Higher School of Commerce

Lectures and Exercises in the Descriptive Statistics

First-year preparatory students

Prepared by: Dr. Boucherite Kamel

Contents

Chapter1:
Basic concepts of statistics

Chapter1: Basic concepts of statistics.

I- **Introduction**:

The science of statistics emerged with the birth of the first social structures, when the earliest written texts dealt with the census of livestock and information about their prices. Evidence of census operations has been found in China in the 22nd century BC and later in Egypt in the 17th century BC, and this system of data collection continued for a long time. During the Middle Ages, the census was used by the state to record members of society for the purpose of forming armies, documenting agricultural production, collecting taxes, and counting births and deaths.

The development of societies has generated the need to organize and summarize the collected data and data in a clear form that is easy to read and use, and this is what was later known as the science of the state or the science of statistics. The word statistic is derived from the Latin word status and the Italian word STATICA, which means state.

Statistical techniques have evolved based on mathematics, especially in the field of probability and inferential statistics. with the development of the use of automated media, which greatly facilitated the application of statistical methods in various fields of life and helped greatly in interpreting and understanding many phenomena in a way that is reasonably close to the actual reality, which helps to make good decisions.

II- Concepts and definitions:

1- **The concept of statistics**:

The multiple definitions of the science of statistics generally indicate that statistics is a branch of mathematics that is concerned with studying and understanding various phenomena and behaviors and even predicting them by relying on a set of statistics and information about these phenomena and behaviors. Statistics are numerical data related to the phenomena and subjects of study presented in an organized and exploitable form so that appropriate statistical methods can be applied to analyze them and extract results that help to understand the studied phenomenon to an acceptable extent and be able to make decisions with the least possible risk.

As examples of data, we find:

- Population statistics for a country;
- Annual production statistics for the industrial sector;
- Statistics of the higher education sector of a country;
- Statistics on the amounts of national consumption and national income of a society.

2- **Types of Statistics.**

Statistics is divided into two main sections:

2.1- **Descriptive statistics**:

It is the branch of statistics that is concerned with describing and understanding a particular phenomenon by preparing figures and graphs and calculating some numerical indicators based on numerical data and data about the studied phenomenon after they have been collected and organized in a logical and effective manner.

2.2- **Inferential Statistics:**

It is the branch of statistics that looks for how to infer results about a certain phenomenon for the studied community based on data and data taken from a sample drawn from this community according to scientific rules based mainly on the theory of probability, such as estimation methods, hypothesis testing, etc. This branch of statistics also helps in making decisions in cases of uncertainty.

3- **Concepts of some basic terms.**

3.1- **Statistics**:

Statistics is the science of data. This involves collecting, classifying, organizing, summarizing, analyzing and interpreting numerical information and data. It is both the science of uncertainty and the technology of extracting information from data. Used to help us make decisions.

3.2- **Statistical population and statistical sample** :

a)- **Statistical population** :

A population is the collection or set of all objects or measurements (individuals, things , animals , events) that are of interest to the collector (interest

to research or analysis) .This is based on set of shared properties that some grouping have in common.

The statistical population is the entire group that we want to draw conclusion about it.

b/ **Statistical sample**:

Statistical sample is the specific group that we will collect data from, the size of sample is always less than the total size of the population.

Example: all countries of the world is the population, but the countries with published data available on birth rates and gross domestic product (GDP) since 2000 is a sample.

3.3- Statistical variable:

A variable is a measure of character we want study, such as the tallness, the weight, prices of goods, academic level,

a / **Types of variables** : we distinguish two types of variables;

- **Qualitative** (categorical) **variable**:

It expresses traits (adjectives) of individuals, such as eyes color, hair color and academic level.

- **Quantitative** (numerical) **variable**:

It expresses the numerical values of individuals, such as height, weight, income, number of cars,

In this case, we have two categories:

a.1/ **Continuous variable**: it can take its values in a domain like wages, height,

a.2/ **Discontinuous (discrete) variable**: it cannot take its values in a domain like number of students.

b/ **Level (scales) of measurements** :

We have four level of measurements according to the type of variable, which are:

*** nominal level (scale)** : compatibility of adjectives that we can't rank , such as eyes color, sex , hair color, … .

*** Ordinal level**: compatibility of adjectives that we can rank, such as academic level.

*** interval level**: it is related to the variables that take values that can be compared (the difference between them has a meaning) but the ratio between

them has no meaning , because zero (0) does not express non-existence (the absence of the phenomenon). For example, in the case of temperatures, zero degrees does not mean the absence of heat.

*** Ratio level**: it is related to the variables that take values and meaningful zero point. Like height, weight, income, number of cars, the above can be summarized in the following table:

	level	variability	order	Similar interval	Meaningful zero point
Qualitative Variables (categorical)	- nominal - ordinal	yes yes	No yes	No No	No No
Quantitative variables	- interval - ratio	yes yes	Yes yes	Yes yes	No yes

4- Data types:

There are many types of data, such as:

4.1- Cross-sectional data:

It is a collection of data related to statistical units about a specific variable recorded at a specific moment in time (time is constant). For example, the distribution of commerce students according to the baccalaureate section for the year 2024. Is given in the following table:

BAC section	mathematics	Scientifics	management and economics	others	total
students number	95	910	695	55	1735

4.2- **Time series data:**

These are the data recorded about a statistical variable during successive points in time, those data that show the evolution of the values of the studied statistical variable over time (time is not constant).

For example, the distribution of the number of students at the higher school of commerce for the period 2019 to 2024 is given in the following table:

year	2019	2020	2021	2022	2023	2024
students number	1426	1405	1430	1420	1475	1735

4.3- panel data:

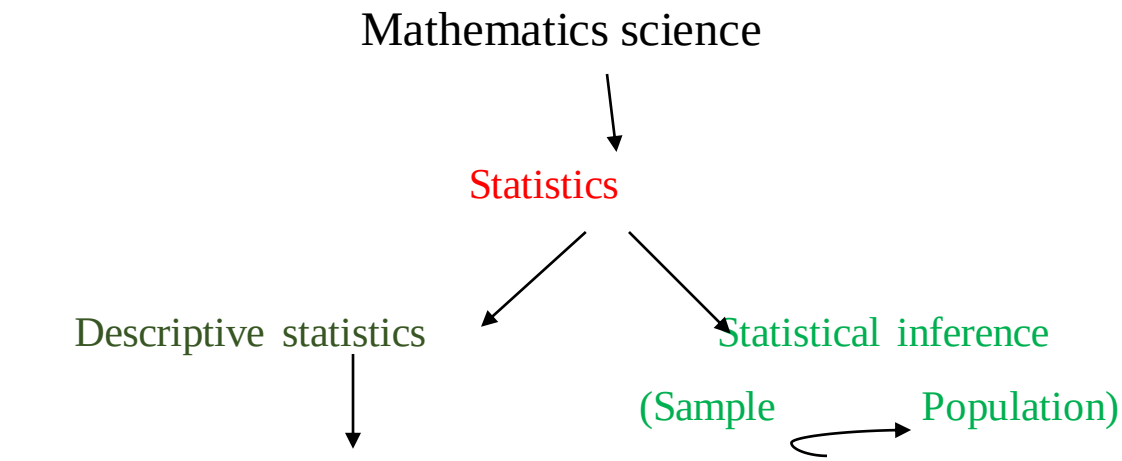
This type of data includes the first two types at the same time. The data shows the development of the components (categories) of a statistical variable over time. That is, for each total value of the variable at a specific moment in time, we give the values of its constituent elements. Panel data is a subset of longitudinal data where observations are for the same subjects each time.

For example, data showing the development of the number of students at the Higher School of Commerce for the years 2019 -2024 according to the baccalaureate division.

year / BAC division	2019	2020	2021	2022	2023	2024
mathematics	88	82	80	83	81	95
Scientifics	720	789	758	702	700	910
management	568	494	550	580	640	695
others	50	40	42	55	54	55
Total	1426	1405	1430	1420	1475	1735

Summary:

The most important elements of the above can be summarized in the form of a organogram as follows:



Definition: Statistics = collecting, classifying, organizing, summarizing, analyzing and interpreting numerical information and data

Basic concepts:

- **statistical population**
- **statistical sample**
- **statistical variables :** **Qualitative**
  Quantitative → **Continuous**
  Discontinuous (discrete)

- **Level (scales) of measurements**

Nominal level (scale)	Ordinal level	Interval level	Ratio level

Example 1:

In each of the following cases, indicate the statistical variable, its type, and the corresponding measurement scale.

1- Distribution of the number of patients according to the degree of response to treatment: no improvement, partial improvement, total improvement.

2- Distribution of the number of families in a city according to the number of children.

3- Marital status of Higher School of Commerce workers: married, single, divorced, widowed.

4- Distribution of the number of states in the country according to the number of traffic accidents during a month.

5- Distribution of students in a particular school according to blood type.

6- Distribution of the number of cities according to temperature

Solution:

We can represent the answer in the table as follows:

case	the variable	type of variable	scale
1	the degree of response to treatment	qualitative	ordinal
2	the number of children in the family	quantitative (discrete)	ratio
3	Marital status	qualitative	nominal
4	the number of traffic accidents	quantitative (discrete)	ratio
5	blood type	qualitative	nominal
6	temperature	quantitative (continuous)	interval

Example 2:

In each of the following cases, identify the studied characteristic, the statistical population, the studied variable, and its type:

1- The educational level of first-year preparatory students at the Higher School of Commerce for the 2024/2025 academic year, based on the students' annual GPA.

2- The standard of living in the city of Algiers in 2020, based on the average family income.

3- The cultural level of workers at the university pole in Kolea for the year 2024, based on the average number of books read by each worker.

Solution:

case	**Statistical population**	**Studied variable**	**type of variable**	**studied characteristic**
1	1st preparatory students at HSC at 2024-2025	the students annual GPA	continuous	the educational level
2	the Residents of the city of Algiers in 2020	the average family income	continuous	the standard of living
3	workers at the university pole in Kolea for the year 2024	average number of books	discrete	The cultural level of workers

Chapter2: Tabular and Graphical Representation of Data

Chapter2: Tabular and Graphical Representation of Data

After collecting of data, and for presenting it in an expressive way, we use statistical tables and graphs.

1- TABULAR REPRESENTATION OF DATA.

The tabular representation called the statistical distribution.

1.1- **qualitative variables**:

Categorical or qualitative variables allow us to put data into categories (X_i) and the absolute frequency (n_i). The frequency (n_i) is the number of elements of the sample that belong to that category (X_i).

The relative frequency of category (f_i) is the ratio of the frequency and total number of observations in the sample (n):

$$f_i = n_i/n$$

Example:

A group of production factory workers in BISKRA in May 2016 is distributed according to family status as follows:

Married, married , married , divorced, married ,widowed , single , divorced, widowed, married, single , divorced , married , single , married , single , married , single , married , married.

category x_i	frequency n_i	relative frequency f_i	f_i (%)
single	5	5/20=0.25	25
married	10	10/20=0.5	50
divorced	3	3/20=0.15	15
widower	2	2/20=0.1	10
total	20	1	100

***- The Ascending Cumulative Frequency (ACF):**

The Ascending Cumulative Frequency (ACF) is the sum of all previous frequencies up to the current point; it is often referred to as the running total of the frequencies.

The ascending cumulative frequency of a value " x " can be found by adding all the frequencies less or equal the frequency of "x".

If we symbolize the ascending cumulative frequency by $N(x_k)$, we have :

$$N(x_k) = \sum_{i=1}^{k} (n_i) \quad , where \ k = 1, \dots, K$$

$and \ K \ reepresents \ the \ number \ of \ possible \ values \ for \ the \ variable \ X.$

$N(x_k)$: indicates the frequency (number of individuals in the sample) of values that are less or equal to x_k .

***- The descending cumulative frequency (DCF):**

The descending cumulative frequency (DCF) of the value " x " indicates the frequency of values that are greater than or equal to " x " .

If we symbolize the descending cumulative frequency by $N^*(x_k)$, we have :

$$N^*(x_k) = n - \sum_{i=1}^{k} (n_{i-1}) \quad , where \ n_0 = 0 \ \ and \ \ k = 1, \dots, K$$

$N^*(x_k)$: indicates the frequency (number of individuals in the sample) of values that are greater than or equal to x_k .

Example:

The department of the industry and commerce of SETIF conducted a qualitative survey in the second trimester of 2016 on the development of the status of industrial activity compared to the previous trimester for 20 industrial establishments.

The results of the response to the question:

- is the level of activity of your establishment :

to retreat (decrease) , stability , to improve

the results are:

- stability , stability , improve , improve , decrease , improve , decrease , improve , stability , stability , decrease , decrease , improve , stability , stability , improve , decrease, stability, stability, improve.

Required :

1- What is the statistical population studied?
2- What is the property studied?
3- What is a variable? What type? What is the measurement scale?
4- Display these data in a statistical (calculate ACF and DCF).

- **Answer**:

1- Statistical population: industrial establishments in the second trimester of 2019 in SETIF.
2- The property studied the status of the industrial activity.
3- The variable: a qualitative assessment of the status of the industrial activity.
   - type: qualitative
   - measurement scale : ordinal
4- Statistical distribution.

x_i	n_i	f_i	f_i %	$N(x_i)$	$N^*(x_i)$
decrease	5	0.25	25	5	20
stability	8	0.40	40	5+8=13	20-5=15
improve	7	0.35	35	13+7=20	15-8=7
total	20	1	100		

1.2- quantitative variables :

a)- frequency distribution grouped without classes.

in the general case if we have the following preliminary statistical series (x_1 , x_2 , …., x_k) with the number of observations per value (x_j) it is (n_j), this series can be displayed in the form of statistical distribution as follow:

values x_j	frequencies n_j
x_1	n_1
x_2	n_2
.	.
.	.
x_j	n_j
.	.
.	.
x_k	n_k
total	n

This statistical distribution is generally used in the case of discrete variable.

Example:

The statistical distribution of the number of rooms in each dwelling in a city in SETIF is as follow:

1 , 1 , 1 , 1 , 2 , 2 , 2 , 2 , 2 , 2 , 2 , 2 , 2 , 2 , 3 , 3 , 3 , 3 , 3 , 3, 3 , 3 , 3 , 3 , 3 , 3 , 3 , 3 , 3 , 3 , 4 , 4 , 4 , 4 , 4 , 4 , 4 , 5 , 5 , 5.

1- Show this data in the form of a statistical distribution.

2- Calculate relative frequency, ascending cumulative frequency, descending cumulative frequency.

Answer:

x_j	1	2	3	4	5	total
n_j	4	10	16	7	3	40
f_j	0.1	0.25	0.4	0.175	0.075	1
$N(x_j)$	4	14	30	37	40	
$N^*(x_j)$	40	36	26	10	3	

b)- grouped frequency distribution with classes:

Often in many statistical studies, we find a large and frequent number of data or observations; it is impractical to use these data individually. In these cases, we summarize these many data in the form of classes. these classes are intervals that divide the length of data series (the range) to continuous portions that relate all the sightings collected from the thoughtful sample expressing variable values or data values (continuous quantitative variable)and each class we give the appropriate frequency which represents the number of variable values that belong to this class (interval).

To accomplish this we follow the following steps:

- We choose the number of classes **(J)**.
- we calculate a range (R) :

$$\mathbf{R=X_{max} - X_{min}}$$

- We select from 5 to 20 classes that in general are no overlapping intervals of equal length, so as to cover the entire range of data. The goal is to use enough classes to show the variation in the data, but not so many that there are only a few data points in many of the classes ,the class width should be slightly larger than the ratio **R/J :**

$$\textbf{Length of class (h) } > \textbf{R/J}$$

For every class "j" , $[\, x_j^- \; ; \; x_j^+ [$ we have:

- The lower boundary x_j^- ;
- The upper boundary x_j^+ ;
- The midpoint of class $x_j = \frac{x_j^+ + x_j^-}{2}$;

- The length class h_j ;
- The frequency of class n_j ;

Example:

The following data refer to a certain type of chemical impurity measured in parts per million in 25 drinking water samples randomly collected from different areas a country.

11 , 19 , 24 , 30 , 12 , 20 , 25 , 29 , 15 , 21 , 24 , 31 , 16 , 23 , 25 , 26 , 32 , 17 , 22 , 26 , 35 , 18 , 24 , 18 , 27.

- make a frequency table displaying class intervals (5 classes)
- Calculate the relative frequencies, ACF and DCF.

Solution:

We will use 5 classes (J=5)

- the maximum value : X_{max} =35
- the minimum value : X_{min} =11
- the range : R= 35-11=24
- R/J= 24/5= 4.8 → then the length of class is **h=5**

Hence, we shall take the class width to be 5. The lower boundary of the first class interval will be chosen to be 10.5

The upper boundary of the fifth class becomes 35.5

We can now construct the frequency table as follow:

N° of class	primary class	real class	frequency n_j	midpoint of class x_j	f_j (%)	ACF $N(x_j)$	DCF $N^*(x_j)$
1	[11 ; 15]	[10.5 ; 15.5[	3	13	12	3	25
2	[16 ; 20]	[15.5 ; 20.5[	6	18	24	9	22
3	[21 ; 25]	[20.5 ; 25.5[	8	23	32	17	16
4	[26 ; 30]	[25.5 ; 30.5[	5	28	20	22	8
5	[31 ; 35]	[30.5 ; 35.5[	3	33	12	25	3
total			25		100		

2-GRAPHICAL REPRESENTATION OF DATA.

Now we shall introduce some ways of graphically representing both qualitative and quantitative data.

2.1 Qualitative variable:

*- **Bar graph**: It is a graph of bars whose heights represent the frequencies (or relative frequencies) of respective categories.

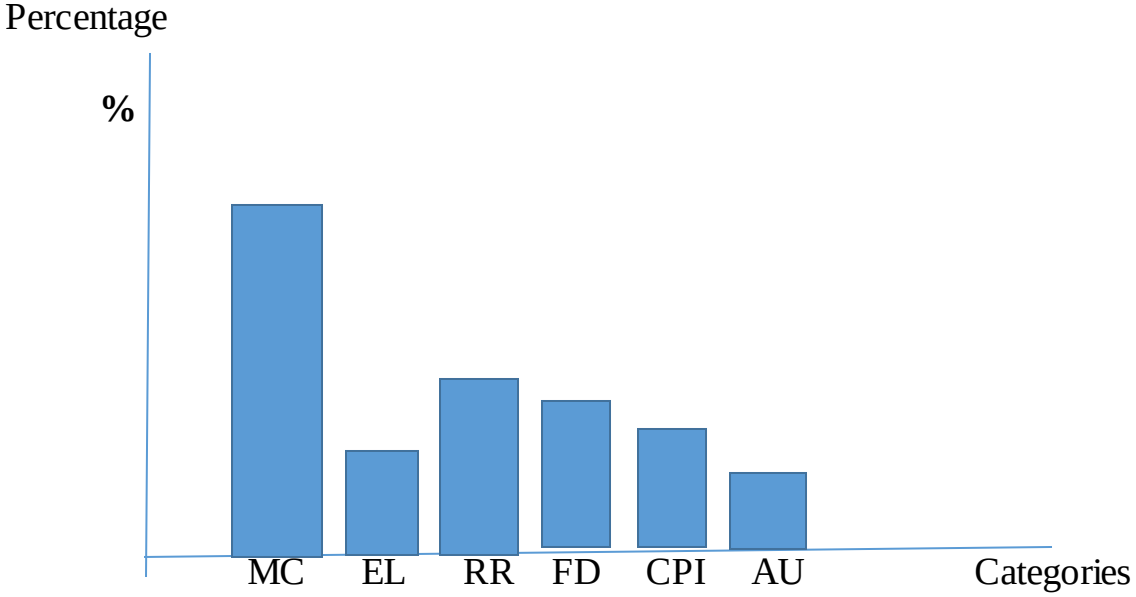
Example:

The data in next table represent the percentages of price increases of some consumer goods and services for the period 12.1990 to 12.2000 in a certain city.

medical care (MC)	electricity (EL)	residential rate (RR)	Food (FD)	consumer price index (CPI)	Apparel upkeep (AU)
83.3 %	22.1%	43.5 %	41.1 %	35.8 %	21.2 %

The graphic representation of this data is given as follow:

- The Bar Graph -



***- Pie chart:**

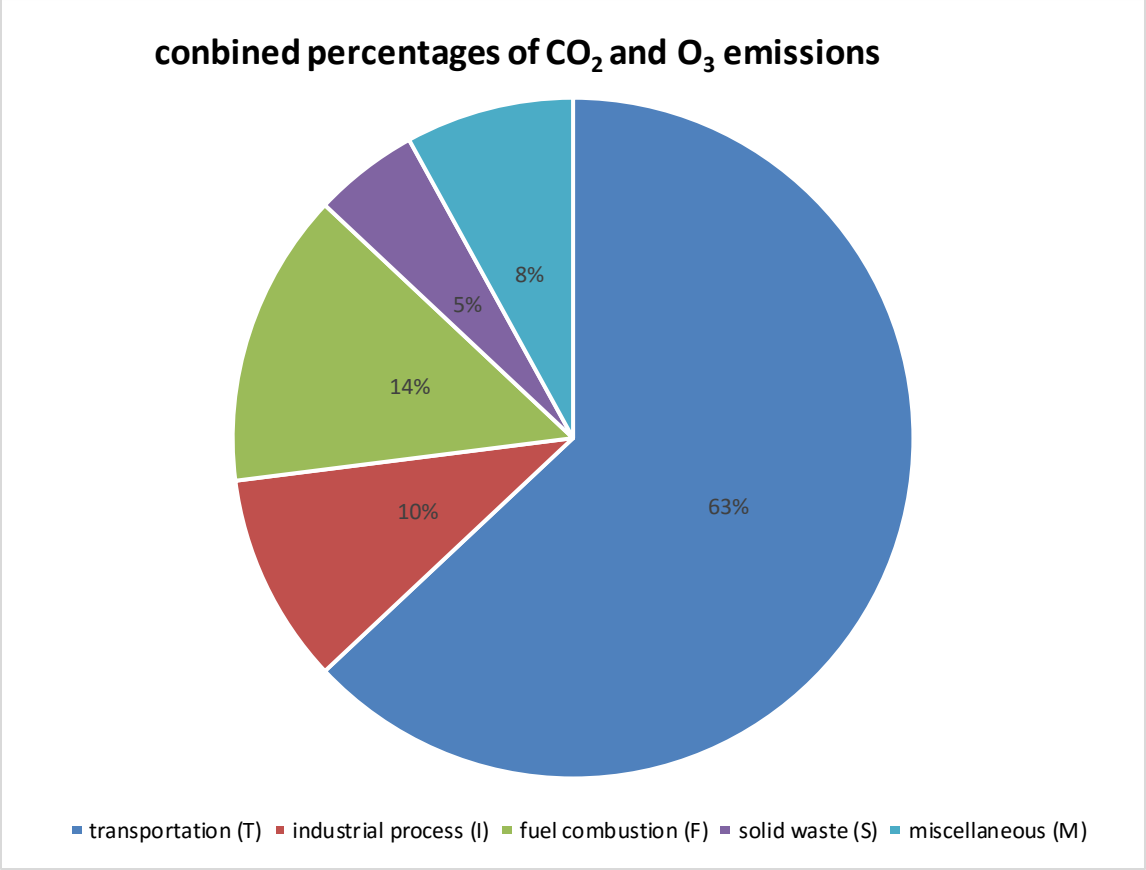
It is a circle divided into sectors that represent the percentages of population or a simple that Belongs to different categories.

