$

Example 5:

Consider the following table:

X	1	2	3	4	5
Y	2	3	4	5	6

Step 1: Split data into two halves

- First half: (1,2), (2,3) → mean:

$$x_1 = \frac{1 + 2}{2} = \frac{3}{2} = 1.5$$

and

$$y_1 = \frac{2 + 3}{2} = \frac{5}{2} = 2.5$$

- Second half: (3,5), (4,4), (5,6) → mean:

$$x_2 = \frac{3 + 4 + 5}{3} = \frac{12}{3} = 4$$

and

$$y_2 = \frac{5+4+6}{3} = \frac{15}{3} = 5$$

Step 2: Equation of the line through these points

- Slope

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{5 - 2.5}{4 - 1.5} = \frac{2.5}{2.5} = 1$$

- Intercept

$$a = y_1 - bx_1 = 2.5 - 1 \times 1.5 = 1$$

Or,

$$a = y_2 - bx_2 = 4 - 1 \times 5 = 1$$

Step 3: Mayer Line Equation

$$Y = a + bX = 1 + X.$$

5. Linear fitting:

- If $\rho \cong \pm 1$, the points of the cloud are almost on a straight line.
- Our goal is to find the closest straight line that passes through the largest number of points of the cloud. This line is called the adjustment line (or line of best fit).
- This line is determined by specifying its equation.

Adjustment Line:

An adjustment line of Y in terms of X has the equation

$$Y = aX + b$$

where,

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var_M(X)} \\ b = \bar{Y}_M - a\bar{X}_M \end{cases}$$

An adjustment line of X in terms of Y has the equation

$$X = cY + d$$

where,

$$\begin{cases} c = \dfrac{Cov(X,Y)}{Var_M(Y)} \\ d = \bar{X}_M - c\bar{Y}_M \end{cases}$$

Note:

- The point $M(\bar{X}_M, \bar{Y}_M)$ belongs to the two adjustment lines.
- Position of the Adjustment Line: They intersect at the point $M(\bar{X}_M, \bar{Y}_M)$.
- They are perpendicular if $\rho = 0$, i.e., there is no linear correlation between Y and X.
- They are identical if $\rho \cong \pm 1$, i.e., there is a strong linear correlation between Y and X.
- So, if $\rho = 1$ or $\rho = -$, the two regression lines coincide and only in this case, we can solve one equation to determine the other one.
- We have,

$$a.c = \rho^2$$

Example: We study the example 4 mentioned above.

Regression lines of Y on X is given by $Y = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var_M(X)} \\ b = \bar{Y}_M - a\bar{X}_M \end{cases}$$

Or,

$$a = \frac{Cov(X,Y)}{Var_M(X)} = \frac{0.0597}{0.0683} \approx 0.874$$

and

$$b = \bar{Y}_M - a\bar{X}_M = 0.625 - 0.874 \times 1.431 = 0.625 - 1.250 \approx -0.625$$

Thus, the regression lines of Y on X is given by

$$Y = aX + b = 0.874\,\mathrm{X} - 0.625$$

Regression line of X on Y is given by $X = cY + d$, where

$$\begin{cases} c = \dfrac{Cov(X,Y)}{Var_M(Y)} \\ d = \bar{X}_M - c\bar{Y}_M \end{cases}$$

Or,

$$c = \frac{Cov(X,Y)}{Var_M(Y)} = \frac{0.0597}{0.0594} \approx 1.005$$

and

$$d = \bar{X}_M - c\bar{Y}_M = 0.625 - 1.005 \times 1.431 = 0.625 - 1.4381 = -0.8131$$

Thus, the regression lines of Y on X is given by

$$X = cY + d = 1.005\text{Y} - 0.8131.$$

6. Regression Curve:

A regression curve models the relationship between variables when it is not linear. In some situations, the scatter plot may suggest other forms of functions; for example, exponential, logarithmic, or polynomial regression can be used.

7. Regression Corridor:

Definition:

A Regression Corridor (or confidence band around the regression curve) is the band around the regression curve that shows the uncertainty of the predictions. It is usually drawn as a shaded band around the line.

Interpretation:

- If the corridor is narrow
  - The data points are close to the regression curve.
  - There is a strong relationship between the variables.
  - Predictions are more accurate.

- If the corridor is wide
  - The data points are scattered far from the curve.
  - The relationship between the variables is weak.
  - Predictions are less accurate.

Use of the regression corridor: The regression corridor helps create:

- Confidence intervals (uncertainty about the estimated line).
- Prediction intervals (uncertainty about new future observations).
- It shows where new data points are likely to fall.

8. Functional adjustment:

Functional adjustment is the process of finding a mathematical function that best represents the relationship between variables in a given dataset. The goal is to minimize the difference between the function's predicted values and the observed data, often using methods like least squares**,** regression analysis, or other optimization techniques.

8.1. Fitting the type $Y = c \times d^X$:

We will fit Y on X using the equation $Y = c \times d^X$, where $c,\ d > 0$are constants. In order to determine the constants c and d, we linearise the equation by the natural logarithm

$$Y = c \times d^X \Rightarrow \ln(Y) = \ln(c \times d^X)$$
$$\Rightarrow \ln(Y) = \ln(c) + \ln(d)X$$

If we let, $Z = \ln(Y),\ b = \ln(c)$ and $a = \ln(d)$, it follows that there is a regression line of Z on X is given by $Z = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Z)}{Var_M(X)} \\ b = \bar{Z}_M - a\bar{X}_M \end{cases}$$

We calculate a and b, and then construct c and d as shown below

$$\begin{cases} a = \ln(d) \Rightarrow d = e^a \\ b = \ln(c) \Rightarrow c = e^b \end{cases}$$

Example 6: Consider the following table:

X	-2	-1	0.01	1	2
Y	12	2.75	0.05	3.5	12.25

Statistical table: $N = 5$

	X_i	Y_i	X_i^2	Y_i^2	X_iY_i
	-2	12	4	144	-24
	-1	2.75	1	7.5625	-2.75
	0.01	0.05	0.0001	0.0025	0.0005
	1	3.5	1	12.25	3.5
	2	12.25	4	150.0625	24.5
Σ	0.01	30.55	10	313.878	1.2505
$\Sigma/_N$	0.002	6.11	2.0001	62.7755	0.25

Calculate $\bar{X}_M$:

From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{N} X_i = 0.002$$

Calculate $\bar{Y}_M$:

From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{i=1}^{N} Y_i = 6.11$$

Calculate $Var_M(X)$:

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = 2.00002 - (0.002)^2 = 2.00002 - 0.000004 \approx 2$$

Calculate $Var_M(Y)$:

$$Var_M(Y) = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}_M{}^2$$

From the table, we have

$$Var_M(Y) = 62.7755 - (6.11)^2 = 62.7755 - 37.3321 = 25.4434$$

Calculate $\sigma_M(X)$:

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{2} = 1.414$$

Calculat $\sigma_M(Y)$:

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{25.4434} = 5.044$$

Calculate $Cov(X,Y)$:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}_M\bar{Y}_M = 0.2501 - 0.002 \times 6.11$$

So,

$$Cov(X,Y) = 0.2501 - 0.0122 \approx 0.2379$$

Calculate $\rho(X,Y)$:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)} = \frac{0.2379}{1.414 \times 5.044}$$

So,

$$\rho(X,Y) = \frac{0.2379}{7.132} = 0.0334$$

Note that, the correlation coefficient is $\rho(X,Y) = 0.0334$, which is very close to 0. This means there is almost no linear relationship between X and Y in this data set.

We will fit Y on X using the equation $Y = c \times d^X$, where $c,\ d > 0$are constants. In order to determine the constants c and d, we linearise the equation by the natural logarithm

$$Y = c \times d^X \Rightarrow \ln(Y) = \ln(c \times d^X)$$
$$\Rightarrow \ln(Y) = \ln(c) + \ln(d)X$$

If we let, $Z = \ln(Y),\ b = \ln(c)$ and $a = \ln(d)$, it follows that there is a regression line of Z on X is given by $Z = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Z)}{Var_M(X)} \\ b = \bar{Z}_M - a\bar{X}_M \end{cases}$$

We calculate a and b, and then construct c and d as shown below

$$\begin{cases} a = \ln(d) \Rightarrow d = e^a \\ b = \ln(c) \Rightarrow c = e^b \end{cases}$$

We use the following statistical table

	X_i	Y_i	$Z_i = \ln(Y_i)$	Z_i^2	X_iZ_i
	-2	12	2.4849	6.175	−4.970
	-1	2.75	1.0116	1.023	1.012
	0.01	0.05	−2.9957	8.974	−0.029957
	1	3.5	1.2528	1.569	1.253
	2	12.5	2.5066	6.283	5.013
Σ	0	30.75	4.2602	24.024	0.2546
$\Sigma/_N$	0	6.15	0.852	4.8052	0.0509

Calculate $\bar{Z}_M$:

From the table, we have

$$\bar{Z}_M = \frac{1}{N}\sum_{i=1}^{N} Z_i = 0.852$$

Calculate $Var_M(Z)$:

$$Var_M(Z) = \frac{1}{N}\sum_{i=1}^{N} Z_i^2 - \bar{Z}_M{}^2$$

From the table, we have

$$Var_M(Z) = 4.805 - (0.852)^2 = 4.805 - 0.726 = 4.079$$

Calculate $\sigma_M(Z)$:

$$\sigma_M(Z) = \sqrt{Var_M(Z)} = \sqrt{4.079} = 2.02$$

Calculate $Cov(X,Z)$:

$$Cov(X,Z) = \frac{1}{N}\sum_{i=1}^{N} X_i Z_i - \bar{X}_M \bar{Z}_M = 0.0509 - (0.002 \times 0.852)$$

So,

$$Cov(X,Z) = 0.0509 - 0.001704 \approx 0.0492$$

Determine a and b

$$a = \frac{Cov(X,Z)}{Var_M(X)} = \frac{0.0492}{2} \approx 0.0246$$

and

$$b = \bar{Z}_M - a\bar{X}_M = 0.852 - 0.0246 \times 0.002 = 0.852 - 0.0000492 = 0.8519508$$

Thus, the regression lines of Y on X is given by

$$Z = aX + b = 0.0246\,\mathrm{X} + 0.852$$

Determine c and d

$$\begin{cases} a = \ln(d) \Rightarrow d = e^a \\ b = \ln(c) \Rightarrow c = e^b \end{cases}$$

So,

$$d = e^a = e^{0.0246} \approx 1.025$$

and

$$c = e^b = e^{0.852} \approx 2.344$$

So, the equation of Y on X is given by

$$Y = c \times d^X = 2.344 \times 1.025^X$$

8.2. **Fitting the type $X = a \times b^Y$**:

We will fit Y on X using the equation $Y = a \times b^X$, where $a, b > 0$are constants. In order to determine the constants a and b, we linearise the equation by the natural logarithm

$$X = a \times b^Y \Rightarrow \ln(X) = \ln(a \times b^Y)$$
$$\Rightarrow \ln(X) = \ln(a) + \ln(b)Y$$

If we let, $T = \ln(X)$, $d = \ln(a)$ and $c = \ln(b)$, it follows that there is a regression line of Z on X is given by $Z = cY + d$, where

$$\begin{cases} c = \dfrac{Cov(Y,T)}{Var_M(Y)} \\ d = \bar{T}_M - c\bar{Y}_M \end{cases}$$

We calculate a and b, and then construct c and d as shown below

$$\begin{cases} c = \ln(b) \Rightarrow b = e^c \\ d = \ln(a) \Rightarrow a = e^d \end{cases}$$

Example**:** We study the example 6 mentioned above.

We use the following statistical table:

	X_i	Y_i	$T_i = \ln(\lvert X_i \rvert)$	T_i^2	$Y_i T_i$
	-2	12	0.6931	0.4804	8.317
	-1	2.75	0	0	0
	0.01	0.05	−4.6052	21.2079	−0.2303
	1	3.5	0	0	0
	2	12.5	0.6931	0.4804	8.487
Σ	0	30.75	−3.219	22.1687	16.5737
$\Sigma/_N$	0	6.15	−0.6438	4.4337	3.315

Calculate $\bar{T}_M$:

From the table, we have

$$\bar{T}_M = \frac{1}{N}\sum_{i=1}^{N} T_i = -0.6438$$

Calculate $Var_M(T)$:

$$Var_M(T) = \frac{1}{N}\sum_{i=1}^{N} T_i^2 - \bar{T}_M{}^2$$

From the table, we have

$$Var_M(T) = 4.4337 - (-0.6438)^2 = 4.4337 - 0.4145 = 4.0192$$

Calculate $\sigma_M(T)$:

$$\sigma_M(T) = \sqrt{Var_M(T)} = \sqrt{4.0192} = 2.0048$$

Calculate $Cov(Y,T)$:

$$Cov(Y,T) = \frac{1}{N}\sum_{i=1}^{N} Y_i T_i - \bar{Y}_M \bar{T}_M = 3.315 - (-0.6438 \times 6.15)$$

So,

$$Cov(X,T) = 3.315 + 3.958 \approx 7.273$$

Determine c and d

$$c = \frac{Cov(Y,T)}{Var_M(Y)} = \frac{7.273}{25.442} = 0.285$$

and

$$d = \bar{T}_M - c\bar{Y}_M = -0.6438 - 0.285 \times 6.11 = -0.6438 - 1.740 \approx -2.384$$

Thus, the regression lines of T on Y is given by

$$T = cY + d = 0.285\,\text{Y} - 2.384$$

Determine a and b

$$\begin{cases} c = \ln(b) \Rightarrow b = e^c \\ d = \ln(a) \Rightarrow a = e^d \end{cases}$$

So,

$$b = e^c = e^{0.285} \approx 1.33$$

and

$$a = e^d = e^{-2.384} \approx 0.092$$

So, the equation of X on Y is given by

$$|X| = a \times b^Y = 0.092 \times 1.33^Y$$

Remark:

- The coefficient of determination between two statistical variables X and Y, denoted by $\rho^2(X,Y)$, can be used to identify the most suitable type of fit for a given dataset.
- The highest value among $\rho^2(X,Z)$ indicates the best fitting model to accurately describe the corresponding scatter plot.