Pie chart are especially useful for presenting Categorical data. The pie slices are drown such that they have an area proportional to the frequency.

Example:

The combined percentages of Carbon monoxide (CO_2) and ozone(O_3) emissions from different sources are listed in table as follow:

transportation (T)	industrial process (I)	fuel combustion (F)	solid waste (S)	miscellaneous (M)
63%	10%	14%	5%	8%



2.3- **Quantitative variable**:

a)- **discrete variable** :

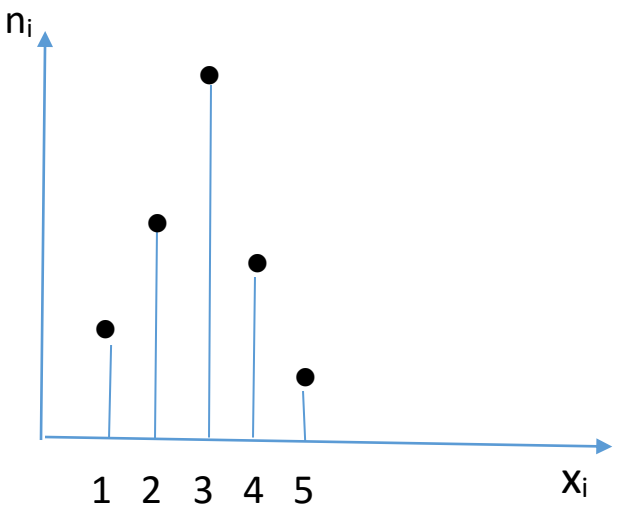
We can representing a frequency distribution of discrete variable by sticks (needle chart; rod graph).

Example:

The statistical distribution of the number of rooms in each dwelling in a city in SETIF is as follow:

x_j	1	2	3	4	5	total
n_j	4	10	16	7	3	40

-The Bar chart of the number of rooms in dwelling in a city in SETIF-



b)- continuous variable:

We use a histogram for representation a statistical distribution of continuous variable.

A histogram is a graph in which classes are marked on the horizontal axis and either the frequencies are represented by the heights on the vertical axis in a histogram, the bars are drawn adjacent to each other without any gaps.

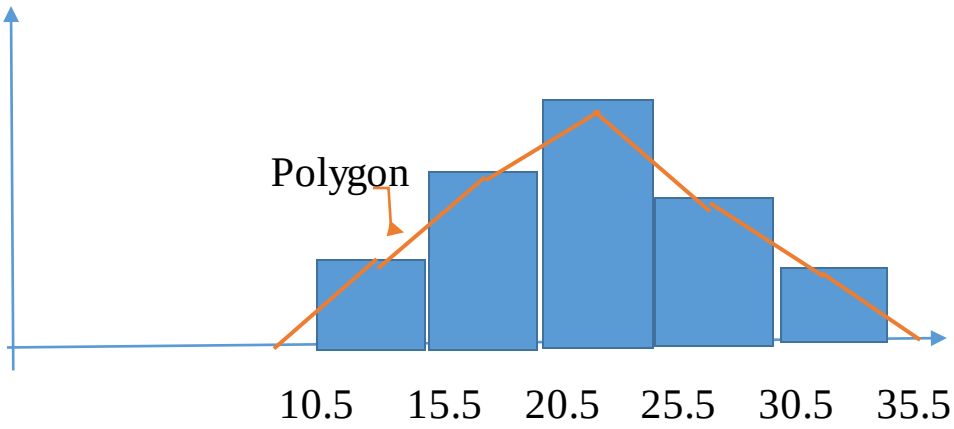
Example:

We consider the frequency distribution as follow:

class	frequencies n_j
[10.5 ; 15.5[	3
[15.5 ; 20.5[	6
[20.5 ; 25.5[	8
[25.5 ; 30.5[	5
[30.5 ; 35.5[	3
total	25

- The Histogram and polygon -

frequencies n_i



- **Note**

if the length of the classes are different, the frequencies should be adjusted for draw the histogram. The adjusted frequency is calculate using the following formula:

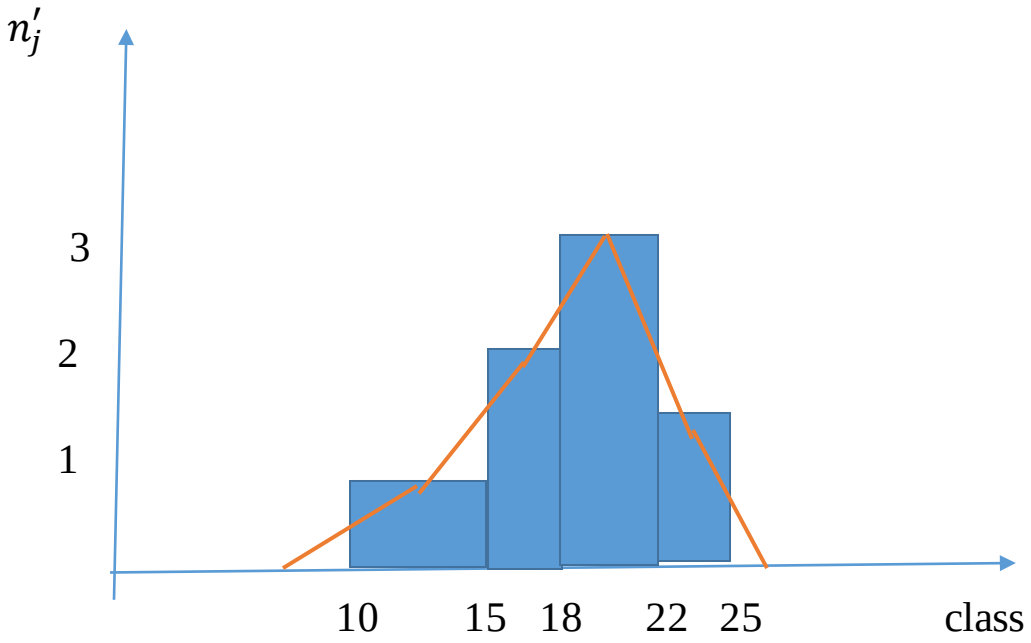
$$n'_j = \frac{n_j}{h_j}$$

Example:

We consider the frequency distribution as follow:

classes	frequency n_j	length of class h_j	adjusted frequency n'_j
$[\,10-15[$	4	5	4/5= 0.8
$[\,15-18[$	6	3	6/3= 2
$[\,18-22[$	12	4	12/4= 3
$[\,22-25[$	5	3	5/3= 1.3
total	23		

- The histogram and Polygon -



2- The Cumulative Distribution Function (CDF):

The cumulative distribution function of « x » denoted by $F(x)$ gives the relative frequency of individuals with values less than or equal "x ", therefore:

$$F: IR \rightarrow [0\ ,1\ ]$$

$$x \mapsto F(x) = \frac{N(x)}{n}$$

Where:

- $F\ is\ an\ incresing\ function\ of\ x$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to +\infty} F(x) = 1$$

- **Discrete variable:**

In this case, the CDF is fixed in certain areas (intervals) as follow:

$$F(x) = \begin{cases} 0 & if\ \ x < x_1 \\ \frac{N(x_j)}{n} & if\ \ x_j \le x < x_{j+1}\ \ ;\ \ j = 1, \dots, (K-1) \\ 1 & if\ \ x \ge x_K \end{cases}$$

Example: The statistical distribution of the number of rooms in each dwelling in a city in SETIF is as follow:

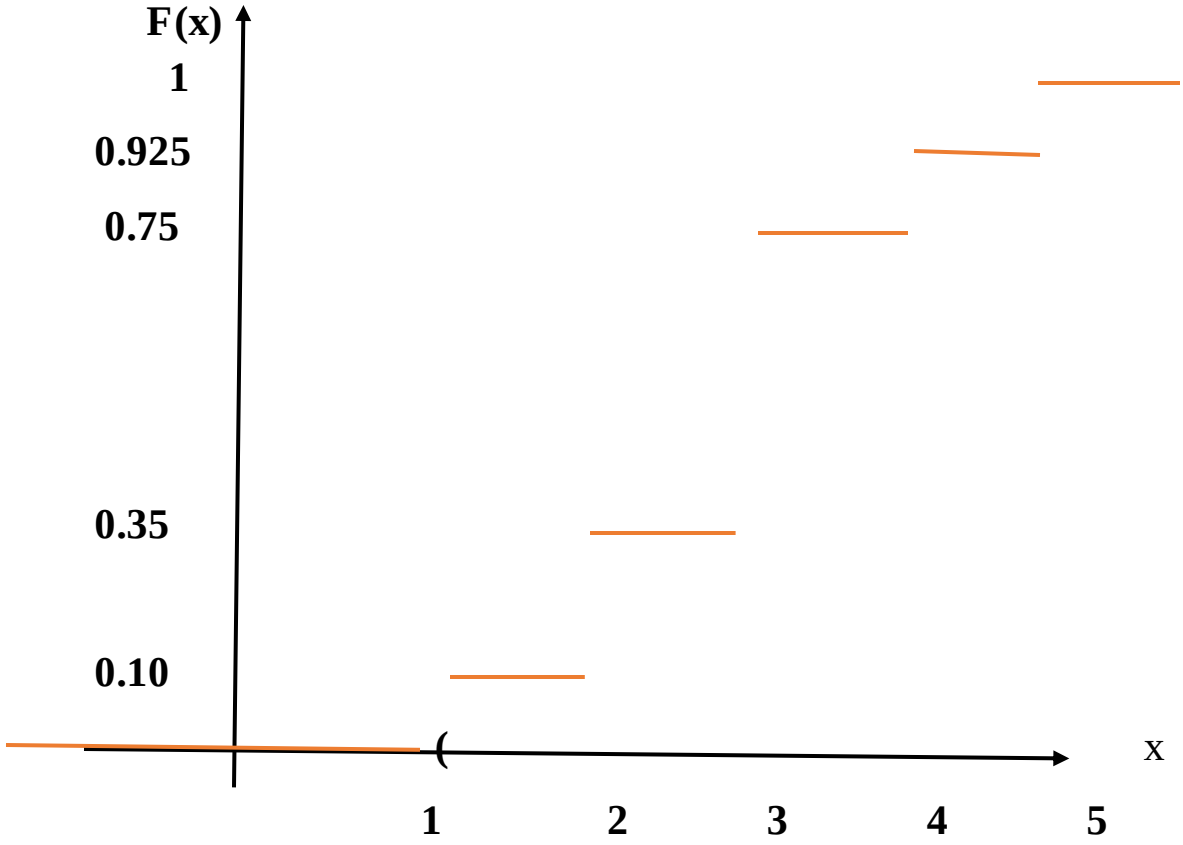
x_j	1	2	3	4	5	total
n_j	4	10	16	7	3	40

the cumulative distribution function (CDF).

x_j	1	2	3	4	5	total
n_j	4	10	16	7	3	40
N(x)	4	14	30	37	40	
F(x)	0.1	0.35	0.75	0.925	1	

$$F(x) = \begin{cases} 0 & if\ \ x < 1 \\ 0.10 & if\ 1 \le x < 2 \\ 0.35 & if\ \ 2 \le x < 3 \\ 0.75 & if\ \ 3 \le x < 4 \\ 0.925 & if\ 4 \le x < 5 \\ 1 & if\ \ x \ge 5 \end{cases}$$

- The CDF graph -



- **Continuous variable.**

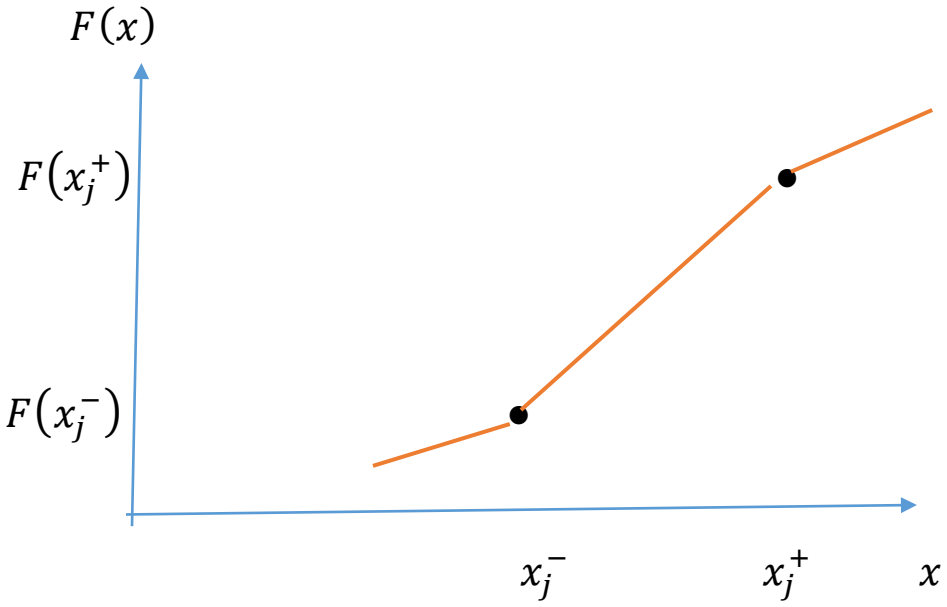
In this case, the CDF is continues and increasing in certain class (intervals).

We assume that the relationship between the cumulative frequency F(x) and the variable (x) is linear.

In each class « j » we assume that the relationship as follow:

$$F(x) = a_j\, x + b_j \quad ,$$

$$where:\ j = 1, \dots, K$$



If we use one of the methods of calculating the rectal equation, we find:

- $a_j = \frac{f_j}{h_j}$, where: f_j represent the relative frequency of class "j".

  h_j Represent the length of class "j".

- $b_j = F(x_j^-) - \frac{f_j}{h_j}\ x_j^-$,

  we note, $F(x_1^-) = 0$ and $F(x_{j+1}^-) = F(x_j^+)$

Finally, we write the CDF as follow;

$$F(x) = \begin{cases} 0 & if\ \ x < x_1^- \\ a_j x + b_j & if\ \ x_j^- \le x < x_j^+\ \ ;\ \ j = 1 \dots K \\ 1 & if\ \ x \ge x_K^+ \end{cases}$$

Example: We give the frequency distribution as follow:

classes $[x_j^- \ ; \ x_j^+[$	frequencies n_j	$N(x_j^+)$	$F(x_j^+)$	relative frequencies (f_j)
$[10.5 \ ; 15.5[$	3	3	0.12	0.12
$[15.5 \ ; 20.5[$	6	9	0.36	0.24
$[20.5 \ ; 25.5[$	8	17	0.68	0.32
$[25.5 \ ; 30.5[$	5	22	0.88	0.20
$[30.5 \ ; 35.5[$	3	25	1	0.12
total	25			1

The Cumulative Distribution Function:

- the class 1, (j=1):

we have $F(x) = a_1 \, x + b_1$ where :

$$a_1 = \frac{f_1}{h_1} = \frac{0.12}{5} = 0.024$$

$$b_1 = F(x_1^-) - \frac{f_1}{h_1} \ x_1^- = 0 - 0.024 * 10.5 = -0.252$$

in this class the cdf is : $F(x) = 0.024 \, x - 0.252$

- the class 2, (j=2):

we have $F(x) = a_2 \, x + b_2$ where :

$$a_2 = \frac{f_2}{h_2} = \frac{0.24}{5} = 0.048$$

$$b_2 = F(x_2^-) - \frac{f_2}{h_2} \ x_2^- = 0.12 - 0.048 * 15.5 = -0.624$$

in this class the cdf is : $F(x) = 0.048 \, x - 0.624$

- the class 3 ,(j=3):

we have $F(x) = a_3\, x + b_3$ where :

$$a_3 = \frac{f_3}{h_3} = \frac{0.32}{5} = 0.064$$

$$b_3 = F(x_3^-) - a_3 \;\; x_3^- = 0.36 - 0.064 * 20.5 = -0.952$$

in this class the cdf is : $F(x) = 0.064\, x - 0.952$

- the class 4 ,(j=4):

we have $F(x) = a_4\, x + b_4$ where :

$$a_4 = \frac{f_4}{h_4} = \frac{0.20}{5} = 0.04$$

$$b_4 = F(x_4^-) - a_4 \;\; x_4^- = 0.68 - 0.04 * 25.5 = -0.34$$

in this class the cdf is : $F(x) = 0.04\, x - 0.34$

- the class 5 ,(j=5):

we have $F(x) = a_5\, x + b_5$ where :

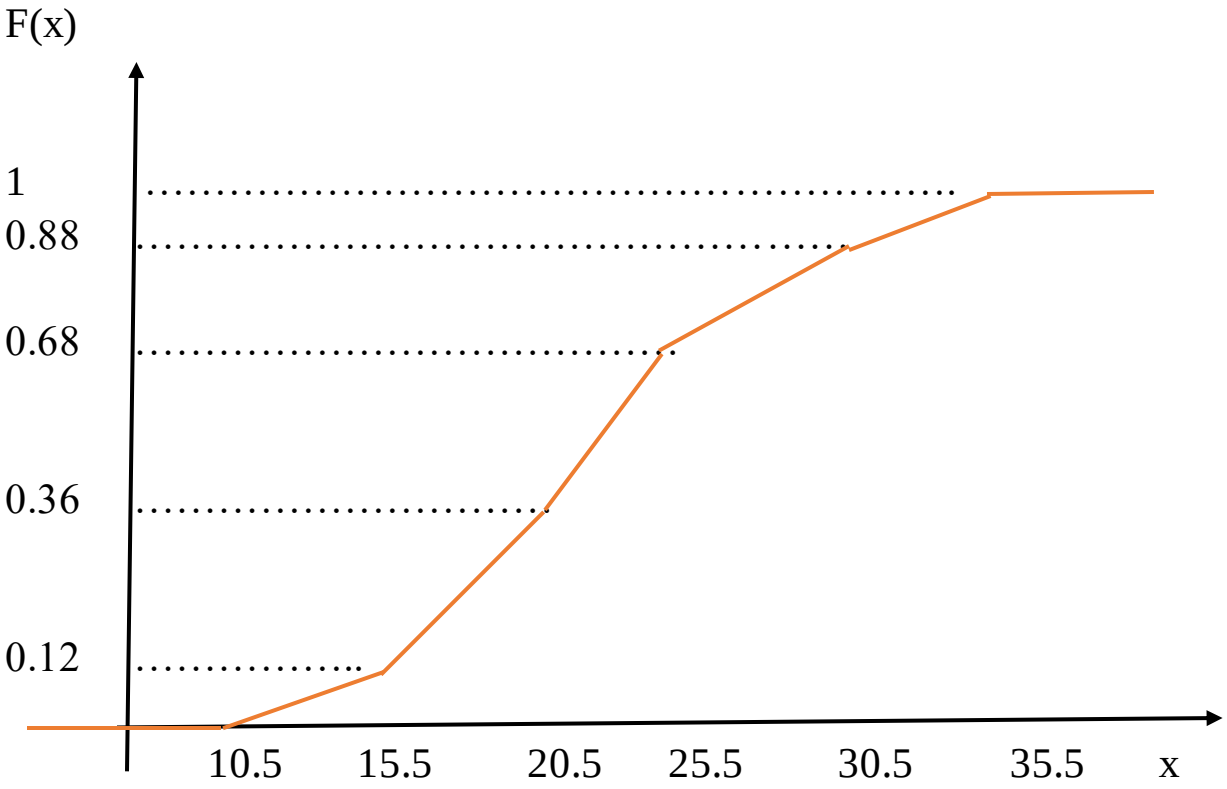
$$a_5 = \frac{f_5}{h_5} = \frac{0.12}{5} = 0.024$$

$$b_5 = F(x_5^-) - a_5 \;\; x_5^- = 0.88 - 0.024 * 30.5 = 0.148$$

In this class the CDF is : $F(x) = 0.024\, x + 0.148$
Then we write the CDF as follow:

$$F(x) = \begin{cases} 0 & if\ x < 10.5 \\ 0.024\,x - 0.252 & if\ x \in [10.5 - 15.5[\\ 0.048\,x - 0.624 & if\ x \in [15.5 - 20.5[\\ 0.064\,x - 0.952 & if\ x \in [20.5 - 25.5[\\ 0.040\,x - 0.340 & if\ x \in [25.5 - 30.5[\\ 0.024\,x + 0.148 & if\ x \in [30.5 - 35.5[\\ 1 & if\ x \geq 35.5 \end{cases}$$

- the graphical representation of CDF:



Exercises.