9. **Series of Exercises 02:**

Exercise 01:
We are given the following table:

X \ Y	0	1	2	$n_{i.}$
[0,2[	10	4	a	17
[2,4[	15	10	b	.
[4,6[	3	6	3	.
$n_{.j}$	.	.	.	$N = 60$

1. Find the values of a and b and complete the table.
2. Draw the scatter plot of the points.
3. Calculate the marginal characters.

Solution 01:

1. Values of a and b:

We have,

$$a + 4 + 10 = 17$$

So,

$$a = 17 - 14 = 3$$

Then, the table

X \ Y	0	1	2	$n_{i.}$
[0,2[	10	4	3	17
[2,4[	15	10	b	25+b
[4,6[	3	6	3	12
$n_{.j}$	28	20	6+b	$N = 60$

So,

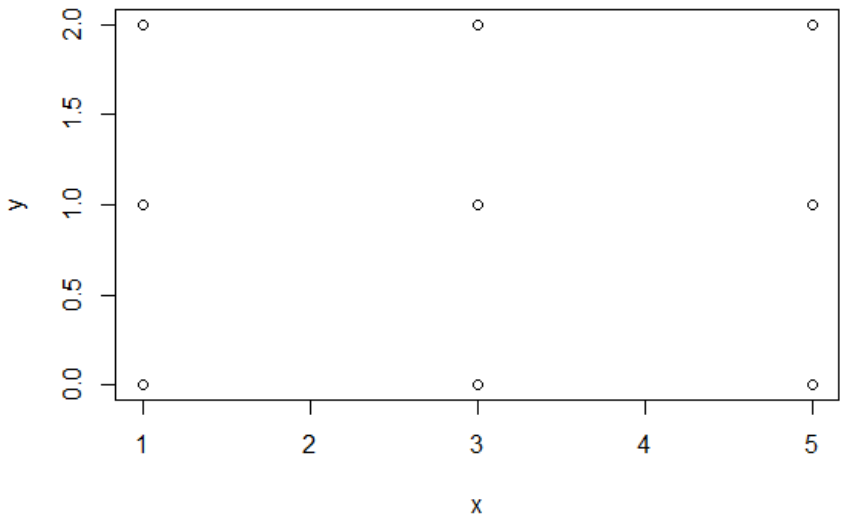
$$17 + 25 + b + 12 = 60$$

Thus,

$$b = 60 - 54 = 6$$

Y / X	0	1	2	$n_{i.}$
[0,2[	10	4	3	17
[2,4[	15	10	6	31
[4,6[	3	6	3	12
$n_{.j}$	28	20	12	$N = 60$

2. Draw the scatter plot of the points:



Statistical table: ($N = 60$)

Y / X	0	1	2	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$	$n_{ij}\,X_iY_j$
1	0 10	4 4	6 3	17	17	17	10
3	0 15	30 10	36 6	31	93	279	66
5	0 3	30 6	30 3	12	60	300	60
$n_{.j}$	28	20	12	$N = 60$	$\sum_{i=1}^{k} n_{i.}X_i$ =170	$\sum_{i=1}^{k} n_{i.}X_i^2$ =596	
$n_{.j}\,Y_j$	0	20	24	$\sum_{j=1}^{p} n_{.j}\,Y_j$ =44			

$n_{.j}\,Y_j^2$	0	20	48	$\sum_{j=1}^{p} n_{.j}\,Y_j^2 =$ 68			
$n_{ij}\,X_iY$	0	64	72				$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\,X_iY_j$ =136

3. Calculate the marginal characters:

Calculate $\bar{X}_M$: From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i = \frac{170}{60} = 2.8333$$

Calculate $Var_M(X)$:

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i.}X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = \frac{596}{60} - (2.8333)^2 = 9.9333 - 8.0276 = 1.9057$$

Calculate $\sigma_M(X)$:

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{1.9057} = 1.38$$

Calculate $\bar{Y}_M$: From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{p} n_{.j}\,Y_j = \frac{44}{60} = 0.7333$$

Calculate $Var_M(Y)$:

$$Var_M(Y) = \frac{1}{N}\sum_{i=1}^{p} n_{.j}Y_i^2 - \bar{Y}_M{}^2$$

From the table, we have

$$Var_M(Y) = \frac{68}{60} - (0.7333)^2 = 9.9333 - 8.0276 = 0.5956$$

Calculate $\sigma_M(Y)$:

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{0.5956} = 0.772$$

Exercise 02:

The number of hours spent on physical exercise by 70 individuals has been recorded before and after a wellness program. Let X represent the number of hours spent on exercise before the program, and Y represent the number of hours spent on exercise after the program. The following results have been obtained:

X \ Y	[90,100[	[100,110[	[110,120[	[120,130[	[130,140[
[60, 70[	4	2	1	0	0
[70, 80[	2	8	10	4	0
[80, 90[	0	6	12	7	1
[90,100[	0	0	5	5	1
[100,110[	0	0	0	1	1

1. Determinate the marginal distributions of X and Y.
2. Calculate the marginal means and the marginal standards deviations of X and Y.
3. Determine the conditional distribution of X given $Y \in [120,130[$.
4. Calculate $\bar{X}_{/Y\in[120,130[}$ and $\sigma(X/Y \in [120,130[)$.
5. Determine the conditional distribution of Y given $X \in [70,100[$.
6. Calculate $\bar{Y}_{/X\in[70,100[}$ and $\sigma(Y/X \in [70,100[)$.
7. Calculate $Cov(X, Y)$.

Solution 2:

1. The marginal distribution of X.

X_i	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$
65	7	455	29575
75	24	1800	135000

85	26	2210	187850
95	11	1045	99275
105	2	210	22050
Total	70	5720	473750
Total $/N$		81.714	6767.857

The marginal distribution of Y:

Y_j	$n_{.j}$	$n_{.j}Y_j$	$n_{.j}Y_j^2$
95	6	570	54150
105	16	1680	176400
115	28	3220	370300
125	17	2125	265625
135	3	405	54675
Total	70	8000	921150
Total $/N$		114.286	13159.286

2. Calculate the marginal means and the marginal standards deviations of X and Y.

Calculate $\bar{X}_M$: From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{\boldsymbol{k}} n_{i.}X_i = 81.714$$

Calculate $Var_M(X)$:

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{\boldsymbol{k}} n_{i.}X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = \frac{473750}{70} - (81.714)^2 = 90.679$$

Calculate $\sigma_M(X)$:

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{90.679} = 9.523$$

Calculate $\bar{Y}_M$: From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{j=1}^{p} n_{.j}\, Y_j = 114.286$$

Calculate $Var_M(Y)$:

$$Var_M(Y) = \frac{1}{N}\sum_{i=1}^{p} n_{.j} Y_i^2 - \bar{Y}_M{}^2$$

From the table, we have

$$Var_M(Y) = \frac{921150}{70} - (114.286)^2 = 97.996$$

Calculate $\sigma_M(Y)$:

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{97.996} = 9.899$$

3. The conditional distribution of X given $Y \in [120,130[$.

X_i	n_{i4}	$n_{i4}X_i$	$n_{i4}X_i^2$
65	0	0	0
75	4	300	22500
85	7	595	50575
95	5	475	45125
105	1	105	11025
Total	17	1475	129225
Total /17		86.765	7601.471

4. Calculate $\bar{X}_{/Y\in[120,130[}$ and $\sigma(\mathrm{X}/Y \in [120,130[)$.
We have

$$\bar{X}_{/Y\in[120,130[} = \frac{1}{17}\sum_{i=1}^{5} n_{i4}X_i = 86.765$$

and

$$\mathrm{Var}(\mathrm{X/Y} \in [120,130[\,) = \frac{1}{17}\sum_{i=1}^{5} n_{i4}X_i^2 - \left(\bar{X}_{/Y\in[120,130[}\right)^2 = 73.306$$

So,

1. $\sigma(X/Y \in [120,130[) = \sqrt{\mathrm{Var}(X/Y \in [120,130[)} = 8.562$

5. Determine the conditional distribution of Y given $X \in [70,100[$.

We have

$Y/X \in [70,100[$ / X	[90,100[	[100,110[	[110,120[	[120,130[	[130,140[
[70,80[	2	8	10	4	0
[80,90[	0	6	12	7	1
[90,100[	0	0	5	5	1
$(n_{j/X\in[70,100]})$ Total	2	14	27	16	2

So, the conditional distribution of Y given $X \in [70,100[$ is as follows.

Y_j	$n_{j/X\in[70,100[}$	$n_{j/X\in[70,100[}Y_j$	$n_{j/X\in[70,100[}Y_j^2$
95	2	190	18050
105	14	1470	154350
115	27	3105	357075
125	16	2000	250000
135	2	270	36450
Total	61	7035	815925
Total /61		115.328	13375.820

6. Calculate $\bar{Y}_{/X\in[70,100[}$ and $\sigma(\mathrm{Y}/X \in [70,100[)$.
We have

$$\bar{Y}_{/X\in[70,100[} = \frac{1}{61}\sum_{j=1}^{5} n_{j/X\in[70,100[} Y_j = 115.328$$

$$\mathrm{Var}\left(Y/X \in \left[70,100[\right) = \frac{1}{61}\sum_{j=1}^{5} n_{j/X\in[70,100[}\, Y_j^2 - \left(\bar{Y}_{/X\in[70,100[}\right)^2 = 75.272$$

$$\text{and } \sigma(Y/X \in [70,100[) = \sqrt{\mathrm{Var}(Y/X \in [70,100[)} = 8.676$$

7. Calculate $Cov(X, Y)$.
We have

X \ Y	95	105	115	125	135	$\sum_{j=1}^{5} n_{ij}X_iY_j$
65	4 [24700]	2 [13650]	[7475] 1	[0] 0	[0] 0	45825
75	[14250] 2	[63000] 8	[86250] 10	[37500] 4	[0] 0	201000
85	[0] 0	[53550] 6	[117300] 12	[74375] 7	[11475] 1	256700
95	[0] 0	[0] 0	[54625] 5	[59375] 5	[12825] 1	126825
105	[0] 0	[0] 0	[0] 0	[13125] 1	[14175] 1	27300
$\sum_{i=1}^{5} n_{ij}x_iy_j$	38950	130200	265650	184375	38475	657650

So

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{5}\sum_{j=1}^{5} n_{ij}X_iY_j - \bar{X}_M\bar{Y}_M = \frac{657650}{70} - (81.714)(114.286) = 56.234$$

Exercise 03:

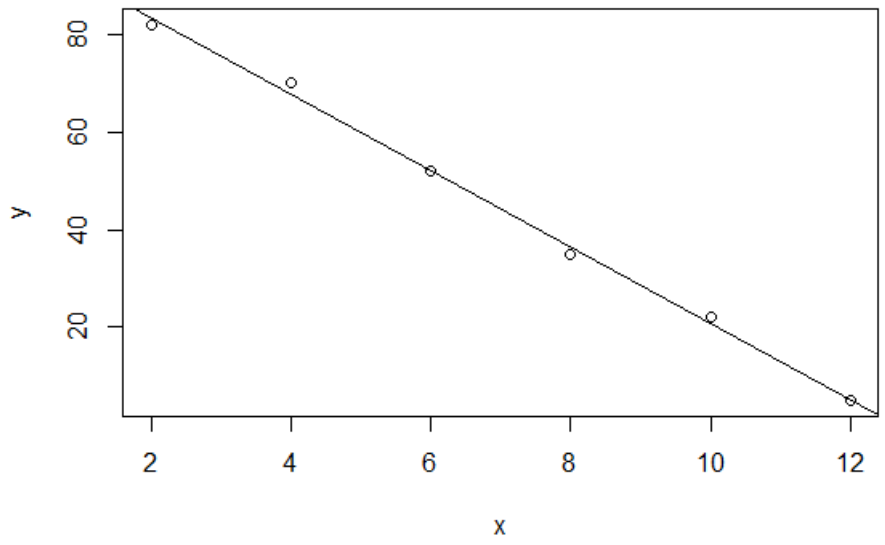
A pair of data was observed, and the following results were obtained:

X	2	4	6	8	10	12
Y	82	70	52	35	22	5

1. Plot this statistical series on a graph.
2. What do you notice and what can you conclude?
3. Find the regression line Y as a function of X and plot it with the scatter plot.
4. Give an estimate of Y when $X = 9$.

Solution 03:

1. Plot the statistical series on a graph:



2. What do you notice and what can you conclude?

   The plotted points lie roughly along a straight line.

   The line has a negative slope, meaning as X increases, Y decreases.

   The data points are close to the line, showing a strong linear relationship between Y and X.

3. Regression line Y as a function of X is given by $Y = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var_M(X)} \\ b = \bar{Y}_M - a\bar{X}_M \end{cases}$$

So, we need the calculate $\bar{Y}_M$, $\bar{X}_M$, $Var_M(X)$ and $Cov(X,Y)$.

Statistical table: ($N = 6$)

	X_i	Y_i	X_i^2	X_iY_i
	2	82	4	164
	4	70	16	280
	6	52	36	312
	8	35	64	280
	10	22	100	220
	12	5	144	60
Σ	42	266	364	1316
Σ/N	7	44.333	60.667	219.333

Calculate $\bar{X}_M$:

From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{N} X_i = 7$$

Calculate $\bar{Y}_M$:

From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{i=1}^{N} Y_i = 44.333$$

Calculate $Var_M(X)$:

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = 60.667 - (7)^2 = 60.667 - 49 = 11.667$$

Calculate $Cov(X,Y)$:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i - \bar{X}_M \bar{Y}_M = 219.333 - (7 \times 44.333)$$

So,

$$Cov(X,Y) = 219.333 - 310.331 = -90.998$$

Then,

$$a = \frac{Cov(X,Y)}{Var_M(X)} = \frac{-90.998}{11.667} \approx -7.8$$

and

$$b = \bar{Y}_M - a\bar{X}_M = 44.333 + 7.8 \times 7 = 44.333 + 54.6 = 98.933$$

Thus, the regression lines of Y on X is given by

$$Y = aX + b = -7.8\text{X} + 98.933$$

4. Estimate Y when $X = 9$:

$$Y = -7.8 \times 9 + 98.933 == -70.2 + 98.933 = 28.733$$

So, when $X = 9$, $Y \approx 28.733$.