Exercise 1:

Let the distribution of 100 students according to the number of brothers, is given by the following table:

number of brothers	1	2	3	4	5	Total
number of students	12	18	34	26	10	100

1- Give a graphical representation of this distribution.
2- Calculate the ascending cumulative frequency (ACF) , represent it graphically.
3- Calculate the descending cumulative frequency (DCF), represent it graphically.

Solution:

1- **the graphical representation** :
in this case we use a needle chart, because the number of brothers is a **discret continues** variable.

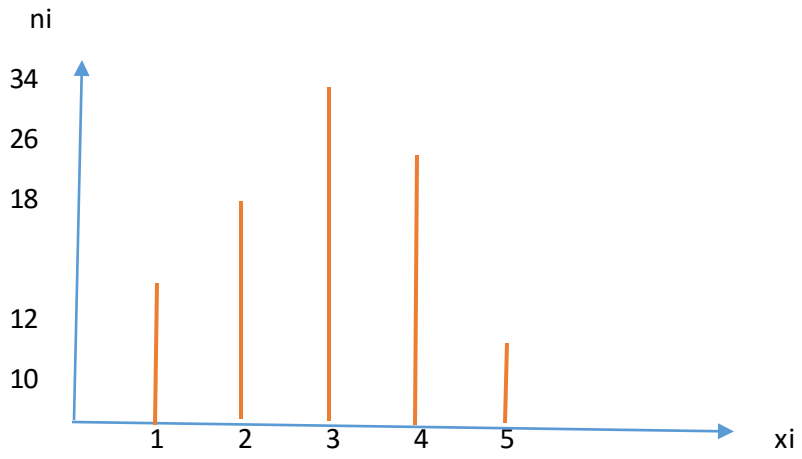


Figure: Needle chart of the distribution of students according to the number of brothers.

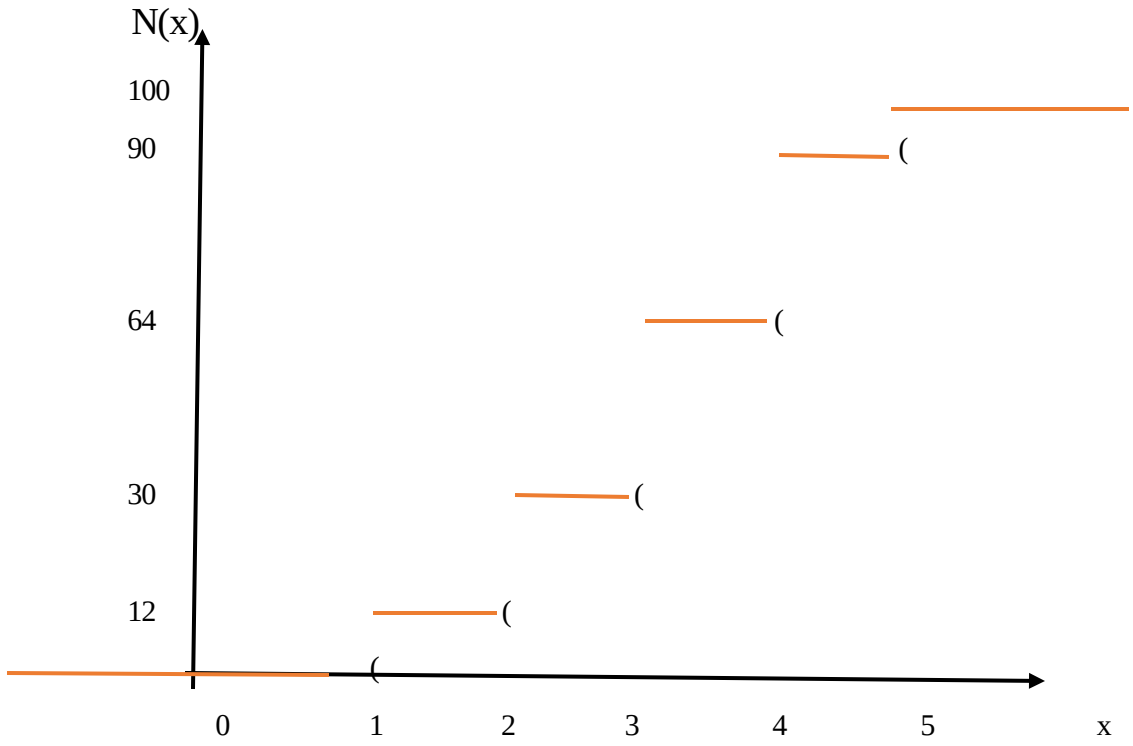
2- Calculate the ACF :

number of brothers (xi)	1	2	3	4	5	Total
number of students (ni)	12	18	34	26	10	100
ACF or N(x)	12	30	64	90	100	

Then the N(x) function is:

$$N(x) = \begin{cases} 0 & if \quad x \in]-\infty\,;1[\\ 12 & if \quad x \in [1\,;2[\\ 30 & if \quad x \in [2\,;3[\\ 64 & if \quad x \in [3\,;4[\\ 90 & if \quad x \in [4\,;5[\\ 100 & if \quad x \in [5\,;+\infty[\end{cases}$$

- The graphical representation of ACF.



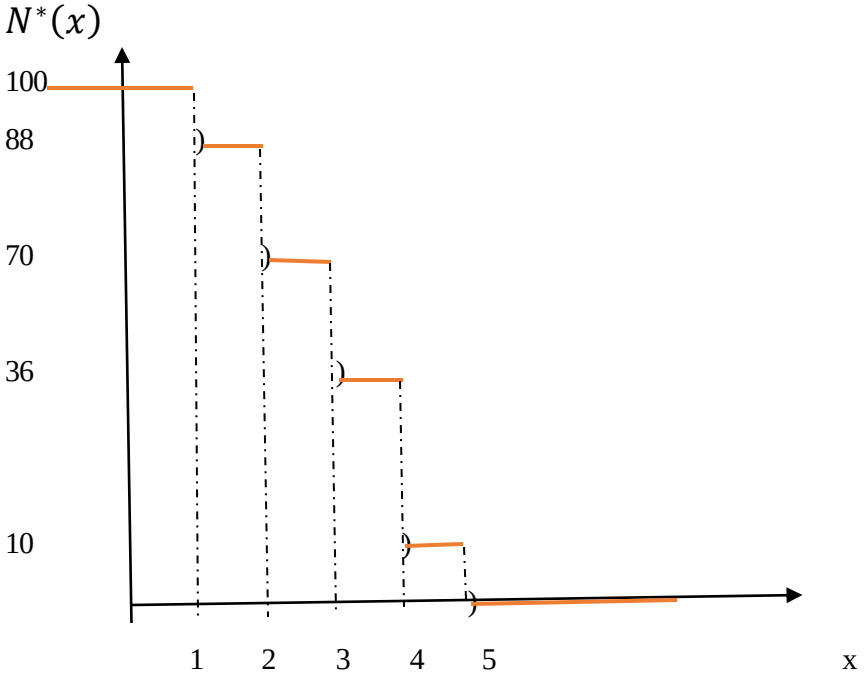
3- The Descending Cumulative Function (DCF)

number of brothers (xi)	1	2	3	4	5	Total
number of students (ni)	12	18	34	26	10	100
DCF or $N^*(x)$	100	88	70	36	10	

Then the DCF function is:

$$N^*(x) = \begin{cases} 100 & if \quad x \in]-\infty ; 1] \\ 88 & if \quad x \in]1 ; 2] \\ 70 & if \quad x \in]2 ; 3] \\ 36 & if \quad x \in]3 ; 4] \\ 10 & if \quad x \in]4 ; 5] \\ 0 & if \quad x \in]5 ; +\infty[\end{cases}$$

- The graphical representation of DCF



Exercises

Exercise 1:

Define the population, the statistical unit (individual), and the variable in the following cases:

1. The life span of televisions produced at Al-Salam Electronics Factory in Al- Burj in 2015
2. Monthly wages for teachers at the National Polytechnic School in 2016
3. The color of the wallet for fourth-year primary school students at Al-Bashir Al-Ibrahimi School in the capital
4. The number of children per family in a neighborhood in the city of Annaba.
5. Number of employees in the organization by job.
6. Ranking of parties according to the number of votes obtained in the elections.

Exercise 2:

Are the following statistical populations precisely defined for the purpose of conducting statistical studies?

1. Students of the Higher School of Commerce.
2. Persons of Algerian nationality
3. Algerian unemployment on April 8, 2012 .
4. Transport sector institutions in Algeria
5. Residents of the Algiers region .
6. Highly paid workers of the SETIF unit.

Exercise 3:

A factory produces 2000 computers annually and employs 500 workers, including 10 engineers, 75 senior technicians, 150 administrators, and 20 specialized machines:

1. Identify statistical population and their appropriate statistical units that can be studied
2. Give examples of characteristics that can be attributed to statistical units.

Exercise 4:

Let the lengths of 4 objects be: 370 - 0.5 - 1.052 - 4.1895

1. Determine the measurement accuracy used to measure each of the five lengths

2. Determine the field of real numbers to which each length can belong.

Exercise 5:

Let the following questions be from a survey of adult residents in the municipality of Muftah in March 2008 about pollution.

- Gender : Female , Male
- Educational level : primary , middle , secondary , university , otherwise
- Age: 18 – 30 ; 30 – 50 ; 50 years and above

1.Define the studied statistical population.

2. What is the feature properties of statistical units?

3. Determine the variables studied, and their names?

4. What the type of each variable in each case and its measurement scale?

Exercise 6:

The statistical investigation carried out by the National Office of Statistics on the costs of living for families in the city of Algiers in all its social occupational categories in the year 2000, using a sample size of 750 families, showed that the monthly expenditure of families living in the city of Algiers is distributed on average among the spending groups in percentages as follows:

Spending groups	(%)
Nutrition and non-alcoholic beverages	43.09
Clothes and shoes	7.45
Housing and its maintenance	9.29

Home Furnishing	4.96
Health and prevention	6.2
Transport and communication	15.85
Education and culture	4.52
Other spending	8.64
Total	100

1. Define the studied characteristic, the variable adopted to measure it?

2. What is the type of variable and its measurement scale?

3. Represent graphically this statistical distribution.

Exercise 7:

In order to determine the extent of students' keenness to follow lessons in the classrooms, the Department of Studies at the Higher School of Commerce followed up students' absence from lessons by surveying a sample of 25 first-year preparatory students during the month of November and recorded the following results:

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

1. Determine the studied characteristic, the variable adopted to measure it.

2. Define the type of variable and its measurement scale.

3. Represent the initial data in the table.
4. Represent graphically this distribution.
6. Determine the cumulative distribution function and Represent it graphically.

Exercise 8:

To determine the level of economic activity in Algeria, the Directorate of Economic Accounts at the National Office of Statistics calculates a macroeconomic indicator called the Gross Domestic Product (GDP),measures the level of economic production achieved within the country's economic region by economic agents residing in this region during a specific period of time, usually one year.

The following table shows the evolution of the level of the value of the GDP

Year	2011	2012	2013	2014
GDP (billiar dinars)	14588.5	16208.7	16643.8	17205.1

Source: ONS

The following table shows the distribution of (GDP) achieved in 2014 according to sectors of activity

Sectors of activity	**value added**
Agriculture, forestry and fishing	1771,5
Fuel	4657.8
Industry	837.0
Construction and public works	1794.0
Services	6906.4
Value added tax and customs duties	1238.4
Total	17205.1

1. What is the statistical population studied?
2. What is the studied property?
3. What is the variable adopted to measure this characteristic?
4. What type of statistical series is in the first table?
5. Represent graphically this series.
6. What is the type of statistical series in the case of the second table? Represent it graphically.
7. Explain the difference between the two series.

Exercise 9:

The Secretary of the Office of the Labor Medicine Service at the level of an institution recorded, within 50 days, the number of individuals who applied to the Service for medical reasons and the results were as follows:

5, 4 ,3,2, 4 , 4 , 3, 4, 3, 2, 4, 5, 2 ,1, 4, 0, 4 ,0 ,2 , 3, 2, 4 ,2 , 3, 4, 2, 3,1, 4, 4, 4, 1, 4 ,2, 2, 4, 5, 4 ,3, 4, 4 ,3, 5 ,4, 4, 3, 0, 4, 3, 4.

1. Determine the type of this series.
2. Represent these data in the form of an ordered statistical series
3. Represent these data in the form of an aggregated statistical distribution.
4. Represent graphically this distribution.
5. Calculate the descending cumulative frequency and represent it graphically.
6. Find the cumulative distribution function and represent it graphically.

Exercise 10:

Define the type of statistical variable and its measurement scale in each following cases:

1. The person's age.
2. The person's family status.
3. The strength of the voice.
4. The social and professional status
5. The temperature in the hall
6. 7- The individual's intelligence
7. The price of bread
8. The mother tongue.

Exercise 11: The Employment Service of an organization with 800 employees conducted a statistical survey of 30 workers regarding the time it takes them to reach the workplace and obtained the following results (in minutes):

40 , 45 , 40 , 15 , 25 , 20 , 5 , 30 , 34 , 50 , 40 , 10 , 26 , 35 , 15 , 20 , 15 , 20 , 25 , 31 , 30 , 35 , 20 , 35 , 30 , 12 , 25 , 20 , 10 , 2 .

1- Display this raw series as a pooled distribution with 6 classes.

2- Calculate the relative frequency and percent relative frequency.

3- Represent graphically this distribution and show the area of the frequency polygon.

4- Calculate the ascending cumulative frequency and the descending cumulative frequency.

5- Find the cumulative distribution function and represent it graphically.

Exercise 12:

In order to form an idea about the standard of living of the population of a city, the Bureau of Statistical Studies conducted a statistical investigation that included 100 families, and data on monthly consumption spending was collected and presented in the form of an aggregate statistical distribution as follows:

Expenditure in thousand DZD Number of families	
20 - 40	15
40 - 60	20
60 - 100	20
100 - 200	45

1. Define the studied property, the statistical variable and this measurement scale.

2. Represent graphically this distribution.

3. Find the cumulative distribution function and represent it graphically.

Chapter 3 : Measures of Central Tendency (or Location).

Chapter 3 : Measures of Central Tendency (or Location).

The Sample data can be summarized, based on a number of descriptive statistics. The measurements of central tendency calculated from sample data are called statistics. If the statistics are calculated for an entire population, they are called parameters rather than statistics.

In this chapter, we discuss the most important statistics for location as arithmetic mean, median, mode; the number of statistics that can be calculated depends on the nature of the data. If we want to talk about the location of nominal data, the numerical information is limited to the frequencies (the greater frequency) .For ordinal data, we can take into account the order in the data. Therefore, it can make sense to speak, for example, about the middle element of a sample. For variables measured on an interval scale or ratio scale, the arithmetic mean plays an important role.

The focus in this chapter is mainly on unvaried descriptive statistics for quantitative variables.

1- **Arithmetic mean** :

The arithmetic mean is undoubtedly the best-known measure of central location; it is also called the sample mean

- the arithmetic measure $\bar{X}$ of observations $x_1 , x_2 , \dots\dots , x_n$ is defined by :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

If we have observations $x_1 , x_2 , \dots\dots , x_J$, with respective frequencies $n_1 , n_2 , \dots\dots , n_J$ (grouped data), then:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j \quad \text{or} \quad \bar{X} = \sum_{j=1}^{J} f_j x_j$$

Where x_j is the midpoint of the j^{-th} class, n_j is the absolute frequency (f_j is the relative frequency) of the j^{-th} class, "J" is the number classes and "n" is number of observations (sample size).

example1: suppose that a sample consists of the following 10 observations: 6 , 3 , 4 , 7 , 4 , 6 , 7 , 6 , 5 , 3.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(6+3+4+7+4+6+7+6+5+3) = 5.1$$

or :

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{1}{10}(2*3+2*4+1*5+3*6+3*7) = 5.1$$

- **example2:** we assume the statistical distribution as follow :

classes	frequency n_j	midpoint of class x_j	$n_j\ x_j$
[10.5 ; 15.5[	3	13	39
[15.5 ; 20.5[	6	18	108
[20.5 ; 25.5[	8	23	184
[25.5 ; 30.5[	5	28	140
[30.5 ; 35.5[	3	33	99
total	25		570

The arithmetic mean

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{570}{25} = 22.8$$

- **Properties of the mean:**
  - The mean is affected by a few extremely large or extremely small observations (outliers).
  - The sum of all observations is equal to the arithmetic mean multiplied by the simple size :

$$\sum_{j=1}^{J} n_j x_j = n\bar{X}$$

  - The sum of the deviations of the observations from the mean equals zero :

$$\sum_{j=1}^{J} n_j (x_j - \bar{X}) = 0$$

Proof: we have

$$\sum_{j=1}^{J} n_j(x_j - \bar{X}) = \sum_{j=1}^{J} n_j x_j - \sum_{j=1}^{J} n_j \bar{X}$$

$$= \sum_{j=1}^{J} n_j x_j - \bar{X} \sum_{j=1}^{J} n_j = n\bar{X} - \bar{X}n = 0$$

- The arithmetic mean of a sample of constant values "a" , equals the constant value itself: $\bar{X} = a$
- If $y_i = ax_i + b$ (linear transformation), then :

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

Proof: we have;

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^{J} n_j y_j = \frac{1}{n} \sum_{j=1}^{J} n_j \left(ax_j + b\right)$$

$$\bar{Y} = a\frac{1}{n} \sum_{j=1}^{J} n_j x_j + \frac{1}{n} \sum_{j=1}^{J} n_j b$$

$$\bar{Y} = a\bar{X} + \frac{1}{n} nb$$

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

2- Geometric mean:

In certain contexts, the geometric mean makes more sense than the arithmetic mean.

The geometric mean G of a set of observations $x_1 , x_2 , \ldots\ldots , x_n$ is :

$$\mathbf{G} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

For the set of observations $x_1 , x_2 , \ldots\ldots , x_J$, with respective frequencies $n_1 , n_2 , \ldots\ldots , n_J$ (grouped data), then:

$$\mathbf{G} = \sqrt[n]{\prod_{j=1}^{J} x_j^{n_j}} = \left(\prod_{j=1}^{J} x_j^{n_j}\right)^{\frac{1}{n}}$$

so:

$$\log(\mathbf{G}) = \frac{1}{n} \sum_{j=1}^{J} n_j \log\left(x_j\right) = \sum_{j=1}^{J} f_j \log\left(x_j\right)$$

Therefore, we conclude, the logarithm of geometric mean is the arithmetic mean of logarithms of observations.

This definition implies that the geometric mean can only be calculated for strictly positive observations. The geometric mean is always smaller or equal to the arithmetic mean ($\mathbf{G \leq \bar{X}}$).

Example 1:

For the following sample 1 , 3 , 5 , 6 , and 8, we obtain:

- **The arithmetic mean:**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}(1+3+5+6+8) = 4.6$$

- **The geometric mean:**

$$\mathbf{G} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[5]{1*3*5*6*8} = 3.73$$

We remark that;

$$\mathbf{G < \bar{X}}$$

Example 2:

We assume a financial institution that provides interest on capital for five successive years as follows: 3%, 5%, 5%, 4%, and 6%.

-Calculate the average annual return rate (interest rate) for this period.

Solution:

We assume that an investor who deposited a sum of S_0 in her account on 1st year, ended up with a total return after 5 years as follow:

- after the 1st year :
  $S_1=S_0+S_0*r_1= S_0(1+r_1) = S_0(1+0.03)$
- after the 2ndyear :
  $S_2=S_1+S_1*r_2= S_1(1+r_2) = S_0(1+r_1)(1+r_2) = S_0(1.03)(1.05)$
- after the 3rdyear:
  $S_3=S_0(1+r_1)(1+r_2)(1+r_3) = S_0(1.03)(1.05)(1.05)$
- after the 4thyear:
  $S_4=S_0(1+r_1)(1+r_2)(1+r_3)(1+r_4) = S_0(1.03)(1.05)(1.05)(1.04)$
- after the 5thyear:
  $S_5=S_0(1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5) = S_0(1.03)(1.05)(1.05)(1.04)(1.06)$

If the average interest rate equals “r”, then $S_5=S_0(1+r)^5$

Moreover, we have:

$S_5=S_0(1+r_1)\ (1+r_2)\ (1+r_3)\ (1+r_4)\ (1+r_5)$

So

$S_0(1+r)^5 = S_0(1+r_1)\ (1+r_2)\ (1+r_3)\ (1+r_4)\ (1+r_5)$

Then:

$(1+r)^5 = (1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)$

Alternatively:

$$r = \sqrt[5]{(1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)} - 1$$

$$r = \sqrt[5]{(1.03)(1.05)(1.05)(1.04)(1.06)} - 1$$

$r = 0.046$; The average interest rate equals 4.6%

Therefore, the mean (the average) of a set of rates $\{r_1, r_2, \ldots., r_n\}$ is defined by:

$$r = \sqrt[n]{(1+r_1)(1+r_2)\ldots\ldots(1+r_n)} - 1$$

In addition, the average growth rate of the value S_0 to become equal to the value S_n during period 'n', is given by the following relation:

$$r = \sqrt[n]{\frac{S_n}{S_0}} - 1$$

3- HARMONIC MEAN.