Exercise 04: (Continuation of Exercise 01)

1. Calculate the linear correlation coefficient $\rho(X,Y)$ and interpret the result.
2. Find the regression line of Y in terms of X.

Solution 04:

1. Calculate the linear correlation coefficient $\rho(X,Y)$:

Calculate $Cov(X,Y)$:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} X_i Y_j \ - \bar{X}_M \bar{Y}_M = \frac{136}{60} - (2.8333 \times 0.7333)$$

So,

$$Cov(X,Y) = 2.2667 - 2.0776 = 0.1891$$

Then, the correlation coefficient ρ between X and Y is given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\sigma_M(Y)} = \frac{0.1891}{1.38 \times 0.772} = \frac{0.1891}{1.0653} \approx 0.1775$$

Interpret the result:

The correlation coefficient $\rho \approx 0.177$, which indicates a weak positive correlation.

2. Regression line of Y in terms of X: is given by $Y = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var_M(X)} \\ b = \bar{Y}_M - a\bar{X}_M \end{cases}$$

So, we need the calculate $\bar{Y}_M$, $\bar{X}_M$, $Var_M(X)$ and $Cov(X,Y)$.

$$a = \frac{Cov(X,Y)}{Var_M(X)} = \frac{0.1891}{1.9057} \approx 0.0993$$

and

$$b = \bar{Y}_M - a\bar{X}_M = 0.7333 - 0.0993 \times 2.8333 = 0.4522$$

Thus, the regression lines of Y on X is given by

$$Y = aX + b = 0.0993\text{X} + 0.4522$$

Exercise 05:

To study the effect of a chemical substance on the spike-stem length of a wheat variety, the following experiment was carried out. The average stem length was measured for each experimental unit, giving the following table.

X Quantity of chemical	2	4	6	8	10
Y Average stem length	70.2	72.1	74.2	75.9	77.8

1. Compute the marginal arithmetic means $\bar{X}_M$ and $\bar{Y}_M$.
2. Compute the marginal standard deviations $\sigma_M(X)$ and $\sigma_M(Y)$.
3. Compute the correlation coefficient ρ between X and Y.
4. Plot the data graphically.

5. Determine the regression line of X on Y and the regression line of Y on X, and plot them.

Solution 05:

Statistical table: $(N = 5)$

	X_i	Y_i	X_i^2	Y_i^2	X_iY_i
	2	70.2	4	4928.04	140.4
	4	72.1	16	5198.41	288.4
	6	74.2	36	5505.64	445.2
	8	75.9	64	5760.81	607.2
	10	77.8	100	6052.84	778.0
Σ	30	370.2	220	27445.74	2259.2
$\Sigma/_N$	6	74.04	44	5489.148	451.84

1. Marginal **a**rithmetic Means:

Mean of X:

From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{N} X_i = 6$$

Mean of Y:

From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{i=1}^{N} Y_i = 74.04$$

2. Marginal standard deviations:

Standard deviation of X:

Firstly, we calculate the marginal variance of X.

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = 44 - (6)^2 = 8$$

Thus,

$$\sigma_M(X) = \sqrt{Var_M(X)} = \sqrt{8} = 2.828$$

Firstly, we calculate the marginal variance of Y.

$$Var_M(Y) = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}_M{}^2$$

From the table, we have

$$Var_M(Y) = 5489.148 - (74.04)^2 = 7.226$$

Thus,

$$\sigma_M(Y) = \sqrt{Var_M(Y)} = \sqrt{7.226} = 2.688$$

3. Correlation coefficient ρ between X and Y:

Firstly, we calculate the covariance of X and Y

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i\, Y_i - \bar{X}_M \bar{Y}_M = 451.84 - (6 \times 74.04)$$

So,

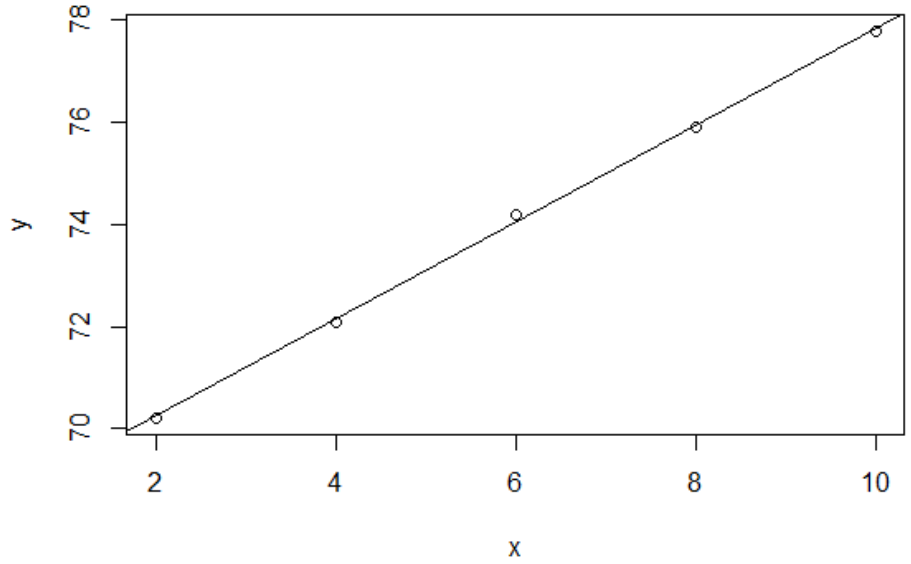
$$Cov(X,Y) = 451.84 - 444.24 = 7.60$$

Finally, the correlation coefficient ρ between X and Y is given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)} = \frac{7.60}{2.828 \times 2.688} = \frac{7.60}{7.6016} \approx 0.9998$$

Thus, there is an almost perfect positive correlation.

4. Plot the data graphically:



5. Regression lines:

Regression lines of Y on X is given by $Y = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var_M(X)} \\ b = \bar{Y}_M - a\bar{X}_M \end{cases}$$

Or,

$$a = \frac{Cov(X,Y)}{Var_M(X)} = \frac{7.60}{2.828} \approx 0.95$$

and

$$b = \bar{Y}_M - a\bar{X}_M = 74.04 - 0.95 \times 6 = 74.04 - 5.70 = 68.34$$

Thus, the regression lines of Y on X is given by

$$Y = aX + b = 0.95X + 68.34$$

Regression line of X on Y is given by $X = cY + d$, where

$$\begin{cases} c = \dfrac{Cov(X,Y)}{Var_M(Y)} \\ d = \bar{X}_M - c\bar{Y}_M \end{cases}$$

Or,

$$c = \frac{Cov(X,Y)}{Var_M(Y)} = \frac{7.60}{7.226} \approx 1.052$$

and

$$d = \bar{X}_M - c\bar{Y}_M = 6 - 1.052(74.04) = 6 - 77.86 = -71.86$$

Thus, the regression lines of Y on X is given by

$$X = cY + d = 1.052\text{Y} - 71.86$$

Exercise 06:

1. X, Y, Z and T are statistical (random) variables such that

$$\sigma_M(X) = 150,\ \sigma_M(Y) = 109,\ Cov(X,Y) = 1050$$

and

$$\sigma_M(T) = 0.6,\ \sigma_M(Z) = 0.8 \text{ and } Cov(T,Z) = 0.41$$

   a. What can be said about the relationship between (X, Y)?
   b. What can be said about the relationship between (T, Z)?

2. If $\sigma_M(X) = 0.9,\ \sigma_M(Y) = 1.2,\ Cov(X,Y) = 2.3$. Are these data correct or incorrect? Justify your answer.

Solution 06:

1. a. Relationship between (X, Y):

Given

$\sigma_M(X) = 150,\ \sigma_M(Y) = 109$ and $Cov(X,Y) = 1050$

Compute the correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)} = \frac{1050}{150 \times 109} \approx 0.064$$

Nothe that, the correlation is very small and positive (0.064). Therefore, X and Y have a very weak positive linear relationship**.**

b. Relationship between (Z, T):

Given

$$\sigma_M(T) = 0.6,\ \sigma_M(Z) = 0.8 \text{ and } Cov(T,Z) = 0.41$$

Compute the correlation

$$\rho(T,Z) = \frac{Cov(T,Z)}{\sigma_M(T)\,\sigma_M(Z)} = \frac{0.41}{0.6 \times 0.8} \approx 0.854$$

Nothe that, the correlation is strong and positive (0.854). Therefore, T and Z have a strong positive linear relationship.

2. Are the following data valid?

Given

$$\sigma_M(X) = 0.9,\ \sigma_M(Y) = 1.2 \text{ and } Cov(X,Y) = 2.3$$

Compute correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_M(X)\,\sigma_M(Y)} = \frac{2.3}{0.9 \times 1.2} \approx 2.13$$

Or, the correlation must satisfy

$$-1 \le \rho(X,Y) \le 1$$

But here

$$\rho(X,Y) > 1$$

So, these data are impossible / invalid.

Note that, in this time the covariance is too large relative to the standard deviations. Such values cannot exist because they imply a correlation greater than 1.

Exercise 07:

1. Consider the following table

Y / X	1	3
-1	0	4
0	5	0
1	0	4

1.1. Draw a scatter plot of Y against X.
1.2. Is there a linear relationship between X and Y?
1.3. Determine the equation of the regression line Y as function of X.

2. Let $Z = X^2$. Complete the following table:

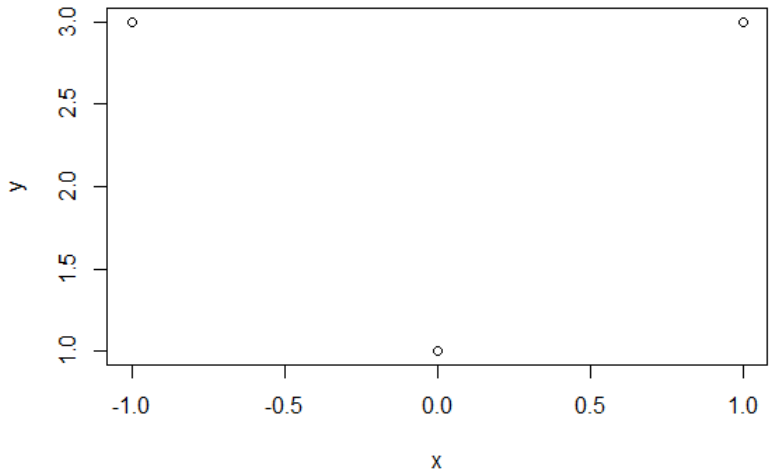
Y / Z	*1*	*3*

2.1. Calculate the arithmetic means of Y^2, Z and Z^2.
2.2. What is the type of relationship between Z and Y?
2.3. Deduce the type of relationship between X and Y.

Solution 07:

1. For X and Y:

1.1. Scatter plot of Y against X:



1.2. Is there a linear relationship between X and Y? No, there is not a linear relationship between X and Y (see the scatter plot).

1.3. Regression line Y as a function of X is given by $Y = aX + b$, where

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var_M(X)} \\ b = \bar{Y}_M - a\bar{X}_M \end{cases}$$

So, we need the calculate $\bar{Y}_M$, $\bar{X}_M$, $Var_M(X)$ and $Cov(X,Y)$.

Statistical table: ($N = 13$)

X \ Y	1	3	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$	$n_{ij}\,X_iY_j$
-1	0 0	-12 4	4	-4	4	-12
0	0 5	0 0	5	0	0	0
1	0 0	12 4	4	4	4	12
$n_{.j}$	5	8	$N = 13$	$\sum_{i=1}^{k} n_{i.}X_i = 0$	$\sum_{i=1}^{k} n_{i.}X_i^2 = 8$	
$n_{.j}\,Y_j$	5	24	$\sum_{j=1}^{p} n_{.j}\,Y_j = 29$			
$n_{.j}\,Y_j^2$	5	72	$\sum_{j=1}^{p} n_{.j}\,Y_j^2 = 77$			
$n_{ij}\,X_iY_j$	0	0				$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\,X_iY_j = 0$

Calculate $\bar{X}_M$:

From the table, we have

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{p} n_{i.}X_i = 0$$

Calculate $\bar{Y}_M$:

From the table, we have

$$\bar{Y}_M = \frac{1}{N}\sum_{i=1}^{p} n_{.j}\,Y_j = \frac{29}{13} = 2.2307$$

Calculate $Var_M(X)$:

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{\boldsymbol{p}} n_{i.}X_i^2 - \bar{X}_M{}^2$$

From the table, we have

$$Var_M(X) = \frac{8}{13} - (0)^2 = 0.6154 \ - 0 = 0.6154$$

Calculate $Cov(X,Y)$:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij}\, X_i Y_j \ - \bar{X}_M\bar{Y}_M = \frac{0}{13} - (0 \times 2.2307)$$

So,

$$Cov(X,Y) = 0 - 0 = 0$$

Then,

$$a = \frac{Cov(X,Y)}{Var_M(X)} = \frac{0}{0.6154} = 0$$

and

$$b = \bar{Y}_M - a\bar{X}_M = 2.2307 - 0 \times 0 = 2.2307 - 0 = 2.2307$$

Thus, the regression lines of Y on X is given by

$$Y = aX + b = 0\,\mathrm{X} + 2.2307$$

So,

$$Y = 2.2307$$

2. Let $Z = X^2$. Complete the following table:

Z \ Y	1	3	$n_{i.}$	$n_{i.}Z_i$	$n_{i.}Z_i^2$	$n_{ij}\,X_iZ_j$
1	0 0	24 8	8	8	8	24
0	0 5	0 0	5	0	0	0
$n_{.j}$	5	8	$N = 13$	$\sum_{i=1}^{2} n_{i.}Z_i$ =8	$\sum_{i=1}^{2} n_{i.}Z_i^2$ =8	
$n_{.j}\,Y_j$	5	24	$\sum_{j=1}^{2} n_{.j}\,Y_j$ =29			
$n_{.j}\,Y_j^2$	5	72	$\sum_{j=1}^{2} n_{.j}\,Y_j^2$ =77			
$n_{ij}\,Z_iY_j$	0	24				$\sum_{i=1}^{k}\sum_{j=1}^{2} n_{ij}\,Z_iY_j$ =24

2.1.Arithmetic means of Y^2, Z and Z^2.