The harmonic mean "H "of a set of data $x_1, x_2, \ldots\ldots., x_n$ is the reciprocal of the arithmetic mean of the reciprocals of the data:

$$\frac{1}{H} = \frac{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}{n}$$

Or:

$$H = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

For the set of observations $x_1, x_2, \ldots\ldots, x_J$, with respective frequencies $n_1, n_2, \ldots\ldots, n_J$ (grouped data), then:

$$H = \frac{n}{\sum_{ij=1}^{J}\left(\frac{n_j}{x_j}\right)}$$

$or;$

$$H = \frac{1}{\sum_{ij=1}^{J}\left(\frac{f_j}{x_j}\right)}$$

With, $x_j \neq 0$, x_j: $midle\ point\ of\ class$ " j", n_j: frequency of class j .

We also note that:

$$\boldsymbol{H \leq G}$$

Example 1: the harmonic mean for the sample, 1 ,3 , 5 , 6 , 8 , is:

$$H = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)} = \frac{5}{\frac{1}{1}+\frac{1}{3}+\frac{1}{5}+\frac{1}{6}+\frac{1}{8}} = \frac{5}{1.825} = 2.74$$

Example 2: calculate the harmonic mean for the statistical distribution as follow:

classes	4 - 6	6 – 10	10 - 14	14 – 18	total
x_j	5	8	12	16	
n_j	3	5	4	2	14
$\frac{n_j}{x_j}$	0.6	0.625	0.333	0.125	1.683

$$H = \frac{n}{\sum_{j=1}^{K}\left(\frac{n_j}{x_j}\right)} = \frac{14}{1.683} = 8.31$$

Example 3:

A person traveled a distance of 200 km, such that he covered the first 110 km at a speed of 120 km/h and covered the next 90 km at a speed of 70 km/h. Calculate the average speed at which this person traveled.

Solution:

We denote the average speed by "S", the total distance by "D" , the first distance by "D_1"and the second distance by "D_2", such as $D=D_1 + D_2$.

We denote the first speed by "S_1" in time t_1 , the second speed by "S_2" in time t_2 .

Such as the total time of this trip is $T=T_1+T_2$

$D_1=110$: $S_1=120$; t_1 $D_2=90$; $S_2=70$; t2

$D=200$; $T=t_1+t_2$; S=?

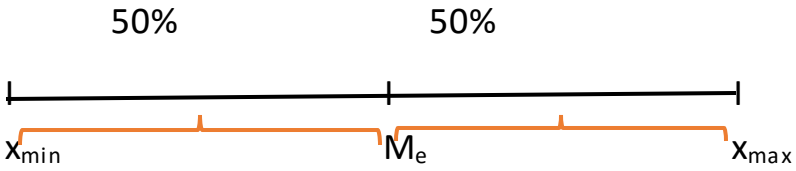
The average speed = the covered distance / the total time

$$S = \frac{D}{T} = \frac{D_1 + D_2}{t_1 + t_2} = \frac{D_1 + D_2}{\frac{D_1}{S_1} + \frac{D_2}{S_2}}$$

$$S = \frac{D_1 + D_2}{\frac{D_1}{S_1} + \frac{D_2}{S_2}} = \frac{110 + 90}{\frac{110}{120} + \frac{90}{70}} = \frac{200}{2.2024} = 90.81\ km/h$$

3- MEDIAN:

For a set of data $x_1, x_2, \ldots\ldots, x_n$ organized into an array, the median of the data "M_e" is the value that divides the array into two equal parts; there are as many data values below the median as above it.



The "odd-even rules" can be used to find the median of such an array. We have two cases:

- If there is an **odd number** of values in an array, then the median is the middle value of the array :

$$M_e = x_{\left(\frac{n+1}{2}\right)}$$

Example: Let the following data: 9,1,2,4,8,5,6.

- Classification in order of values: 1,2,4,5,6,8,9 .
- we have n=7 is an odd number, then
-

$$M_e = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{7+1}{2}\right)} = x_{(4)} = 5$$

- If the number of elements (n) is **even**, the median equal to the arithmetic mean of two middles values of the array:

$$M_e = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

Example: Suppose that a sample consists of the following 10 observations, 6, 3 ,4,7,4,6,7,6,5,3.

We have that n=10 is an even number;

classification in order of values :3,3,4,4,5,6,6,6,7,7,.

- the median $\mathrm{M_e} = \frac{1}{2}\left(\mathrm{x}_{\left(\frac{n}{2}\right)} + \mathrm{x}_{\left(\frac{n}{2}+1\right)}\right) = \frac{1}{2}\left(\mathrm{x}_{\left(\frac{10}{2}\right)} + \mathrm{x}_{\left(\frac{10}{2}+1\right)}\right)$

$$\mathrm{M_e} = \frac{1}{2}\left(\mathrm{x}_{(5)} + \mathrm{x}_{(6)}\right) = \frac{1}{2}(5+6) = 5.5$$

- <u>case 3</u>: For the grouped statistical distribution with classes,
  - We determine the median class :

  The median class (class number "J ") is the class corresponding to the ascending cumulative frequency that is equal to or greater than n/2 (or, $\mathrm{J} = \inf\left\{\mathrm{j} : \mathrm{F}\left(\mathrm{x}_{\mathrm{j}}^{+}\right) \geq 0.5\right\}$).
  - we calculate the median by applying the cumulative distribution function F(x) as follow :

    We know that the "CDF" in median class (class 'J')

$$\mathrm{F(x)} = \mathrm{a_J x} + \mathrm{b_J}$$

$$a_J = \frac{f_J}{h_J} \quad ; \quad b_J = F(x_J^-) - \frac{f_J}{h_J} x_J^-$$

then :

$$F(x) = \frac{f_J}{h_J} x + F(x_J^-) - \frac{f_J}{h_J} x_J^-$$

and we know : $F(M_e) = 0.5$

Then:

$$0.5 = \frac{f_J}{h_J} . M_e + F(x_J^-) - \frac{f_J}{h_J} x_J^-$$

Or:

$$M_e = x_J^- + \frac{0.5 - F(x_J^-)}{f_j} . h_J$$

where:

x_J^- : the lower boundary of median class.

$F(x_J^-)$: the ascending cumulative relative frequency of class befor the median class.

f_J : relative frequency of the median class.

h_j : lengh of the median class.

note; we can use the the ascending cumulative frequency $N(x)$ for calculate the median as follow:

$$M_e = x_J^- + \frac{\frac{n}{2} - N(x_{J-1})}{n_j} . h_J$$

x_J^- : the lower boundary of median class.

n: sample size (number of elements).

$N(x_{J-1})$: the ascending cumulative frequency of class befor the median class.

n_J : frequency of the median class.

h_j : lengh of the median class.

We can determine the median value using the graph of cumulative distribution function $F(x)$, where $(F(M_e) = 0.5)$, or by the graph of ascending cumulative frequency or the graph of descending cumulative frequency, where,

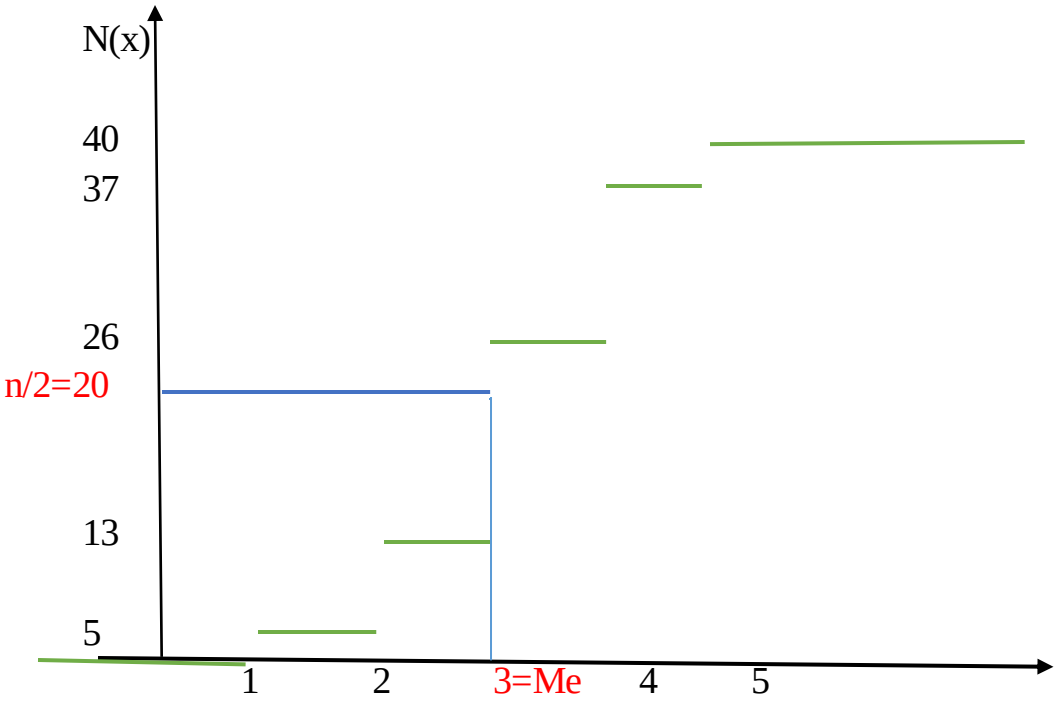
$$N(M_e) = N^*(M_e) = n/2$$

Example1:

We assume the statistical distribution of a discrete variable (X) as follow:

x_i	1	2	3	4	5	total
n_i	5	8	13	11	3	40
$N(x_i)$	5	13	26	37	40	

We draw the ascending cumulative frequency N(x).
We have n =40 (n is even number) and n/2 = 20 then Me=3 .



Example2: We assume the frequency distribution as follow:

classes	n_j	x_j	f_j	$F(x_j^+)$	ACF $N(x_j^+)$	DCF $N^*(x_j^-)$
[10.5 ;15.5[	3	13	0.12	0.12	3	25
[15.5 ;20.5[	6	18	0.24	0.36	9	22
[20.5 ;25.5[	8	23	0.32	0.68	17	16
[25.5 ;30.5[	5	28	0.20	0.88	22	8
[30.5 ;35.5[	3	33	0.12	1	25	3
total	25		1			

We have n/2 = 25/2=12.5
In addition, we have:

$number\ of\ medien\ cless\ J = \inf\left\{j : N\left(x_j^+\right) \geq \frac{n}{2} = 12.5\right\} = inf\{3,4,5\} =$ 3 Then J=3 (third class is the median class)

Then: $M_e \in [20.5\ ;\ 25.5]$

We have:

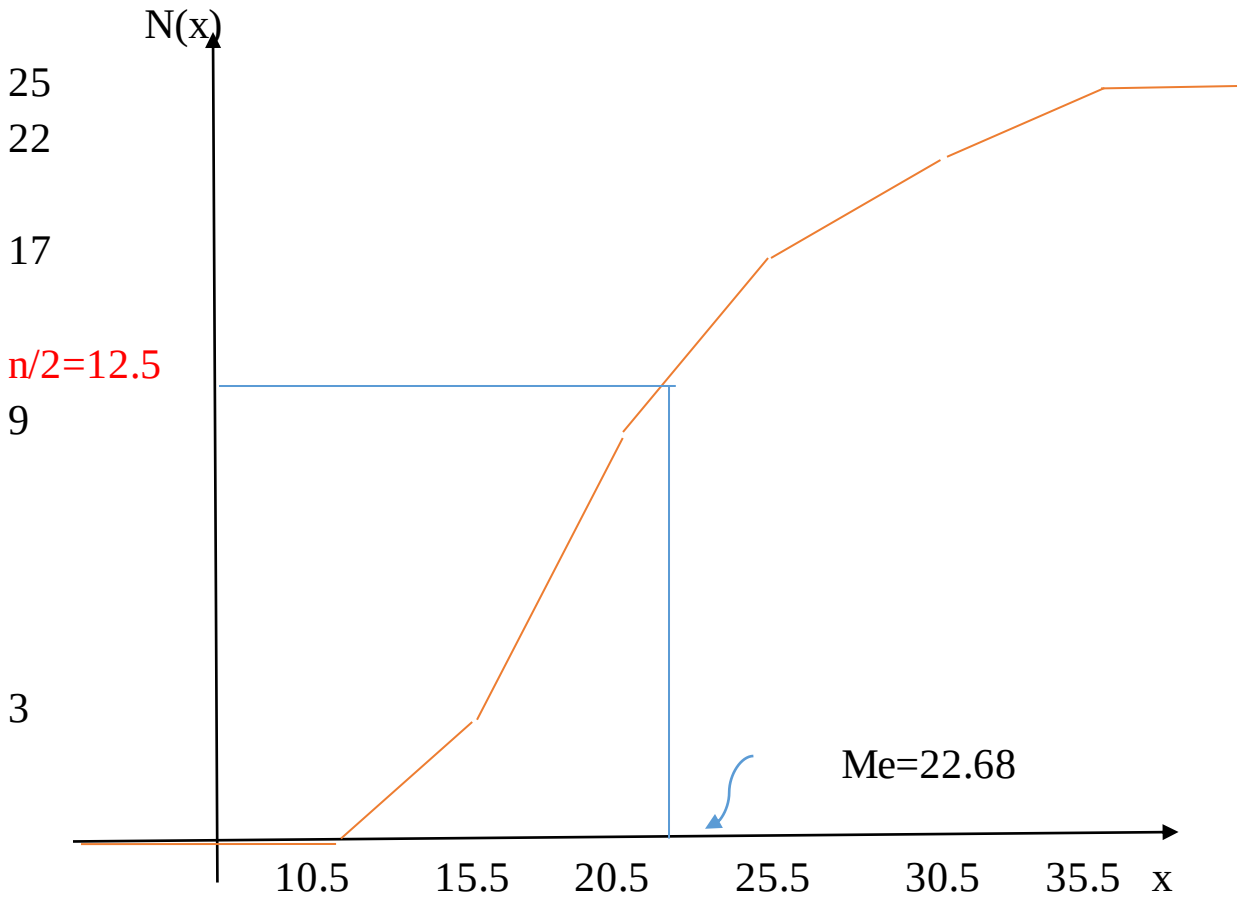
$$M_e = x_J^- + \frac{\frac{n}{2} - N(x_{J-1})}{n_j}.h_J = 20.5 + \frac{\frac{25}{2} - 9}{8}.5 = 22.6875$$

On the other hand:

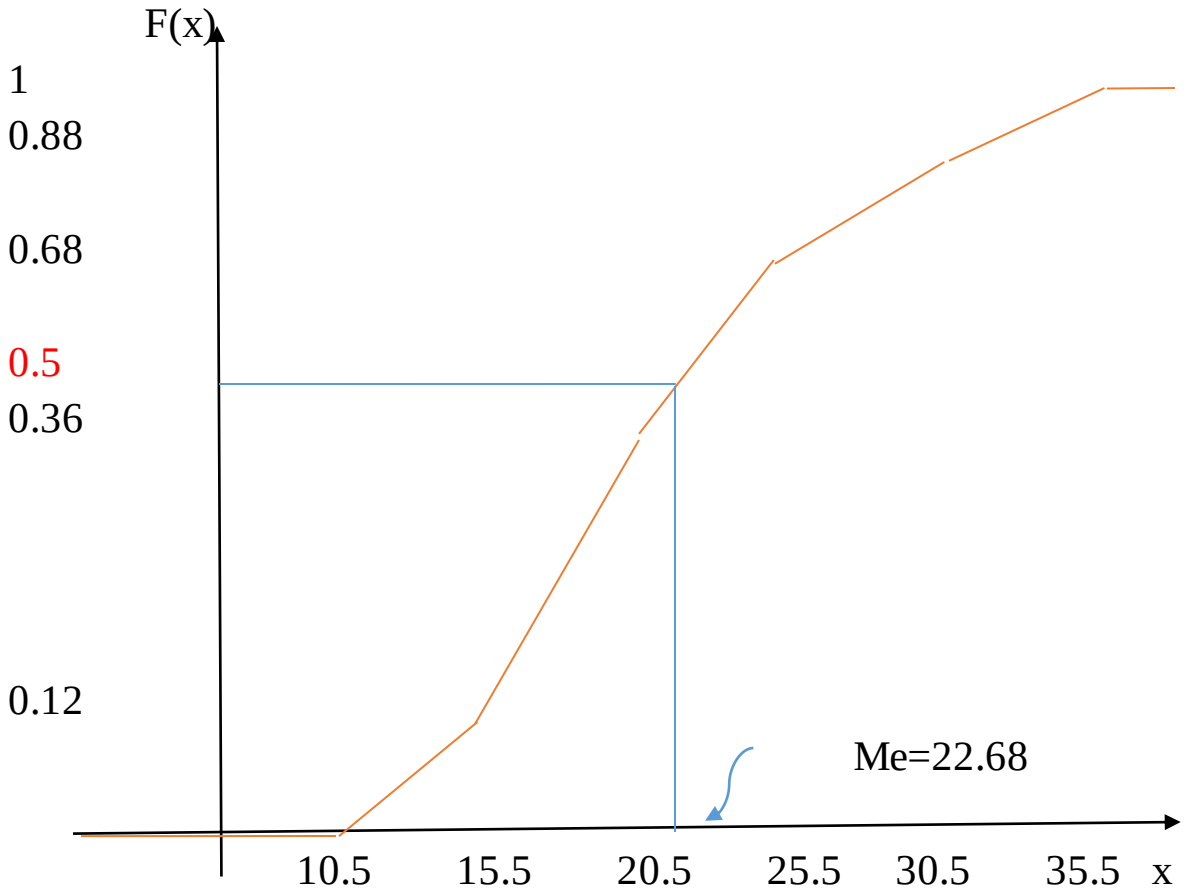
$$M_e = x_J^- + \frac{0.5 - F(x_{J-1})}{f_j}.h_J = 20.5 + \frac{0.5 - 0.36}{0.32}.5 = 22.6875$$

- **the median Graphically:**

  We draw the ascending cumulative frequency N(x):



or by using the cumulative distribution function F(x) we have:



- Some properties of the median :
  - About 50% of the observations are below (or above) the median.
  - The median is not affected by a few extremely large or extremely small observations.
  - the sum of the absolute deviations of the observations x_i from a constant " c" ($\sum_{i=1}^{n}|x_i - c|$) is minimal if $c = M_e$

4- **Percentile and Quartile**:

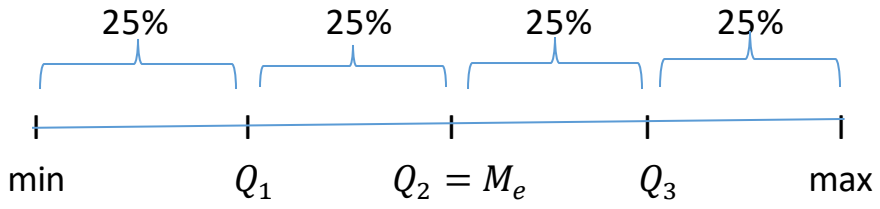
- **Percentile**

The p^{th} percentile (C_P) of a set of data, where 0<p<1 , is a real number that is greater than (about) 100*P% of the observations, and smaller than (about) 100*(1-p)% of the observations.

- **Quartiles.**

The first / second / third quartile (Q_1, Q_2, Q_3) is the 25 / 50 / 75th percentile of the simple in other words,

$$Q_1 = C_{0.25}, Q_2 = C_{0.50}, Q_3 = C_{0.75}$$



Q_1 : The lower quartile is the middle number the half of the data below the median.

Q_3 : The upper quartile is the middle number the half of the data above the median.

Example: for the data set representing the number of children in a random sample of 10 families in a neighborhood: 1,5,2,4,3,3,9,1,6,8.

- In the first, we classify the data set in ascending order. we obtain: 1, 1, 2, 3, 3, 4, 5, 6, 8, 9.

a) - the median:

We have n=10 (is an even number), then

$$M_e = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) = \frac{3+4}{2} = 3.5$$

b) - The first quartile (Q_1):

Q_1 is the median of the first half (50%) data, in this case is the median of following data : 1;1;2;3;3. we have n=5 (is an odd number) then $Q_1 = \left(x_{\left(\frac{n+1}{2}\right)}\right) = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 2$.

c)- the upper (3rd) quartile(Q_3):

Q_3 is the median of the 2nd half (50%) data, in this case is the median of following data : 4 ;5 ;6 ;8 ;9 . we have n=5 (is an odd number) then $Q_3 = \left(x_{\left(\frac{n+1}{2}\right)}\right) = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 6$.