Calculate $\bar{Z}_M$:

From the table, we have

$$\bar{Z}_M = \frac{1}{N}\sum_{i=1}^{2} n_{i.}Z_i = \frac{8}{13} = 0.6153$$

Calculate $\overline{Y^2}$:

From the table, we have

$$\overline{Y^2}_M = \frac{1}{N}\sum_{j=1}^{2} n_{.j}\,Y_j^2 = \frac{77}{13} = 5.9230$$

Calculate $\overline{Z^2}$:

From the table, we have

$$\overline{Z^2}_M = \frac{1}{N}\sum_{j=1}^{2} n_{i.}Z_i^2 = \frac{8}{13} = 0.6153$$

2.2.What is the type of relationship between Z and Y?

As Z increases from $0 \rightarrow 1$, Yincreases from $1 \rightarrow 3$.

This is a linear relationship:

$$Y = 2Z + 1$$

2.3.Deduce the type of relationship between X and Y.

Since $Z = X^2$ and Y is linear in Z, so,

$$Y = 2Z + 1 = 2X^2 + 1$$

Thus, Y is a **quadratic function** of X. (see also the above Scatter plot).

Exercise 08:
Consider the following set of data.

X	2	3	4	5	6	7	8	9	10
Y	1.1	0.8	0.45	0.28	0.3	0.26	0.22	0.2	0.15

1. Fit this set of data using a hyperbola of the form $Y = \frac{1}{aX+b}$.
2. Fit the same set of data using a logarithmic function of the form $Y = c\ln(X) + d$.
3. Estimate the value of Y when X takes the value 12, using the two models.
4. Which estimation is the most accurate?

Solution 08:

1. Fit the set of data using a hyperbola of the form $Y = \frac{1}{aX+b}$:

   We have

$$Y = \frac{1}{aX+b} \Longrightarrow \frac{1}{Y} = aX + b$$

We set $Z = \frac{1}{Y}$, so $Z = aX + b$.
Therefore,

$$\begin{cases} a = \dfrac{Cov(X,Z)}{Var_M(X)} \\ b = \bar{Z}_M - a\bar{X}_M \end{cases}$$

X_i	Y_i	Z_i	X_i^2	X_iZ_i	T_i	T_i^2	T_iY_i	Z_i^2	Y_i^2
2	1.1	0.909	4	1.818	0.693	0.480	0.762	0.826	1.21
3	0.8	1.25	9	3.75	1.099	1.207	0.879	1.562	0.64
4	0.45	2.222	16	8.888	1.386	1.922	0.624	4.937	0.202
5	0.28	3.571	25	17.855	1.609	2.590	0.451	12.752	0.078
6	0.3	3.333	36	19.998	1.792	3.210	0.538	11.109	0.09
7	0.26	3.846	49	26.922	1.946	3.787	0.506	14.792	0.068

8	0.22	4.545	64	36.36	2.079	4.324	0.457	20.657	0.048
9	0.2	5	81	45	2.197	4.828	0.439	25	0.04
10	0.15	6.667	100	66.67	2.303	5.302	0.345	44.449	0.022
54	3.76	31.343	384	227.261	15.104	27.65	5.001	136.084	2.398
6	0.418	3.483	42.667	25.251	1.678	3.072	0.556	15.120	0.266

From the table, we get

$$\bar{X}_M = \frac{1}{N}\sum_{i=1}^{9} X_i = 6, \qquad \bar{Z}_M = \frac{1}{N}\sum_{i=1}^{9} Z_i = 3.483$$

$$Var_M(X) = \frac{1}{N}\sum_{i=1}^{9} X_i^2 - (\bar{X}_M)^2 = 6.667$$

$$Cov(X,Z) = \frac{1}{N}\sum_{i=1}^{9} X_i Z_i - \bar{X}_M \bar{Z}_M = 4.353$$

$$\Rightarrow \begin{cases} a = \dfrac{Cov(X,Z)}{Var_M(X)} = 0.653 \\ b = \bar{Z}_M - a\bar{X}_M = -0.435 \end{cases}$$

Therefore,

$$Y = \frac{1}{aX+b} = \frac{1}{(0.653)X - 0.435}$$

2. Fit the set of data using a logarithmic function of the form $Y = c\ln(X) + d$:

We set $T = \ln(X)$, so $Y = cT + d$.
Therefore

$$\begin{cases} c = \dfrac{Cov(T,Y)}{Var_M(T)} \\ d = \bar{Y}_M - c\bar{T}_M \end{cases}$$

From the table, we have

$$\bar{T}_M = \frac{1}{N}\sum_{i=1}^{9} T_i = 1.678, \quad \bar{Y}_M = \frac{1}{N}\sum_{i=1}^{9} Y_i = 0.418$$

$$Var(T) = \frac{1}{N}\sum_{i=1}^{9} T_i^2 - (\bar{T}_M)^2 = 0.256$$

$$Cov(T,Y) = \frac{1}{N}\sum_{i=1}^{9} T_i Y_i - \bar{T}_M \bar{Y}_M = -0.145$$

$$\Longrightarrow \begin{cases} c = \dfrac{Cov(T,Y)}{Var_M(T)} = -0.566 \\ d = \bar{Y}_M - c\bar{T}_M = 1.368 \end{cases}$$

Thus,

$$Y = c\ln(X) + d = (-0.566)\ln(X) + 1.368$$

3. Estimate the value of Y when X takes the value 12, using the two models:

   - Using the first model:

$$X = 12 \Longrightarrow Y = \frac{1}{(0.653)X - 0.435} = \frac{1}{(0.653) \times 12 - 0.435} = 0.135$$

   - Using the second model:

$$X = 12 \Longrightarrow Y = (-0.566)\ln(X) + 1.368 = -0.038$$

4. Which estimation is the most accurate:

We have to compare $\rho^2(X,Z)$ and $\rho^2(T,Y)$

$$\rho(X,Z) = \frac{Cov(X,Z)}{\sigma_M(X)\sigma_M(Z)}$$

$$Var(Z) = \frac{1}{N}\sum_{i=1}^{9} Z_i^2 - (\bar{Z}_M)^2 = 2.989$$

$$\Longrightarrow \rho(X,Z) = \frac{4.353}{\sqrt{6.667 \times 2.989}} = 0.975$$

$$\Longrightarrow \rho^2(X,Z) = 0.951$$

Moreover,

$$\rho(T,Y) = \frac{Cov(T,Y)}{\sigma_M(T)\sigma_M(Y)}$$

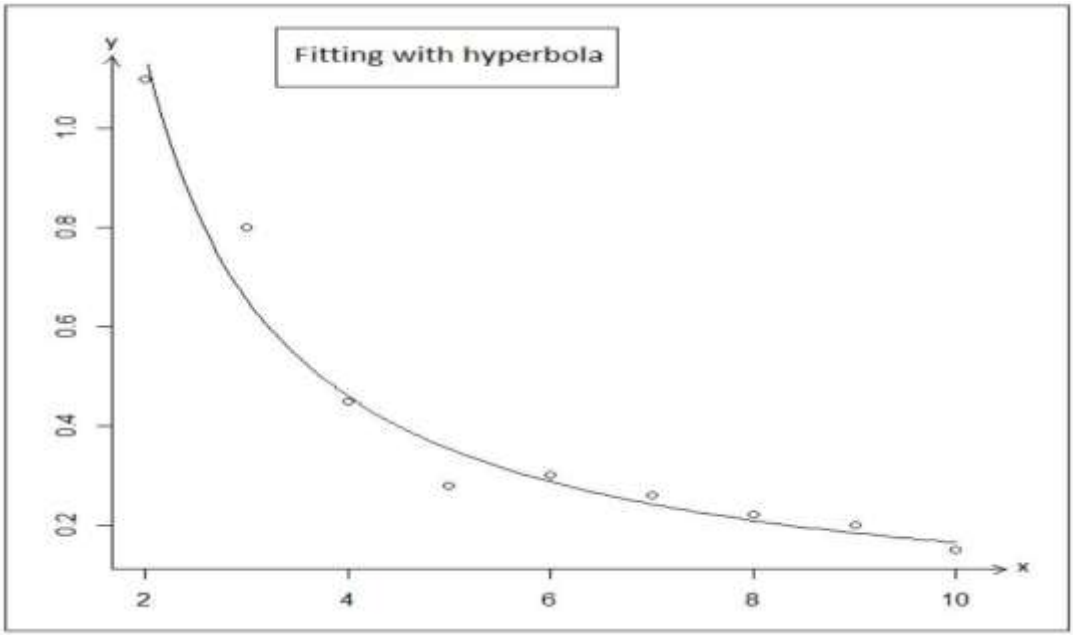
$$Var(Y) = \frac{1}{N}\sum_{i=1}^{9} Y_i^2 - (\bar{Y}_M)^2 = 0.091$$

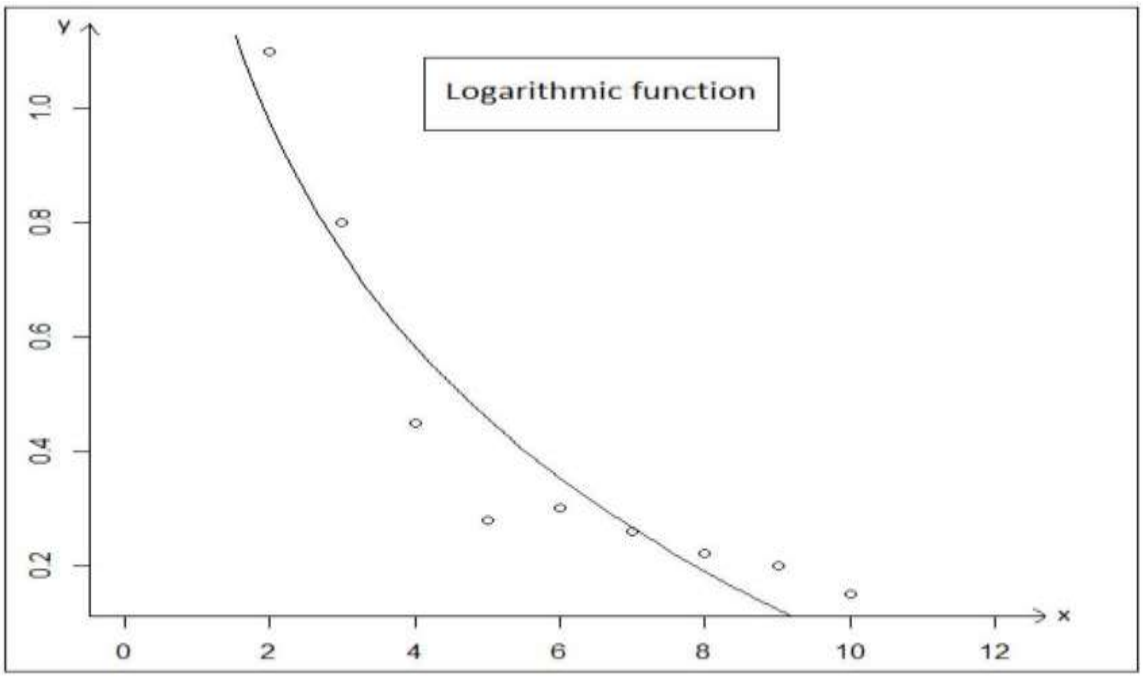
$$\Longrightarrow \rho(T,Y) = \frac{-0.145}{\sqrt{0.256 \times 0.091}} = -0.950$$

$$\Longrightarrow \rho^2(T,Y) = 0.903$$

Note that, $\rho^2(X,Z) = 0.951 > \rho^2(X,Z) = 0.903$, so the estimation by the first model (hyperbola) is the most accurate.

The scatter plot of this set of data as well as the fitting curves:




Logarithmic function
y
x
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12

IV. Bibliography

[1] Bernstein S, and Bernstein R. (1999), *Elements of Statistics I: Descriptive Statistics and Probability* McGraw-Hill Companies.

[2] Boukeloua Mohamed, Polycopie de cours, Probability and Statistique. *Ecole nationale polytechnique Constantine-Malek Bennabi.*

[3] Chibat A, (2004), *Statistique Descriptive,* Universite Mentouri Constantine.

[4] Dagnelie P. Théorie et méthodes statistiques, 1973, Presses agronomiques de Gembloux.

[5] Freund, J (2001) Modern Elementary Statistics 10th Edition, Printice Hall.

[6] Hamon A. and Nicolas J. (2008), *Statistique descriptive : Cours et exercices corriges,* Presses universitaires de Rennes.

[7] Heumann C,, Sehomaker M, and Shalabh (2016), *Introduction to statistics and Data Analysis,* Springer International Publishing, Switzerland.

[8] Neeter, J, Waserman, Whitmare, (1993): Applied Statistics. 4th Edition, Louise Richardson.

[9] Rosen K, H,, Michaels J, G,, Gross J, L,, Grossman J, W, and Shier, D, R, (2000), *Handbook of Discrete and Combinatorial Mathematics,* CRC Press LLC, Boca Raton.

[10] Probabilités et statistique cours et problèmes Série Schaum.