Remark: In the statistical distribution grouped with class, we can use a relationship as the median, as follows:

- **lower quartile :**

$$Q_1 = x_J^- + \frac{\frac{n}{4} - N(x_{J-1})}{n_j}.h_J \quad , \qquad wher: J = \inf\{j : N(x_j^+) \geq n/4\}$$

- **upper quartile**:

$$Q_3 = x_J^- + \frac{\frac{3n}{4} - N(x_{J-1})}{n_j}.h_J \quad , \qquad wher: J = \inf\{j : N(x_j^+) \geq 3n/4\}$$

Example: if we take the previous example:

- we have n/4=25/4=6.25 then J=2 (the class of Q_1 is the class number 2 or the class $[15.5 \ ; \ 20.5[$

$$Q_1 = x_J^- + \frac{\frac{n}{4} - N(x_{J-1})}{n_j}.h_J = 15.5 + \frac{6.25 - 3}{6}.5 = 18.21$$

- we have 3n/4=3*25/4=75/4=18.75 then J=4 (the class of Q_3 is the class number 4 or the class $[25.5 \ ; \ 30.5[$

$$Q_3 = x_J^- + \frac{\frac{3n}{4} - N(x_{J-1})}{n_j}.h_J = 25.5 + \frac{18.75 - 17}{5}.5 = 27.25$$

5- MODE:

The mode is the most frequently occurring member of the data set.The mode (M_O) of a sample is the observation with the highest frequency.

$$M_o = x_s \ \ ; where, \ \ n_s = max\{n_i\} \, ; i = 1 \ldots K$$

Example:

Let the statistical distribution as follow:

x_j	10	11	12	15	16	18	19
n_j	4	6	10	17	11	7	3

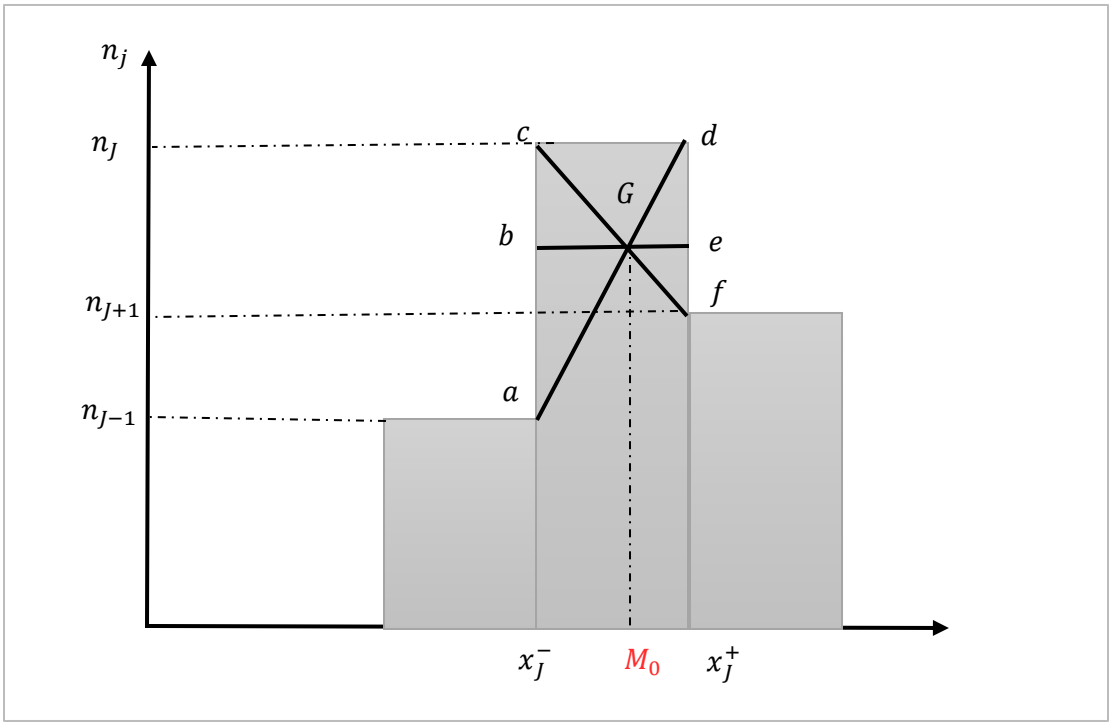
In this distribution, the highest frequency equals 17 corresponding to value x_4=15 then : $M_o = 15$.

- If we have a statistical distribution grouped with classes, the mode can calculated as follow :

- The modal class is the class corresponding to the highest frequency.

-

- $if:\ n_J = max\{n_j\}, j = 1, \dots, K\ \ then\ M_o \in [x_J^-\ ;\ x_J^+[$



We have (acG) and (dfG) two similar triangles, then:

$$\frac{\overline{Gb}}{\overline{ac}} = \frac{\overline{Ge}}{\overline{df}} \ <=> \ \frac{M_o - x_J^-}{n_J - n_{J-1}} = \frac{x_J^+ - M_o}{n_J - n_{J+1}}$$

$$if\ we\ make\ \ n_J - n_{J-1} = \alpha_1\ \ ;and\ \ \ n_J - n_{J+1} = \alpha_2$$

$$then:$$

$$\frac{M_o - x_J^-}{\alpha_1} = \frac{x_J^+ - M_o}{\alpha_2} \Rightarrow (M_o - x_J^-)\alpha_2 = \alpha_1(x_J^+ - M_o)$$

In addition, we know that: $x_J^+ = x_J^- + h_J$

Then

$$(M_o - x_J^-)\alpha_2 = \alpha_1((x_J^- + h_J) - M_o)$$
$$(\alpha_1 + \alpha_2)M_o = \alpha_1(x_J^- + h_J) + \alpha_2 x_J^-)$$

$$\Rightarrow \quad (\alpha_1 + \alpha_2)M_o = (\alpha_1 + \alpha_2)x_J^- + \alpha_1 h_J$$

Then:

$$M_o = x_J^- + \frac{\alpha_1}{\alpha_1 + \alpha_2} h_J$$

On the other hand:

$$M_o = x_J^- + \frac{n_J - n_{J-1}}{2n_J - n_{J-1} - n_{J+1}} h_J$$

Where:

x_J^- : the lower boundary of median class

n_J : frequency of the modal class.

n_{J+1}: the frequency of class after the modal class.

n_{J-1}: the frequency of class befor the modal class.

h_j : lengh of the modal class.

Example: Let the statistical distribution as follow:

classes	n_j
[10.5 ;15.5[	3
[15.5 ;20.5[	6
[20.5 ;25.5[	8
[25.5 ;30.5[	5
[30.5 ;35.5[	3
total	25

We have $\max\{n_j\} = 8$ then J=3 and $M_o \in$ [20.5 ;25.5[.

$$x_J^- = 20.5 \quad ; \quad n_J = 8 \; ; \quad n_{J-1} = 6 \quad ; \quad n_{J+1} = 5 \quad ; \; h_J = 5$$

$$M_o = x_J^- + \frac{n_J - n_{J-1}}{2n_J - n_{J-1} - n_{J+1}} h_J$$

$$M_o = 20.5 + \frac{8-6}{2*8-6-5} 5 = 22.5$$

Remark:

If the lengths of classes are not equal, we must use the adjusted frequencies:

$$(n_j' = \frac{n_j}{h_j})$$

Example: the following table shows the distribution of workers according to monthly wages (1000 DZD)

class	n_j	h_j	n_j'
8 14	36	6	6
14 16	28	2	14
16 20	20	4	5
20 30	16	8	2
total	100	/	/

We have $\max\{n_j'\} = 14$ then $M_o \in [14 \; , \; 16[$

$$M_o = x_J^- + \frac{n_J' - n'_{J-1}}{2n'_J - n'_{J-1} - n'_{J+1}} h_J$$

$$M_o = 14 + \frac{14-6}{2*14-6-5} * 2 = 14.94 \; (1000 \; DZD)$$

Exercises

Exercise 1:

The following data represents the number of years of seniority for 20 employees in a certain institution:

5 8 11 5 2 9 9 5 8 5 10 5 6 5 7 3 12 6 2 1

1. Calculate the average years of seniority.

2. Calculate the mode and the median of this distribution?

3. Find the first and third quartiles.

Exercise 2:

A sample of 110 families distributed according to the number of children as follows:

number of children	0	1	2	3	3	5	Σ
frequency	18	27	27	18	15	5	110

1. Find the average number of children per family.

2. Calculate the mode and median of this distribution.

3. Determine the first and third quartiles.

Exercise 3:

A person raised birds. He had over months, 10, 12, 14, 15 and 17 birds successively.

1. Calculate the monthly average of the number of birds.

2. Calculate the monthly average of the increase in the number of birds.

Exercise 4:

The inflation rates in a certain country are given as follows:

- 2% in the first year

- 5% in the second year

- 12.5% in the third year

1- Calculate the arithmetic and geometric means.

2- Which one is a better summary of the inflation rates?

Exercise 5:

We pulled 800 sheets at a speed of 1000 sheets per hour in one lab, and in another lab, we pulled 2000 sheets at a speed of 700 sheets per hour.

What is the average speed for pulling 3000 sheets?

Exercise 6:

The business revenue for Al-Fateh Textiles Company experienced the following changes over 5 years: 2%, 5%, 2%, -1%, and 4% successively.

1. Calculate the average annual growth rate of the business revenue over 5 years.

2. A second competing company triples its business revenue over 5 years. Calculate the average annual growth rate of the business revenue for this second company.

Exercise 7:

The arable land in the country is distributed across 4 different regions in the following percentages: 1%, 18%, 40%, and 19% of the total arable land in the country. The wheat yields per hectare in each region are as follows: 5, 35, 1, and 9, respectively.

- Calculate the average wheat yield per hectare.

Exercise 8:

The cumulative ascending frequency curve (for the continuous variable (X)) is given by the following relationship:

$$N(x) = \begin{cases} x^2 & if\ 0 \leq x \leq 50 \\ 200\sqrt{12.5(x-37.5)} & if\ 50 \leq x \leq 150 \end{cases}$$

1- Draw the cumulative ascending frequency curve.

2- Identify the median, first quartile (Q_1), and third quartile (Q_3).

3- How many values are enclosed between 25 and 125?

Exercise 9:

The following table represents the distribution of 12,000 workers from Alpha Company based on monthly wages:

number of workers	total monthly wages in thousand DZD
less than 30	2664
30- less than 40	2680
40- less than 45	2156
45- less than 55	2808
55- less than 65	996
65- less than 85	516
85 and more	180
Total	12000

The total wages for the first and last categories are 66.6 million and 16.65 million DZD, respectively.

1. Calculate the average monthly wage.

2. Calculate the mode, and then find the median algebraically and graphically.

Exercise 10:

In order to monitor incentives provided to workers in the telecommunications sector, the Regulatory Authority conducted a study to grant performance bonuses to 50 permanent workers at "Al-Hudhud" Telecommunications Company in the second quarter of 2017. The statistical distribution of the workers according to the bonuses is as follows:

Bonuses in 10^3 DZD	number of workers n_j
10-15	4
15-25	10
25-40	22
40-50	9
50-70	5
Σ	50

1. Calculate the average bonus.

2. Calculate the median, interpret it, and determine it graphically.

3. Calculate the mode of this distribution and interpret it.

4. The company's management decided, in accordance with the income distribution policy enforced by the government, to exempt 20% of workers with low bonuses from mandatory deductions. Determine the bonus range for the workers affected by these mandatory deductions.

5. Furthermore, the company's management, following the previous policy, imposed increasing proportional deductions on 15% of the workers with high bonuses. Determine the bonus range for the workers affected by these proportional deductions.

Exercise 11:

The turnover for Al-Fath Textiles Company defined the following growth rates over 5 years: 2%, 5%, 0%, -1%, and 4%, respectively.

1. What is the appropriate measure to calculate the average annual growth rates for this company's business number?

2. Calculate this average.

3. A competing second company triples its business number over 5 years. Calculate the average annual growth rates for this company.

Assume that the Al-Fath Company's turnover at the beginning of a certain year is $G_1 = 8*10^9$ DA and increases at a rate of (10%) annually.

4. How many years are required for Al-Fath Company's turnover to equal turnover of the competing company, which is 30.5 billion dinars?

Exercise 12:

1. Show that variance equals the mean of the sum of the squares of the variable values minus the square of the mean.

2. Prove the validity of the following statements:

- $n = \sum_{j=1}^{k} n_j \quad \overline{Y} = \frac{1}{n} \sum_{j=1}^{k} Y_j$ with $\sum_{j=1}^{k} (Y_j - \overline{Y}) = 0$
- $\sum_{i=1}^{n} (Y_i - K) = n\overline{Y} - nK$ with K constant

3. Show that the value $\sum_{i=1}^{n} (Y_i - a)$ is the smallest possible when $\overline{Y} = a$

Chapter 04: Measures of Variation and Shape

Chapter 04: Measures of Variation and Shape

I- Measures of Variation (or Spread).

The best-known statistics of variation or spread are for quantitative data. These statistics measure the variation or spread around a central value. Data with the same mean or median may still differ greatly in this respect.

1- **The range**.

The range of a set of observations is the difference between the value of the largest and the smallest observation:

$$R = X_{max} - X_{min}$$

The advantage of the range is its simplicity. A major drawback is that only two observations are used in the calculation. Not all intermediate observations have influence. It is clear that the range is particularly sensitive to extreme values.

- **Example 1:**

Let the following sample data set: 7 , 11 , 11 , 15 , 22 , 22 , 30.

The range:

$$R = X_{max} - X_{min}$$
$$R = 30 - 7 = 23$$

- **Example 2:**

Consider the following statistical distribution:

class	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	total
frequency	4	7	12	8	5	36

$$R = X_{max} - X_{min}$$
$$R = 35 - 10 = 25$$

2- **Interquartile range.**

The interquartile range is defined as the difference between the third and the first quartile:

$$IQ = Q_3 - Q_1$$

Since half of the data is between Q_1 and Q_3 , the interquartile range is a measure of spread for half of the data set. This measure of spread is insensitive to extreme values.

3- Mean Absolute Deviation "MAD":

The mean absolute deviation is the arithmetic mean of absolute values of the deviations of the observations from the arithmetic mean:

$$MAD = \frac{\sum_{i=1}^{n}|x_i - \bar{X}|}{n}$$

or :

$$MAD = \frac{\sum_{j=1}^{J} n_j |x_j - \bar{X}|}{n}$$

Example:

_Let the following Statistical distribution :

class	5 - 9	9 – 13	13 – 17	17 - 21	total
n_j	5	7	10	4	26
x_j	7	11	15	19	
$n_j x_j$	35	77	150	76	338
$\lvert x_j - \bar{X}\rvert$	6	2	2	6	
$n_j \lvert x_j - \bar{X}\rvert$	30	14	20	24	88

- we have $\bar{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j$ = 1/26 (338) = 13
- then $MAD = \frac{\sum_{j=1}^{J} n_j |x_j - X|}{n} = \frac{88}{26} = 3.38$

4- Variance

For sample data measured on a ratio scale or an interval scale, the variance is used more often as a measure of spread than the mean absolute deviation.

The variance V(X) of a set of observations $x_1, x_2, \ldots, x_n$, is the mean of the squared deviations from the arithmetic mean ($\bar{X}$) :

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

$$or: \quad V(X) = \frac{1}{n}\sum_{i=1}^{J} n_j \left(x_j - \bar{X}\right)^2$$

and:

$$V(X) = \frac{1}{n}\sum_{i=1}^{J} n_j \left(x_j - \bar{X}\right)^2 = \frac{1}{n}\sum_{j=1}^{J} (n_j x_j^2 - 2n_j\bar{X} + n_j(\bar{X})^2)$$

$$V(X) = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{X}^2$$

The **standard deviation** is defined as follow:

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

Example:

Calculate the standard deviation of the following statistical distribution of the ages of the sports club members (years):

classes	n_j	X_j	$n_j X_j$	$X_j^2 n_j$
16.5 21.5	4	19	76	1444
21.5 26.5	9	24	216	5184
26.5 31.5	24	29	696	20184
31.5 36.5	9	34	306	10404
36.5 41.5	4	39	156	6084
$\sum$	50	/	1450	43300

We have the arithmetic mean: $\bar{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j = \frac{1450}{50} = 29$ year

then the variance :

$$V(X) = \frac{1}{n}\sum_{j=1}^{J} n_j x_j^2 - \bar{X}^2$$

$$V(X) = \frac{1}{50}(43300) - (29)^2 = 866 - 841 = 25 \ year^2$$

The standard deviation:

$$\sigma = \sqrt{V(X)} = \sqrt{25} = 5 \text{ Years}$$

Among them, the dispersion of the ages of the sports club members is equal to 5 years, which is a weak dispersion compared to the observations.

- **Proprieties.**

- The variance is always positive (same thing for standard deviation).
- The variance of a set of equal values is zero.

$$\text{If } x_i = C \,, \forall i = 1,..,n \;\; then \;\; V(X) = V(C) = 0$$

- The variance is affected by a few extremely large or extremely small observations (outliers).
- If $y_i = ax_i + b$ (linear transformation), then :

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

Proof:

$$V(Y) = \frac{1}{n}\sum_{j=1}^{J} n_j\left(y_j - \bar{Y}\right)^2 = \frac{1}{n}\sum_{j=1}^{J} n_j(ax_i + b - (a\bar{X} + b))^2$$

$$V(Y) = \frac{1}{n}\sum_{j=1}^{J} n_j\left(ax_j - a\bar{X}\right)^2 = a^2\,\frac{1}{n}\sum_{j=1}^{J} n_j\left(x_j - \bar{X}\right)^2 = a^2 V(X)$$

- The variance is the smallest mean square of the deviation.

Proof:

$$eqm = \frac{1}{n}\sum_{j=1}^{J} n_j\left(x_j - \beta\right)^2 = \frac{1}{n}\sum_{j=1}^{J} n_j\left(x_j - \beta + \bar{X} - \bar{X}\right)^2$$

$$= \frac{1}{n}\sum_{j=1}^{J} n_j\left[\left(x_j - \bar{X}\right) - (\beta - \bar{X})\right]^2$$

$$= \frac{1}{n}\sum_{j=1}^{J} n_j\left[\left(x_j - \bar{X}\right)^2 + (\beta - \bar{X})^2 - 2\left(x_j - \bar{X}\right)(\beta - \bar{X})\right]$$

$$= \frac{1}{n}\sum_{j=1}^{J} n_j\left(\left(x_j - \bar{X}\right)^2 + \frac{1}{n}\sum_{j=1}^{J} n_j\,(\beta - \bar{X})^2 - \frac{2}{n}\sum_{j=1}^{J} n_j\left(x_j - \bar{X}\right)(\beta - \bar{X})\right)$$

$$= \frac{1}{n}\sum_{j=1}^{J} n_j \left(\left(x_j - \bar{X}\right)^2 \right) + (\beta - \bar{X})^2 - \frac{2}{n}(\beta - \bar{X}) \sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right) \Big)$$

We know that:

$$\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right) = 0 \text{ , and } \quad \frac{1}{n}\sum_{j=1}^{J} n_j \left(\left(x_j - \bar{X}\right)^2 \right) = V(X)$$

Then: $eqm = V(X) + (\beta - \bar{X})^2$

We get: $eqm \geq V(X)$

If $\beta = \bar{X}$ then: $eqm = V(X)$

5- Coefficient of variation :

Although variance and standard deviation play an extremely important role in statistics, they are not always the best measures of spread

If we consider the following two data sets with eight observations:

- Sample 1: 15 , 20 , 20 , 30 , 35 , 35 , 40 , 45 .
- Sample 2: 1015 , 1020 , 1020 , 1030 , 1035, 1035 ,1040 , 1045 .

The arithmetic means of the two sets are 30 and 1030, while the sample variance is 114.2857 for both samples; therefore, the sample standard deviation is also the same for both samples, equal to 10.69. Nevertheless, it is clear that—relative to the arithmetic mean—the variability in the second sample is considerably smaller than in the first. In this case, we use the coefficient of variation to compare the dispersion between the two samples.

Definition:

The coefficient of variation "CV" is defined as the ratio of the standard deviation and the arithmetic mean:

$$CV = \frac{\sigma}{\bar{X}} \, 100$$

Example:

Taking the data from the two previous samples, we find:

$$\overline{X_1} = 30 \quad , \quad \overline{X_2} = 1030$$
$$V_1(X) = V_2(X) = 114.2857$$
$$\sigma_1(X) = \sigma_2(X) = 10.69$$

Therefore, the coefficients of variation for two samples are:

$$CV_1(X) = \frac{10.69}{30}.100 = 0.356$$

$$CV_2(X) = \frac{10.69}{1030}.100 = 0.0104$$

We remark that $CV_2(X) < CV_1(X)$ .then; we conclude that the dispersion is smallest in the second sample.

The coefficient of variation is unreliable when $\bar{X}$ is very small and it is sensitive to outliers.

The coefficient of variation is useful for comparing spread of data with different means and indispensable when comparing the variation of data with different dimensions (data expressed in different units of measurement).

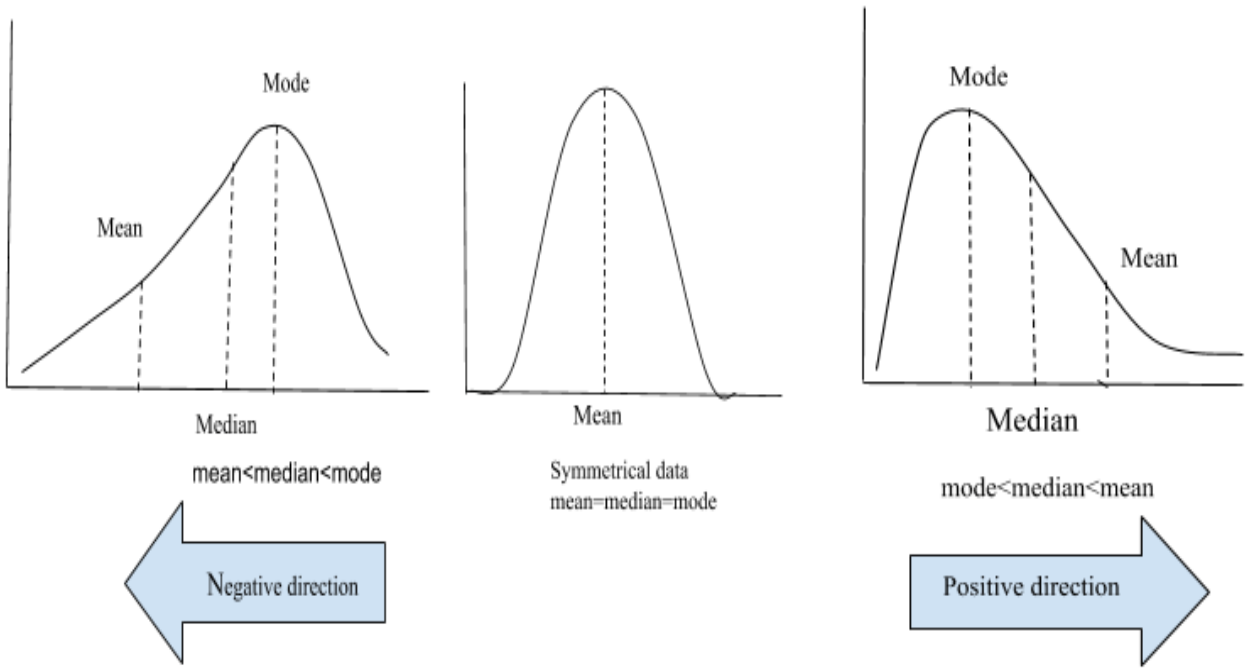
II- **MEASURES OF SHAPE: SKEWNESS and KURTOSIS.**

By studying the shape of a frequency distribution, we mean to compare this distribution with the normal distribution (Symmetrical distribution) in terms of symmetry and length.

For this purpose, we use other two statistical measures that compare the shape to the normal curve called **Skewness** and **Kurtosis**.

1- **SKEWNESS.**

The graphical representation of a frequency distribution of a statistical variable can be symmetric or asymmetric. A histogram that is not symmetric is called skewed. In a histogram that is skewed to the left (or negatively skewed), the left-hand tail is longer than the right-hand tail. In a histogram that is skewed to the right (or positively skewed), the right-hand tail is longer than the left-hand tail. In a unimodal histogram, skewness can be determined based on the positions of the arithmetic mean, the median and the mode. In a perfectly symmetrical histogram, the three statistics (measurements) are identical. In a histogram that is skewed to the left, the mean is smaller than the median, which in turn is smaller than the mode. When a histogram is skewed to the right, the mode is smaller than the median, which itself is smaller than the arithmetic mean. The reason for this is that the arithmetic mean is more sensitive to extremely large or extremely small values than the median.



1.1-Measures of skewness.

a)- **The Pearson's coefficient**:

The Pearson's coefficient of skewness is defined as follow:

$$S_p = \frac{3(\bar{X} - M_e)}{\sigma}$$

- if $S_p > 0$ then we **have a positively skewed** (right-skewed)
- if $S_p < 0$ then we **have a negatively skewed** (left-skewed)
- if $S_p = 0$ then we **have no skew** (symmetric)

b)- **Fisher's coefficient**:

Fisher's coefficient of skewness is defined as follow:

$$S_F = \frac{\mu_3}{\sigma^3}$$

Where: σ: standard deviation.

μ_3: The central moment of third degree,

$$\mu_3 = \frac{1}{n}\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^3$$

Remark:

The central moment of degree "k" is: $\mu_k = \frac{1}{n}\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^k$ and the non-central moment of degree "k" is: $m_k = \frac{1}{n}\sum_{j=1}^{J} n_j \left(x_j\right)^k$

- if $S_F > 0$ then we **have a positively skewed** (right-skewed)
- if $S_F < 0$ then we **have a negatively skewed** (left-skewed)
- if $S_F = 0$ then we **have no skew** (symmetric)

Example:

Let the following statistical distribution of the ages of the sports club members

classes (years)	n_j	X_j	$n_j X_j$	$X_j^2 n_j$	$n_j\left(x_j - \bar{X}\right)^3$	N(X)
16.5 21.5	4	19	76	1444	- 4000	4
21.5 26.5	9	24	216	5184	-1125	13
26.5 31.5	24	29	696	20184	0	37
31.5 36.5	9	34	306	10404	1125	46
36.5 41.5	4	39	156	6084	4000	50
$\sum$	50	/	1450	43300	0	///////

a)- The Pearson's coefficient of skewness is defined as:

$$S_p = \frac{3(\bar{X} - M_e)}{\sigma}$$

we have $\bar{X} = 29$ and $\sigma = 5$

and

$$M_e = X_k^- + \frac{\frac{n}{2} - N(X_{k-1})}{n_k} h_k = 26.5 + \frac{25 - 13}{24}\, 5 = 29\, year$$

then

$$S_p = \frac{3(\bar{X} - M_e)}{\sigma} = \frac{3(29-29)}{5} = 0$$

Then we have a symmetric distribution.

b)- Fisher's coefficient of skewness is defined as:

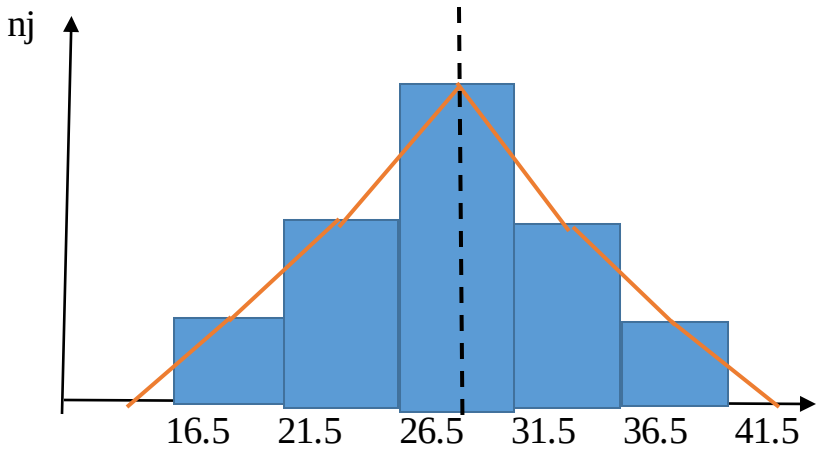
$$S_F = \frac{\mu_3}{\sigma^3}$$

we have

$$\mu_3 = \frac{1}{n}\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^3 = 0$$

and $S_F = 0$ then we have a symmetric distribution.

we can see a symmetric distribution by the histogram as follow :



$$\bar{X} = M_e = M_o$$

2- KURTOSIS:

Kurtosis is a statistical number that tells us if a distribution is taller or shorter than a normal distribution. If a distribution is similar to the normal distribution, the kurtosis value is " 0 " . If kurtosis is greater than " 0 ", then it has a higher peak compared to the normal distribution . If kurtosis is less than "0" , the it is flatter than a normal distribution.

There are three types of distributions:

- Leptokurtic : sharply peaked with fat tails, and less variable .
- Mesokurtic : Medium peaked. (normal)
- Platykurtic : Flattest peak and highly dispersed .

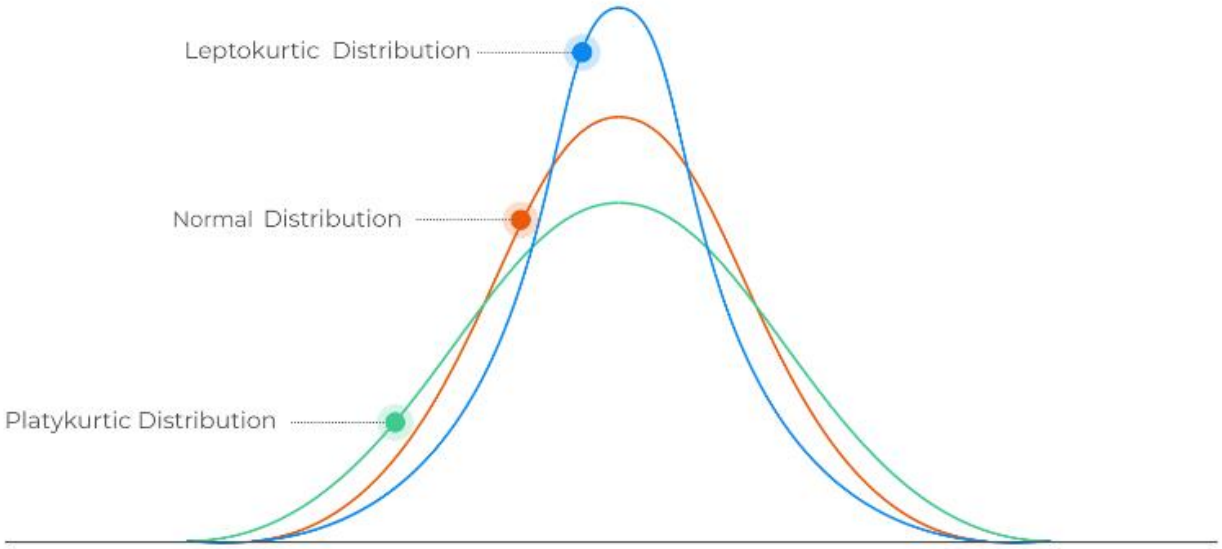
Measures of kurtosis.

we can measure the kurtosis by deferent measures as Fisher's coefficient :

$$\boldsymbol{K_F = \frac{\mu_4}{\sigma^4} - 3}$$

Where: $\mu_4 = \frac{1}{n}\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^4$

- if $K_F > 0$ $,then\ we\ have\ leptokurtic\ distribution.$
- if $K_F < 0$ $,then\ we\ have\ platykurtic\ distribution.$
- if $K_F = 0$ $,then\ we\ have\ mesokurtic\ (normal)\ distribution$



Example:

Study the Kurtosis of the following statistical distribution of the ages of the sports club members (years):

classes	n_j	X_j	$n_j X_j$	$X_j^2 n_j$	$n_j(x_j - \bar{X})^4$
16.5 21.5	4	19	76	1444	40000
21.5 26.5	9	24	216	5184	5625
26.5 31.5	24	29	696	20184	0
31.5 36.5	9	34	306	10404	5625
36.5 41.5	4	39	156	6084	40000
$\sum$	50	/	1450	43300	91250

We have the Fisher's coefficient:

$$K_F = \frac{\mu_4}{\sigma^4} - 3$$

Where: $\mu_4 = \frac{1}{n}\sum_{j=1}^{J} n_j \left(x_j - \bar{X}\right)^4 = \frac{91250}{50} = 1825$

Then:

$$K_F = \frac{\mu_4}{\sigma^4} - 3 = \frac{1825}{(5)^4} - 3 = -0.08$$

We note that the coefficient is negative but close to " 0 " ,so it can be said that this distribution is mesokurtic (normal).

Chapter 05: Measures of Concentration. (Lorenz Curve and GINI Coefficient)

Chapter 05: Measures of Concentration.

(Lorenz Curve and GINI Coefficient)

The term "economic concentration "has been employed in many different sense and "indexes of concentration have been constructed.

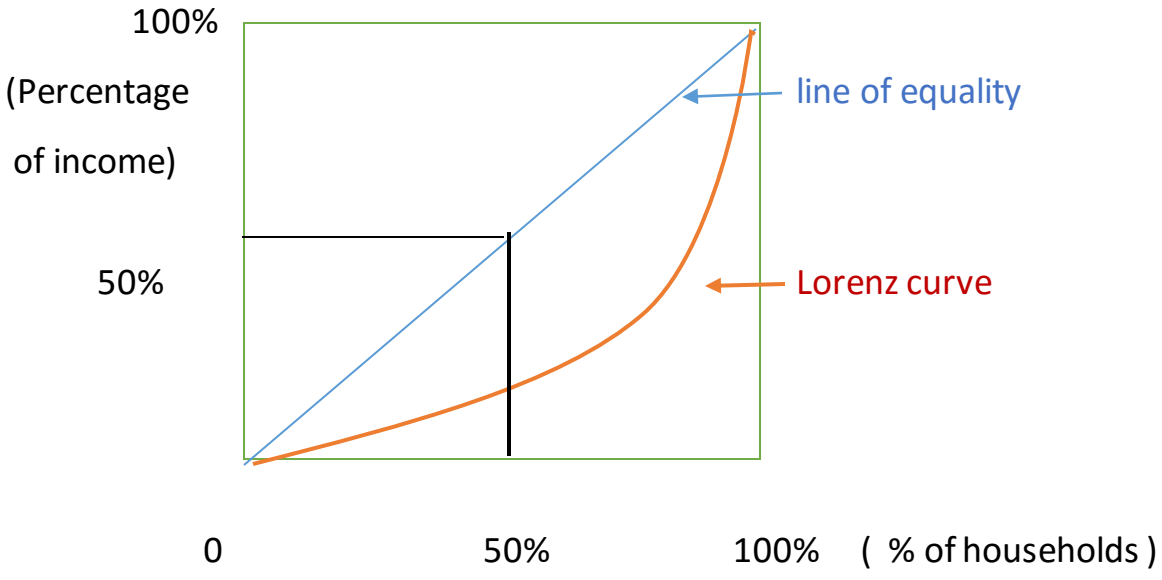
The concentration measures are usually used to determine the degree of equality in the distribution of wealth among members of a people.

1- LORENZ CURVE .

A straight diagonal line, which represents perfect equality in income or wealth distribution, represents the Lorenz curve; the Lorenz curve lies beneath it showing estimated distribution. The Lorenz curve is used to represent economic inequality as well as unequal wealth distribution. The farther away the curved line is way from the straight diagonal line, the higher the level of inequality.

Constructing a Lorenz curve involves fitting a continuous function to some incomplete set of data, there is no guarantee that the values along a Lorenz curve (other than those actually observed in the data) actually correspond to the true distributions of income.

A Lorenz curve plots the cumulative proportion of the population under study ranked by the socioeconomic stratified, beginning with the lowest socioeconomic group of the population (the most vulnerable), against the cumulative proportion of wealth.



2- GINI Coefficient.

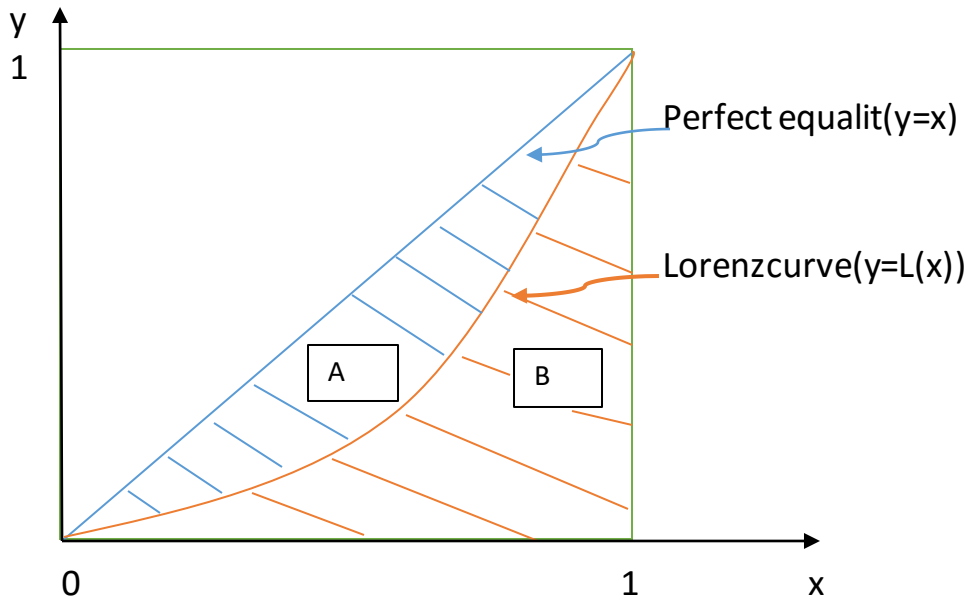
The Gini coefficient is calculated as a ratio of the area on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve equals A , and the area underneath the Lorenz curve is B, than the GINI Coefficient is:

$$G = \frac{A}{A + B}$$

This ratio is expressed as a percentage or as the numerical equivalent of that percentage, which is always a number between 0 and 1.

$$0 \leq G \leq 1$$

- if G=0 : we have a perfect equality.
- if G=1 : we have a perfect inequality.



We have:

$$G = \frac{A}{A + B}$$

Where:

$$A = \int_0^1 \big(x - L(x) \big) dx = \frac{1}{2} - \int_0^1 L(x) dx$$

And: $B = \int_0^1 L(x)dx$

Then:

$$G = \frac{A}{A+B} = \frac{\frac{1}{2} - \int_0^1 L(x)dx}{\frac{1}{2} - \int_0^1 L(x)dx + \int_0^1 L(x)dx} = \frac{1/2 - \int_0^1 L(x)dx}{1/2}$$

$$= 1 - 2\int_0^1 L(x)dx$$

Or: $\boldsymbol{G = 1 - 2\int_0^1 L(x)dx}$

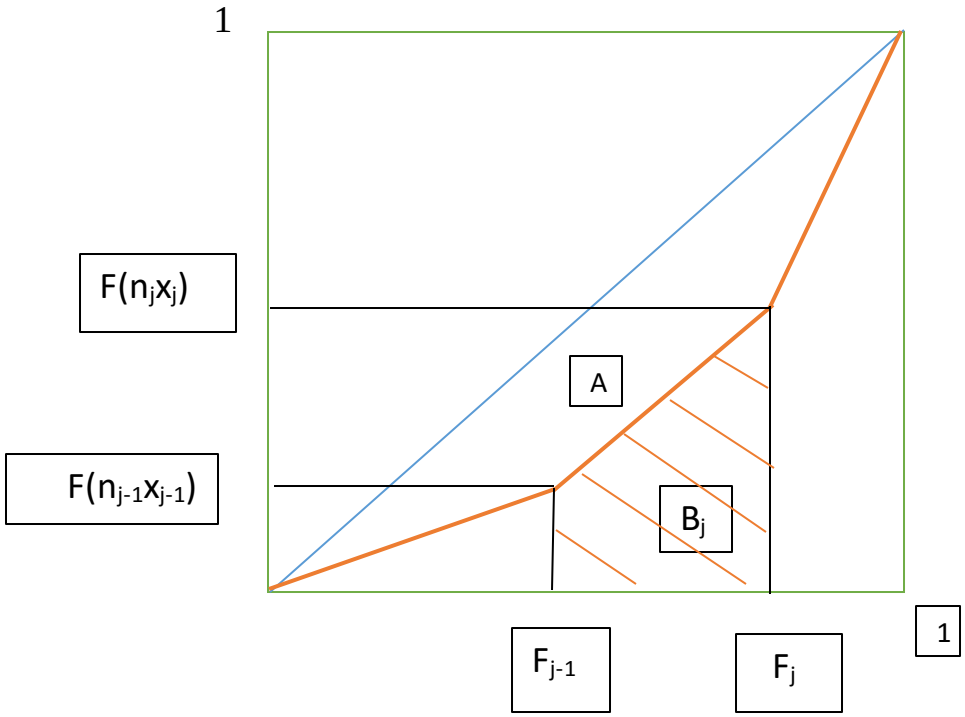
Remark: if we have a statistics distribution of wealth (or income , or wages) , then the Lorenz curve is a broken line as follows :

wages classes	$[x_1^-; x_1^+[$	…….	$[x_j^-; x_j^+[$	…….	$[x_K^-; x_K^+[$	total
(x_j)	x_2	……..	x_j	…….	x_K	
number of workers (n_j)	n_1	…….	n_j	…….	n_K	n
relative frequency (f_j)	f_1	…….	f_j	…….	f_K	1
cumulative frequency (F_j)	F_1	…….	F_j	…….	1	
total wage of class $(n_j.x_j)$	$n_1.x_1$	…….	$n_j.x_j$	…….	$n_K.x_K$	S= $\sum_{j=1}^{K} n_j x_j$
wage ration for each class $f(n_j x_j)$	$f(n_1 x_1)$	…….	$f(n_j x_j)$	……..	$f(n_K x_{jK})$	1
cumulative frequency of wages $(F(n_j x_j))$	$F(n_j x_j)$	…….	$F(n_j x_j)$	…….	1	

where :

$$f(n_j x_j) = \frac{n_j . x_j}{\sum_{j=1}^{K} n_j x_j}$$

- the Lorenz curve represents the relationship between the cumulative frequencies (F_j) and the cumulative frequencies of wages ,or wealth $F(n_j x_j)$, as follow :



$$G = \frac{A}{A + B}$$

$$A = \frac{1 * 1}{2} - B \Rightarrow \quad A = \frac{1}{2} - B$$

where: $B = B_1 + B_2 + \cdots . + B_K = \sum_{j=1}^{K} B_j$

B_j Is the trapezoid area, then:

$$B_j = \frac{\left[F(n_{j-1}x_{j-1}) + F(n_j x_j)\right]}{2} * (F_j - F_{j-1}) = \frac{\left[F(n_{j-1}x_{j-1}) + F(n_j x_j)\right]}{2} * f_j$$

and :

$$B = \sum_{j=1}^{K} \frac{[F(n_{j-1}x_{j-1}) + F(n_j x_j)]}{2} * f_j$$

we have :

$$G = \frac{A}{A+B} = \frac{\frac{1}{2} - B}{\frac{1}{2} - B + B} = 1 - 2B = 1 - 2\sum_{j=1}^{K} \frac{[F(n_{j-1}x_{j-1}) + F(n_j x_j)]}{2} * f_j$$

then:

$$\boldsymbol{G = 1 - \sum_{j=1}^{K} [F(n_{j-1}x_{j-1}) + F(n_j x_j)] * f_j}$$

Example: Let the statistical distribution of monthly wages (1000 AD) for a group of workers, as follow:

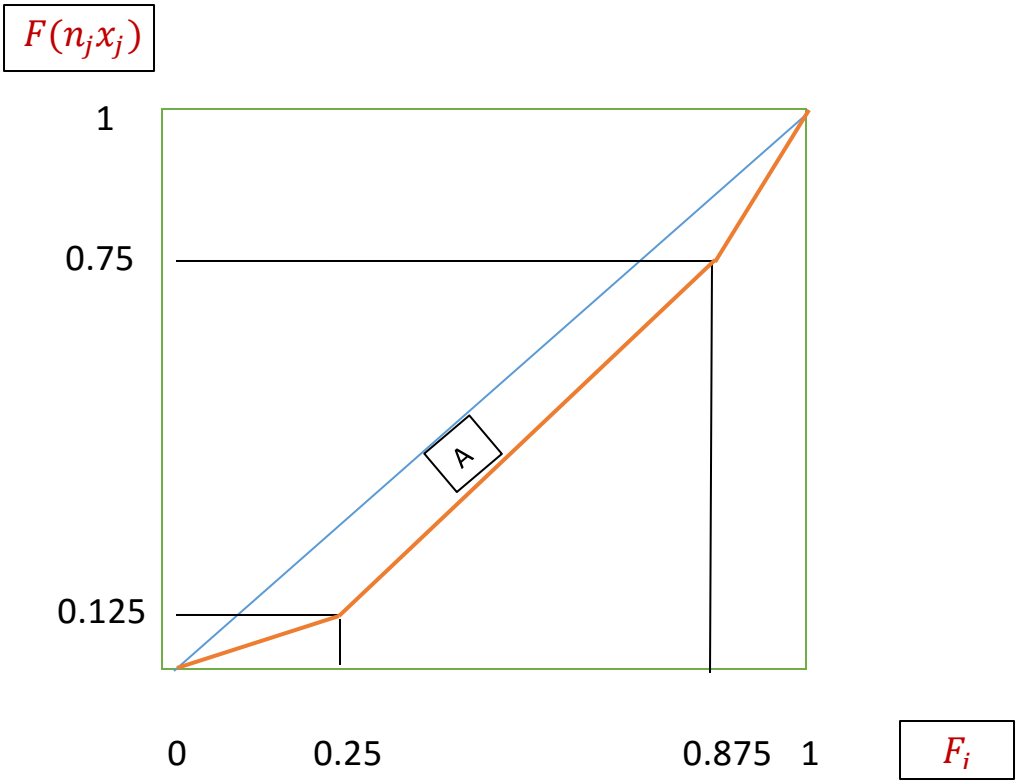
wages	5 - 15	15 - 25	25 - 55	total
frequency	50	125	25	200

We calculate the GINI coefficient.

- the Lorenz curve:

wages	5 - 15	15 - 25	25 - 55	total
frequency (n_j)	50	125	25	200
f_j	0.25	0.625	0.125	1
F_j	0.25	0.875	1	
x_j	10	20	40	
$n_j x_j$	500	2500	1000	S=4000
$f(n_j x_j)$	0.125	0.625	0.25	1
$F(n_j x_j)$	0.125	0.75	1	

The Lorenz curve represent the relationship between F_j and $F(n_j x_j)$.



- **GINI coefficient**:

$$G = 1 - \sum_{j=1}^{K} \left[F\left(n_{j-1} x_{j-1}\right) + F\left(n_j x_j\right) \right] * f_j$$

$$G = 1 - \sum_{j=1}^{3} \left[F\left(n_{j-1} x_{j-1}\right) + F\left(n_j x_j\right) \right] * f_j$$

$$G = 1 - [(0 + 0.125) * 0.25 + (0.125 + 0.75) * 0.625 + (0.75 + 1) * 0.125]$$

G= 0.20=20%

There is weak concentration of wages among workers.

Chapter 06:
Index Numbers

Chapter 06: Index Numbers

An index number is a method of evaluations a variable or group of variables in regards to geographical location, time, and other features.

Index numbers are one of the most widely used statistical tools. Some of the advantages or uses of index numbers are as follows:

- Help in study of trends: index numbers help in the study of trends in variables like export, import, industrial and agricultural production, share prices, and more.
- Help in formulating policies: most of the economic and business decisions and policies are guided by the index numbers.
- Helpful in forecasting: index numbers not only help in the study of past and present behavior, they are also used for forecasting economic and business activities.
- Facilitates comparative study: To make comparisons with respect to time and place especially where units are different, index numbers prove to be very useful.
- Measurement of purchasing power of money to maintain standard of living: index numbers such as cost inflation index help in measuring the purchasing power of money at different times between different regions.

- **Definition**:

  Index number is a percentage ratio, which measures the average change of several variables between two different times, places or situation.

1- Calculating a simple index number.

To convert a time series (or any other data series) to an index series, we first choose one period as the base period (usually labelled "period 0 ") and then for every other observation of the variable, we divide the value of series by the value in the base period. For ease of comparison index numbers are often expressed with the value in the base period set to 100 , the value of the index is multiplied by 100 (this is optional but common).

Let $\boldsymbol{X^t}$ represent the value of a variable **X** at a time "**t**" and $\boldsymbol{X^0}$ represent the value of the variable **X** at a time "**0**". Then, the simple index number for the variable X at the current time "t" with base period "0" is given by:

$$I_{(t/0)} = \frac{X^t}{X^0} . 100$$

From this equation, it can be seen that the index with time "t" set to be the base period "0" will have a value of 100, and if the variable X increase in later time periods, the value of the index will be greater than 100.

Example:

The next table shows the annual turnover for a company "A" between 2002 and 2007 ($ million):

year	2002	2003	2004	2005	2006	2007
turnover	17.9	18.5	18.1	18.6	19.0	19.1

- Calculate a simple index series of turnover for company "A", with 2002 as the base year.
-

Solution:

we have : $I_{t/0} = \frac{X^t}{X^0} . 100$

- for the year 2002, (t=0):

$$I_{0/0} = \frac{X^0}{X^0} . 100 = \frac{17.9}{17.9} . 100 = 100$$

- for the year 2003, (t=1):

$$I_{1/0} = \frac{X^1}{X^0} . 100 = \frac{18.5}{17.9} . 100 = 103.35$$

- for the year 2004, (t=2):

$$I_{2/0} = \frac{X^2}{X^0} . 100 = \frac{18.1}{17.9} . 100 = 101.17$$

- for the year 2005, (t=3):

$$I_{3/0} = \frac{X^3}{X^0} . 100 = \frac{18.6}{17.9} . 100 = 103.91$$

- for the year 2006, (t=4):

$$I_{4/0} = \frac{X^4}{X^0} . 100 = \frac{19.0}{17.9} . 100 = 106.14$$

- for the year 2007, (t=5):

$$I_{5/0} = \frac{X^5}{X^0} . 100 = \frac{19.1}{17.9} . 100 = 106.7$$

These results can be represented in the following table:

year (t)	2002	2003	2004	2005	2006	2007
index number ($I_{t/0}$)	100	103.35	101.17	103.91	106.14	106.7

- Index number and percentage change.

The percentage change (g) in a series between point A and point B is calculated as:

$$g = \frac{B - A}{A} . 100$$

That is, the difference between the levels of the index series at the two time points, divided by the value of the series at the earlier time point, multiplied by 100 in order to express the value as percentage.

In the case of index numbers, we can calculate the growth in an index between two periods, time "s" and time "t" (where s<t) as:

$$g_{(t/s)} = \frac{I_{(t/0)} - I_{(s/0)}}{I_{(s/0)}} . 100$$

or,

$$g_{(t/s)} = \frac{\frac{X^t}{X^0} - \frac{X^s}{X^0}}{\frac{X^s}{X^0}} . 100 = \frac{X^t - X^s}{X^s} . 100 = \left(\frac{X^t}{X^s} - 1\right) . 100$$

$$= \left(\frac{X^t}{X^s} . 100 - 100\right) = I_{(t/s)} - 100$$

if s=0 then:

$$g_{(t/0)} = I_{(t/0)} - 100$$

or:

$$I_{(t/0)} = g_{(t/0)} + 100$$

Example:

We found in the previous example, the index number of the turnover in 2007 (base 2002) equals $I_{(2007/2002)} = 106.7$, this means that the turnover for this company increased by 6.7%, (106.7-100=6.7) in 2007 compared to 2002.

-The proprieties of index numbers.

a- The time reversal propriety.

Let $I_{(t/0)}$ the index number of the current time "t" with base period "0", and $I_{(0/t)}$ the index number of the base time "0" with current period "t", then:

$$I_{(t/0)}.I_{(0/t)} = 100^2$$

or:

$$I_{(t/0)} = \frac{X^t}{X^0}.100 \Rightarrow \frac{1}{I_{(t/0)}} = \frac{1}{\frac{X^t}{X^0}.100} = \frac{X^0}{X^t.100} = \frac{X^0}{X^t}.\frac{100}{100^2}$$

$$= I_{(0/t)}.\frac{1}{100^2} \Rightarrow I_{(t/0)}.I_{(0/t)} = 100^2$$

$$if \quad I_{(\frac{2003}{2002})} = 103.35 \qquad then \; I_{(\frac{2002}{2003})} = \frac{100^2}{103.35} = 96.76$$

Note:

If we consider the index number without multiplying the result by 100 ($I_{(t/0)} = \frac{X^t}{X^0}$) then:

$$I_{(t/0)}.I_{(0/t)} = 1$$

b- Circular property:

This property indicates that :

$$\boldsymbol{I_{(t/0)} = I_{(t/t-1)}.I_{(t-1/t-2)} \ldots\ldots I_{(1/0)}.\frac{1}{(100)^{T-1}}}$$

If we consider the index number without multiplying the result by 100 ($I_{(t/0)} = \frac{X^t}{X^0}$) then:

- $I_{(t/0)} = I_{(t/t-1)}.I_{(t-1/t-2)} \ldots\ldots I_{(1/0)} = \frac{X^t}{X^{t-1}}\frac{X^{t-1}}{X^{t-2}} \ldots\ldots \frac{X^2}{X^1}\frac{X^1}{X^0} = \frac{X^t}{X^0}$

Example:

$$*) \quad I_{(2/0)} = I_{(2/1)}.I_{(1/0)}\frac{1}{100}$$

$$**) \quad I_{(3/0)} = I_{(3/2)}.I_{(2/1)}.I_{(1/0)}.\frac{1}{100^2}$$

- and: $$\boldsymbol{I_{(t/0)}.I_{(t-1/t)}.I_{(t-2/t-1)} \ldots\ldots I_{(0/1)} = 1}$$

where: $\boldsymbol{I_{(t/0)}.I_{(t-1/t)}.I_{(t-2/t-1)} \ldots\ldots I_{(1/2)}\, I_{(0/1)} = \frac{X^t}{X^0}\frac{X^{t-1}}{X^t}\frac{X^{t-2}}{X^{t-1}}\ldots\frac{X^1}{X^2}\frac{X^0}{X^1} = 1}$

Example : $\boldsymbol{I_{(2/0)}.I_{(1/2)}.I_{(0/1)} = \frac{X^2}{X^0}\frac{X^1}{X^2}\frac{X^0}{X^1} = 1}$

Exercise :
Complete the following table representing an index numbers:

2013	2012	2011	2010	base year / current year
…	…	…	100	2010
…	50	100	80	2011
40	100	…	…	2012
100	…	…	…	2013

- solution:

- we have $\boldsymbol{I_{(11/10)} = 80}$ **,then** $\boldsymbol{I_{(10/11)} = \frac{100^2}{I_{(11/10)}} = \frac{10000}{80} = 125}$
- we have $\boldsymbol{I_{(11/12)} = 50}$ **, then** $\boldsymbol{I_{(12/11)} = \frac{10000}{50} = 200}$
- we have $\boldsymbol{I_{(12/13)} = 40}$ **, then** $\boldsymbol{I_{(13/12)} = \frac{10000}{40} = 250}$
- $\boldsymbol{I_{(12/10)} = I_{(12/11)}I_{(11/10)}\frac{1}{100} = 200 * 80 * \frac{1}{100} = 160}$
- $\boldsymbol{I_{(13/10)} = I_{(13/12)} * I_{(12/10)} * \frac{1}{100} = 250 * 160 * \frac{1}{100} = 400}$
- $\boldsymbol{I_{(13/11)} = I_{(13/12)} * I_{(12/11)} * \frac{1}{100} = 250 * 200 * \frac{1}{100} = 500}$
- $\boldsymbol{I_{(10/12)} = \frac{10000}{I_{(12/10)}} = \frac{10000}{160} = 62.5}$
- $\boldsymbol{I_{(10/13)} = \frac{10000}{I_{(13/10)}} = \frac{10000}{400} = 25}$
- $\boldsymbol{I_{(11/13)} = \frac{10000}{I_{(13/11)}} = \frac{10000}{500} = 20}$

2013	2012	2011	2010	base year / current year
25	62.5	125	100	2010
20	50	100	80	2011
40	100	200	160	2012
100	250	500	400	2013

- **Some types of simple index numbers.**

Among the most important types of index numbers, we mention:

1- The Price index number.

The simple price index number is:

$$\boldsymbol{PI_{(t/0)} = \frac{P^t}{P^0}\ 100}$$

where: P^t : represent the level price in current time period (current year)

P^0 : represent the level price in base time period (base year)

2- Quantity index number.

The simple quantity index number is:

$$\boldsymbol{QI_{(t/0)} = \frac{Q^t}{Q^0}\ 100}$$

where: Q^t : represent the level quantity in current time period (current year)

Q^0 : represent the level quantity in base time period (base year)

3- Value index number.

The simple value index number is:

$$\boldsymbol{VI_{(t/0)} = \frac{V^t}{V^0}\ 100 = \frac{\sum_{i=1}^{n} P_i^t\, Q_i^t}{\sum_{i=1}^{n} P_i^0\, Q_i^0}\ 100}$$

where: P_i^t and $\boldsymbol{P_i^0}$: represent the level price of the commodity "I" in current time period (current year) and in the base year respectively.

Q_i^t *and* $\boldsymbol{Q_i^0}$: represent the level quantity of the commodity "I" in current year and in base time period (base year) respectively.

n: represent the number of commodities .

Example :

The prices and quantities of three commodities produced by an company reached the following levels within two years :

comodity	2019		2020	
	Prices (P_i^0)	Quantities (Q_i^0)	Prices (P_i^t)	Quantities (Q_i^t)
A (Kgs)	20	10	22	15
B (Liters)	50	20	55	25
C (m)	25	40	30	50

- calculate the value index number in 2020 base 2019
  - solution:

the value index number in 2020 base 2019 is:

$$VI_{(2020/2019)} = \frac{\sum_{i=1}^{3} P_i^t Q_i^t}{\sum_{i=1}^{3} P_i^0 Q_i^0} 100 = \frac{22*15+55*25+30*50}{20*10+50*20+25*40}.100$$
$$= 145.68$$

the value of total product in this company is increase by 45.68% in 2020 compared by 2019.

2- Complex indices (Weighted index).

If, instead of calculating simple price (or quantity) indices, the aim is to compile an aggregate index that considers the behaviour of prices as a whole, or the general level of prices (or quantities), the problem of aggregation arises: how should heterogeneous products such as oil and bread be added together?

The weighted index numbers are the weighted means (averages) of a simple index numbers. They differ from one another according to the weightings and the type of mean (average) that is used in the calculation. The most important of them are:

2.1- Laspeyres index number.

In this case, we use the weighted arithmetic mean of simple index where the weightings are calculated in base time period.

- **Laspeyres Price Index (LPI):**

The laspeyres price index (LPI), is weighted arithmetic mean of simple price index numbers for a group of goods (n goods) that uses the expenditure shares in the base time period (0) as weights.

We assume a basket of goods (n goods) where each good (i) has a price P_i^0 and quantity Q_i^0 in the base year, but in the current year (t), the prices and quantities of these goods changed to P_i^t and Q_i^t respectively, so we have the following:

- The simple price index number for each goods (i) is:

$$\left(PI_{(t/0)}\right)_i = \frac{P_i^t}{P_i^0}\ 100$$

- The weight for each simple price index number (i) is:

$$W_i^0 = \frac{P_i^0 Q_i^0}{\sum_{i=1}^n P_i^0\ Q_i^0}$$

then the laspeyers price index is:

$$LPI_{(t/0)} = \sum_{i=1}^{n} W_i^0\ \left(PI_{(t/0)}\right)_i$$

$$= \sum_{i=1}^{n} \frac{P_i^0 Q_i^0}{\sum_{i=1}^n P_i^0\ Q_i^0}\left(\frac{P_i^t}{P_i^0}\ 100\right) = \frac{\sum_{i=1}^n P_i^t Q_i^0}{\sum_{i=1}^n P_i^0 Q_i^0}\ 100$$

the laspeyers price index in current year (t) compared to base year (0) is:

$$LPI_{(t/0)} = \frac{\sum_{i=1}^n P_i^t Q_i^0}{\sum_{i=1}^n P_i^0 Q_i^0}\ 100$$

- **Laspeyres Quantity Index (LQI).**

in the same way we find the laspeyers quantity index in current year (t) compared to base year (0) is:

$$LQI_{(t/0)} = \frac{\sum_{i=1}^n Q_i^t P_i^0}{\sum_{i=1}^n Q_i^0 P_i^0}\ 100$$

2.2- Paasche index number.

The paasche price index is a weighted harmonic mean (average) of the price relatives (simple price index) that uses the actual expenditure shares in the current year (period) "t" as weights.

- **Paasche Price Index (PPI)**

We assume a basket of goods (n goods) where each good (i) has a price P_i^0 and quantity Q_i^0 in the base year, but in the current year (t), the prices and quantities of these goods changed to P_i^t and Q_i^t respectively, so we have the following:

- The simple price index number for each goods (i) is:

$$\left(PI_{(t/0)}\right)_i = \frac{P_i^t}{P_i^0}\ 100$$

- The weight for each simple price index number (i) is:

$$W_i^t = \frac{P_i^t Q_i^t}{\sum_{i=1}^n P_i^t\ Q_i^t}$$

then the Paasche price index is the harmonic mean of simple price indexes :

$$PPI_{(t/0)} = \frac{1}{\sum_{i=1}^n \frac{W_i^t}{\left(PI_{(t/0)}\right)_i}} = \frac{1}{\sum_{i=1}^n \left(\frac{\frac{P_i^t Q_i^t}{\sum_{i=1}^n P_i^t\ Q_i^t}}{\frac{P_i^t}{P_i^0}\ 100} \right)}$$

$$= \frac{1}{\sum_{i=1}^n \left(\frac{P_i^t Q_i^t}{\sum_{i=1}^n P_i^t\ Q_i^t} \cdot \frac{P_i^0}{P_i^t 100} \right)} = \frac{1}{\sum_{i=1}^n \left(\frac{Q_i^t}{\sum_{i=1}^n P_i^t\ Q_i^t} \cdot \frac{P_i^0}{100} \right)}$$

$$= \frac{1}{\frac{\sum_{i=1}^n P_i^0 Q_i^t}{\sum_{i=1}^n P_i^t Q_i^t . 100}} = \frac{\sum_{i=1}^n P_i^t Q_i^t .}{\sum_{i=1}^n P_i^0 Q_i^t} 100$$

the Paasche price index in current year (t) compared to base year (0) is:

$$PPI_{(t/0)} = \frac{\sum_{i=1}^n P_i^t Q_i^t}{\sum_{i=1}^n P_i^0 Q_i^t} 100$$

- **Paasche Quantity Index (PQI).**

in the same way we find the paasche quantity index in current year (t) compared to base year (0) is:

$$PQI_{(t/0)} = \frac{\sum_{i=1}^{n} Q_i^t P_i^t}{\sum_{i=1}^{n} Q_i^0 P_i^t} 100$$

2.3- Fisher index number.

The FISHER index number is a geometric average (mean) for the Laspeyres index and the Paasche index as follows:

$$FI_{(t/0)} = \sqrt{LI_{(t/0)} PI_{(t/0)}}$$

- **Fisher price index number:**

The Fisher price index number is:

$$FPI_{(t/0)} = \sqrt{LPI_{(t/0)} PPI_{(t/0)}} = \sqrt{\frac{\sum_{i=1}^{n} P_i^t Q_i^0}{\sum_{i=1}^{n} P_i^0 Q_i^0} \frac{\sum_{i=1}^{n} P_i^t Q_i^t}{\sum_{i=1}^{n} P_i^0 Q_i^t}} .100$$

- **Fisher quantity index number:**

The Fisher quantity index number is:

$$FQI_{(t/0)} = \sqrt{LQI_{(t/0)} PQI_{(t/0)}} = \sqrt{\frac{\sum_{i=1}^{n} Q_i^t P_i^0}{\sum_{i=1}^{n} Q_i^0 P_i^0} \frac{\sum_{i=1}^{n} Q_i^t P_i^t}{\sum_{i=1}^{n} Q_i^0 P_i^t}} .100$$

Example: we consider the previous example, calculate:

- The laspeyers price index in 2020 base 2019.
- The Paasche price index in 2020 base 2019.
- The laspeyers quantity index in 2020 base 2019.
- The Paasche quantity index in 2020 base 2019.
- The Fisher price index in 2020 base 2019.
- The Fisher quantity index in 2020 base 2019.

Solution:

comodity	2019		2020		$P_i^0 Q_i^0$	$P_i^t Q_i^0$	$P_i^t Q_i^t$	$P_i^0 Q_i^t$
	P_i^0	Q_i^0	P_i^t	Q_i^t				
A (Kgs)	20	10	22	15	200	220	330	300
B (Liters)	50	20	55	25	1000	1100	1375	1250
C (m)	25	40	30	50	1000	1200	1500	1250
total					2200	2520	3205	2800

- The laspeyers price index in 2020 base 2019:

$$LPI_{(t/0)} = \frac{\sum_{i=1}^{n} P_i^t Q_i^0}{\sum_{i=1}^{n} P_i^0 Q_i^0} 100 = \frac{2520}{2200}.100 = 114.45$$

Then, the level of prices of goods produced in this company, increased by 14.45% in 2020 compared to 2019.

- The Paasche price index in 2020 base 2019:

$$PPI_{(t/0)} = \frac{\sum_{i=1}^{n} P_i^t Q_i^t}{\sum_{i=1}^{n} P_i^0 Q_i^t} 100 = \frac{3205}{2800}.100 = 114.46$$

Then, the level of prices of goods produced in this company, increased by 14.46% in 2020 compared to 2019.

- The laspeyers quantity index in 2020 base 2019:

$$LQI_{(t/0)} = \frac{\sum_{i=1}^{n} Q_i^t P_i^0}{\sum_{i=1}^{n} Q_i^0 P_i^0} 100 = \frac{2800}{2200}.100 = 127.27$$

Then, the level of quantities of goods produced in this company, increased by 27.27% in 2020 compared to 2019.

- The Paasche quantity index in 2020 base 2019:

$$PQI_{(t/0)} = \frac{\sum_{i=1}^{n} Q_i^t P_i^t}{\sum_{i=1}^{n} Q_i^0 P_i^t} 100 = \frac{3205}{2520}.100 = 127.18$$

Then, the level of quantities of goods produced in this company, increased by 27.18% in 2020 compared to 2019.

- The Fisher price index in 2020 base 2019:

$$FPI_{(t/0)} = \sqrt{LPI_{(t/0)} PPI_{(t/0)}} = \sqrt{114.45 * 114.46}\,.100 = 114.45$$

Then, the level of prices of goods produced in this company, increased by 14.45% in 2020 compared to 2019.

- The Fisher quantity index in 2020 base 2019:

$$FQI_{(t/0)} = \sqrt{LQI_{(t/0)} PQI_{(t/0)}} = \sqrt{127.27 * 127.18}\,.100 = 127.22$$

Then, the level of quantities of goods produced in this company, increased by 27.22% in 2020 compared to 2019.

2.4- The purchasing power index number.

The purchasing power of money refers to the value of a currency in terms of its ability to buy goods and services. It is often measured as an index number, which is a statistical measure that attempts to make meaningful comparisons between different countries or different years.

Purchasing power refers to what a family can obtain with its income. It is therefore related both to income (wages) and to the cost of living (prices of goods and services). The cost of living for households is often measured by the Consumer Price Index (CPI), which is calculated using the prices of a basket of goods and services that represents the items typically consumed by families.

Then we can calculate the purchasing power index by the real wages index, as follows:

$$PPI_{(t/0)} = \frac{NWI_{(t/0)}}{CPI_{(t/0)}}.100$$

Where:

$PPI_{(t/0)}$: represents the purchasing power index in current year "t" compared to base year "0".

$NWI_{(t/0)}$: represents the nominal wages index in current year "t" compared to base year "0".

$CPI_{(t/0)}$: represents the consumer prices index in current year "t" compared to base year "0".

Example:

In 2009 a family distributed its consumer spending on 3 commodities: food, clothing and education in the following proportions respectively; 68% , 18% and 14%. The prices of these three commodities changed (%) during the time period 2010-2013 as follows:

	2010	2011	2012
Food	2.5%	3.5%	1.5%
Clothing	1.0%	2.0%	1.0%
Education	0.5%	1.0%	1.5%

1- Calculate the weighted total price index for all years, base 2009=100.
2- If you know that the wage (salary) of this family develops as follows:

year	2010	2011	2012
wage ($)	1200	1200	1350

- Calculate the purchasing power index for this family based on 2010=100.

Solution:

1- We know that the relationship between simple index ($I_{(t/0)}$) and growth ratio ($g_{(t/0)}$) is:

$$I_{(t/0)} = g_{(t/0)} + 100$$

So for every commodity, we have:

for " FOOD":

$$I_{(10/09)} = g_{(10/09)} + 100 = 2.5 + 100 = 102.5$$
$$I_{(11/10)} = g_{(11/10)} + 100 = 3.5 + 100 = 103.5$$
$$I_{(12/11)} = g_{(12/11)} + 100 = 1.5 + 100 = 101.5$$

for "CLOTHING":

$$I_{(10/09)} = g_{(10/09)} + 100 = 1.0 + 100 = 101.0$$
$$I_{(11/10)} = g_{(11/10)} + 100 = 2.0 + 100 = 102.0$$
$$I_{(12/11)} = g_{(12/11)} + 100 = 1.0 + 100 = 101.0$$

for "EDUCATION":

$$I_{(10/09)} = g_{(10/09)} + 100 = 0.5 + 100 = 100.5$$

$$I_{(11/10)} = g_{(11/10)} + 100 = 1.0 + 100 = 101.5$$
$$I_{(12/11)} = g_{(12/11)} + 100 = 1.5 + 100 = 101.5$$

- Calculating of prices index numbers of every commodities (base 2009=100).

-We use the circular property -

For " FOOD ":

$$I_{(09/09)} = 100.0$$
$$I_{(10/09)} = 102.5$$
$$I_{(11/09)} = I_{(11/10)}.I_{(10/09)}.\frac{1}{100} = 103.5 * 102.5 * \frac{1}{100} = 106.08$$
$$I_{(12/09)} = I_{(12/11)}.I_{(11/09)}.\frac{1}{100} = 101.5 * 106.08 * \frac{1}{100} = 107.67$$

For "CLOTHING":

$$I_{(09/09)} = 100.0$$
$$I_{(10/09)} = 101.0$$
$$I_{(11/09)} = I_{(11/10)}.I_{(10/09)}.\frac{1}{100} = 102.0 * 101.0 * \frac{1}{100} = 103.02$$
$$I_{(12/09)} = I_{(12/11)}.I_{(11/09)}.\frac{1}{100} = 101.0 * 103.02 * \frac{1}{100} = 104.05$$

For "EDUCATION":

$$I_{(09/09)} = 100.0$$
$$I_{(10/09)} = 100.5$$
$$I_{(11/09)} = I_{(11/10)}.I_{(10/09)}.\frac{1}{100} = 101.5 * 100.5 * \frac{1}{100} = 102.00$$
$$I_{(12/09)} = I_{(12/11)}.I_{(11/09)}.\frac{1}{100} = 101.5 * 102.0 * \frac{1}{100} = 103.53$$

Then the prices index numbers for every commodities (base 2009=100) are in the following table:

	2009	2010	2011	2012
Food	100	102.5	106.08	107.67
Clothing	100	101.0	103.02	104.05
Education	100	100.5	102.0	103.53

- Calculate the weighted total price index for all years, base 2009=100: in this case, we use the laspyers price index because we have a spending shares in base year (2009).

$$- \quad LPI_{(10/09)} = \sum_{i=1}^{3} IP^{i}_{(10/09)} W^{i}_{09} =$$

$$= 102.5 * 68\% + 101 * 18\% + 100.5 * 14\% = 101.95$$

$$- \quad LPI_{(11/09)} = \sum_{i=1}^{3} IP^{i}_{(11/09)} W^{i}_{09} =$$

$$= 106.08 * 68\% + 103.02 * 18\% + 102.0 * 14\% = 104.96$$

$$- \quad LPI_{(12/09)} = \sum_{i=1}^{3} IP^{i}_{(12/09)} W^{i}_{09} =$$

$$= 107.67 * 68\% + 104.05 * 18\% + 103.53 * 14\% = 106.44$$

years	2009	2010	2011	2012
total prices index numbers (2009=100) or $CPI_{(t/09)}$	100	101.95	104.96	106.44

2- Calculate the purchasing power index ($PPI_{(t/0)}$)for this family based on 2010=100.

We have:

$$PPI_{(t/0)} = \frac{NWI_{(t/0)}}{CPI_{(t/0)}} . 100$$

in the first, we calculate the wage index number $NWI_{(t/10)}$ (base 2010=100) where:

$$NWI_{(t/10)} = \frac{W_t}{W_{10}} . 100 = \frac{W_t}{1200} 100$$

so:

year	2010	2011	2012
wages (W_t)	1200	1200	1350
$NWI_{(t/10)}$	$\frac{1200}{1200}.100 = 100$	$\frac{1200}{1200}.100 = 100$	$\frac{1350}{1200}.100 = 112.5$

in the 2nd, we calculate the price index number $CPI_{(t/10)}$ (base 2010=100) where:

$$CPI_{(t/10)} = LPI_{(t/10)} = \frac{LPI_{(t/09)}}{LPI_{(10/09)}}$$

Then price index number base 2010:

years	2009	2010	2011	2012
total prices index numbers (2010=100) or $CPI_{(t/09)}$ ss	$\frac{100}{101.95}.100 = 98.08$	$\frac{101.95}{101.95}.100 = 100$	$\frac{104.96}{101.95}.100 = 102.95$	$\frac{106.44}{101.95}.100 = 104.4$

Finally, the purchasing power index numbers base year 2010 $PPI_{(t/10)}$, are:

$$CPI_{(t/10)} = \frac{NWI_{(t/10)}}{CPI_{(t/10)}}.100$$

year	2010	2011	2012
$PPI_{(t/10)}$	$\frac{100}{100}.100 = 100$	$\frac{100}{102.95}.100 = 97.13$	$\frac{112.5}{104.4}.100 = 107.75$

We note that the purchasing power of this family, decreased by 2.87% in 2011 and increased by 7.75% in 2012 compared to 2010.

Chapter 7:
Tow variables statistical analysis (Regression and Correlation)

Chapter 7: Tow variables statistical analysis

(Regression and Correlation)

The statistical researcher often finds himself needing to study two statistical variables at the same time and not being satisfied with studying each variable separately and independently. He also finds great importance in knowing whether there is a relationship between two variables or not and what type of this relationship it is. It is also worth noting that these variables can be qualitative as well as quantitative.

Sometimes we want to know, for example, whether there is a relationship between the score obtained in mathematics and the score obtained in history, or between a family's income and its consumption of goods and services. We can also study the relationship between gender and proficiency in the English language, or the relationship between gender and the average score in the baccalaureate exam

However, before that, we must know how to collect data and present it in a specific table to facilitate the process of exploiting it and extracting the most important statistical indicators that are useful in understanding the phenomenon under study.

1- Data collection and tabulation (contingency table).

Suppose that we want to conduct a statistical study of the variables of weight (X) and height (Y) on a sample of 12 students. The results in their original form were represented in the following table:

student number	01	02	03	04	05	06	07	08	09	10	11	12
weight (Kg)	55	60	55	65	75	65	75	65	65	75	80	80
height (Cm)	165	170	155	175	180	155	155	155	180	175	175	165

The question is how to compile these data into an appropriate table showing the values of the variables and their corresponding frequencies. The answer is shown in the common table as follows:

Y height / weight X	155	165	170	175	180	total
55	01	01	00	00	00	02
60	00	00	01	00	00	01
65	02	00	00	01	01	04
75	01	00	00	01	01	03
80	00	01	00	01	00	02
total	04	02	01	03	02	12

- In general: If we have a statistical distribution of two variables X and Y over n statistical individuals, such that n_{ij} represents the number of individuals who have the value x_i and the value y_j at the same time. The joint table is written as follows

Y / X	Y_1	Y_2	.....	Y_j	.....	Y_K	$n_{i.}$
X_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1K}	$n_{1.}$
X_2	n_{21}	n_{22}		n_{2j}		n_{2K}	$n_{2.}$
. . .	. . .	. . .	.....	. . .	.....	. . .	. . .
X_i	n_{i1}	n_{i2}		n_{ij}		n_{iK}	$n_{i.}$
. . .	. . .	. . .	.....	. . .	.....	. . .	. . .
X_L	n_{L1}	n_{L1}	.....	n_{Lj}	.....	n_{LK}	$n_{P.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	.....	$n_{.j}$	.....	$n_{.K}$	$n_{..}$

Where:

$$n_{i.} = \sum_{j=1}^{k} n_{ij} \quad and \quad n_{.j} = \sum_{i=1}^{L} n_{ij}.$$

$$n = \sum_{i=1}^{L} \sum_{j=1}^{k} n_{ij} = \sum^{k}_{j=1} n_{.j} = \sum_{i=1}^{L} n_{i.}$$

Note: The combined table can be expressed in relative frequencies f_{ij} instead of absolute frequencies n_{ij}.

where:

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

and

$$\sum_{i=1}^{L}\sum_{j=1}^{K} f_{ij} = 1$$

2- Marginal distribution and conditional distribution.

2.1- **Marginal distribution:**

The marginal distribution is taking each variable independently of the other variables. That is, we are interested in distributing the values x_i with the frequencies $n_{i.}$, or the marginal distribution of the variable X is the set of couples$(x_i ; n_{i.})$. This distribution can be represented in a table as follows:

x_i	x_1	…….	x_L	TOTAL
$n_{i.}$	$n_{1.}$	…….	$n_{L.}$	n

- Also, the marginal distribution of the variable Y is the set of couples $(y_j ; n_{.j})$

As measures of central tendency and measures of dispersion, they are calculated in the same way as in the case of a one-dimensional statistical variable.

2.2 - **Conditional distribution**:

We mean by the conditional distribution of a statistical variable, let it be X When the second statistical variable is given (Y= yj), is the set of couples $(x_j ; n_{ij})$. This distribution can be represented in a table as follows:

x_i	x_1	… x_i ….	x_L	TOTAL
n_{i1}	n_{11}	… n_{i1} ….	n_{L1}	$n_{i.}$

- **a)- Numerical features of conditional distributions**:

a.1)-**The conditional mean**:

- The conditional mean of the variable X with the condition Y=yj ,is the average value of the variable X for the statistical individuals who have the value y_j of the variable Y, and we symbolize it with the symbol $\overline{X_J}$ where:

$$\bar{X}_j = \frac{\sum_{i=1}^{L} n_{ij}\, x_i}{n_{.j}}$$

- The conditional mean of the variable Y with the condition $X=x_i$,is the average value of the variable Y for the statistical individuals who have the value x_i of the variable X, and we symbolize it with the symbol $\overline{X_J}$ where:

$$\bar{Y}_i = \frac{\sum_{j=1}^{k} n_{ij}\, y_j}{n_{i.}}$$

a.2)- **conditional variance**:

- The conditional variance of the variable X conditional on $Y=y_j$ is:

$$V_j(X) = \frac{\sum_{i=1}^{L} n_{ij}(x_i - \bar{X}_j)}{n_{.j.}}.$$

- The conditional variance of the variable Y conditional on $X=x_i$ is:

$$V_I(Y) = \frac{\sum_{J=1}^{K} n_{ij}(y_i - \bar{Y}_i)}{n_{i.}}$$

c)- **The relationship between the marginal and conditional mean:**

We have :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{l} n_{i.} x_i$$

and

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

then

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{l} \sum_{j=1}^{K} n_{ij}\, x_i$$

or

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{l} n_{ij}\, x_i$$

then:

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{k} n_{.j} \bar{X}_j$$

- Similarly for Y we find:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{L} n_{i.} \bar{Y}_i$$

Example:

Let the statistical distribution in an organization, according to wages and years of work be, as follows:

Wages (mu) / Years of employment	30	40	45	50
3	3	1	1	0
5	0	3	3	1
8	0	1	4	3

1. Find the marginal distribution of wages, and then calculate the marginal mean and marginal variance.

2. Find the marginal distribution of years of employment, and then calculate the marginal mean and marginal variance.
3. Find the conditional distribution of wages for workers with 5 years of employment, and then calculate the mean and variance.

Solution :

1- The marginal distribution of wages:

Wages (Xi)	**30**	**40**	**45**	**50**	**total**
Frequencies ($n_{i.}$)	**3**	**5**	**8**	**4**	**20**

- **Marginal mean:**

$$\bar{X} = \sum_{i=1}^{4} \frac{n_i x_i}{n} = \frac{3 * 30 + 5 * 40 + 8 * 45 + 4 * 50}{20} = 42.5\, mu$$

- **Marginal variance:**

$$V(X) = \sum_{i=1}^{4} \frac{n_i {x_i}^2}{n} - (\bar{X})^2$$

$$= \frac{3 * 900 + 5 * 1600 + 8 * 2025 + 4 * 2500}{20}$$

$$- (42.5)^2 = 38.75\ mu^2$$

- **Standar deviation :**

$$\sigma(X) = \sqrt{V(X)} = \sqrt{38.75} = 6.22\ mu$$

2- the marginal distribution of years of employment:

years of employment (Yj)	**3**	**5**	**8**	**total**
frequencies ($n_{.j}$)	**5**	**7**	**8**	**20**

- **Marginal mean:**

$$\bar{Y} = \sum_{j=1}^{3} \frac{n_{.j} y_j}{n} = \frac{5 * 3 + 7 * 5 + 8 * 8}{20} = 5.7\ year$$

- **Marginal variance:**

$$V(Y) = \sum_{j=1}^{3} \frac{n_{.j} {y_j}^2}{n} - (\bar{Y})^2 = \frac{5 * 9 + 7 * 25 + 8 * 64}{20} - (5.7)^2$$
$$= 4.11 \ year^2$$

- **Standard deviation :**

$$\sigma(Y) = \sqrt{V(Y)} = \sqrt{4.11} = 2.02 \ year$$

3- The conditional distribution of wages for workers with 5 years of employment.

Wages (Xi)	**30**	**40**	**45**	**50**	**total**
Frequencies ($n_{i2.}$)	**0**	**3**	**3**	**1**	**7**

- **Conditional mean of "X":**

$$\bar{X}_2 = \sum_{i=1}^{4} \frac{n_{i2} x_i}{n_{.2}} = \frac{0 * 30 + 3 * 40 + 3 * 45 + 1 * 50}{7} = 43.57 \ mu$$

- **Marginal variance:**

$$V_2(X) = \sum_{i=1}^{4} \frac{n_{i2} {x_i}^2}{n_{.2}} - (\overline{X_2})^2$$
$$= \frac{0 * 900 + 3 * 1600 + 3 * 2025 + 1 * 2500}{7}$$
$$- (43.57)^2 = 12.36 \ mu^2$$

- **Standard deviation :**

$$\sigma_2(X) = \sqrt{V_2(X)} = \sqrt{12.36} = 3.51 \ mu$$

3- Correlation and regression for two statistical variables.

In many statistical studies, we are interested in the extent to which there is a correlation or relationship between two or more variables and what is the nature of this relationship, whether it is linear or not, and whether it is positive or

inverse. For example, the relationship between the price of a certain commodity and the volume of demand for it, the relationship between income and the volume of consumption, the relationship between the quantities produced of a certain commodity and the number of workers in an production organization .... etc.

3.1- **Correlation coefficient.**

a)- **Rank correlation coefficient (SPEARMAN's coefficient) :**

It measures the degree of correlation between two statistical variables by relying on the ranks of the values of variable X and variable Y . It is calculated as follows:

- We order the observed values of X in ascending order and give each value x_i its corresponding order rg(xi).

- We order the observed values of Y in ascending order and give each value y_i its corresponding order rg(yi).

- We attach each value xi and yi for individual i with their corresponding rg(xi) and rg(yi) respectively.

- We calculate the difference between the orders of the two variables and let di=rg(xi)-rg(yi)

- Finally, we calculate Spearman's correlation coefficient as follows :

$$rs = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

If rs=0 there is no correlation between x and y.
If rs=+1 there is a strong positive correlation between x and y.
If rs=-1 there is a strong inverse correlation between x and y.

Example: In order to study the correlation between the number of years of smoking x and the percentage of retinal damage y, we collected the following data from a sample of patients at a hospital:

x	22	21	48	35	40	14
y	50	50	75	60	70	25

- Calculate Spearman's correlation coefficient.

Solution:

- Arrange the x values in ascending order and give the rank number for each value as follows:

Arrangement of x values	14	21	22	35	40	80
Rank number rg(x)	1	2	3	4	5	6

- We order the values of y in ascending order and give the rank number as follows:

Arrangement of y values	25	50	50	60	70	75
Rank number rg(y)	1	2.5	2.5	2.5	4 5	6

- Back to the original table:

x	22	21	48	35	40	14	total
y	50	50	75	60	70	25	
rg(x)	3	2	6	4	5	1	
rg(y)	2.5	2.5	6	4	5	1	
di=rg(x)-rg(y)	0.5	-0.5	0	0	0	0	
di^2	0.25	0.25	0	0	0	0	0.5

So:

$$rs = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6(0.5)}{6(6^2-1)} = 0.98$$

The correlation coefficient is close to one, so there is a strong positive correlation between the number of years of smoking and lungs damage.

Example:

A large number of university graduates applied for employment in a large company, recruitment is done after passing the written test and the oral test. After the candidates passed the written test, the company's management wanted to study the possibility of abandoning the oral test and just using the results of the written test(X). For this purpose, the company's research department drew a sample of only 08 candidates to take the oral test(Y) and compared its results with the results of the written test to see if the candidates' level remains approximately the same in both tests. The results of the eight candidates in both exams were as follows:

x : Excellen	poor	Good	Acceptable	Very Good	Excellent	Acceptable	VeryGood
y : Good	Good	VeryGood	Poor	Poor	Excellent	Good	Acceptable

- Based on the data, how does the studies department propose that the company should be satisfied with the written exam or should an oral exam be added?

Solution:

To answer the question, we calculate the correlation coefficient between x and y to see whether the results of the written exam and the oral exam are correlated or not. If the correlation is strong and positive, it means that the written test eliminates the need to take the oral exam because the results will remain almost the same. If the correlation is negative or zero, it means that the results of the written test are insufficient to judge the level of the candidates, which requires the oral test.

- Calculate Spearman's correlation coefficient:

total	vergood	acce	excell	verygood	acceptabl	good	poor	Excel	x
	accep	good	excell	poor	poor	verygood	good	good	y
	5.5	2.5	7.5	5.5	2.5	4	1	7.5	rg(x)
	3	5	8	1.5	1.5	7	5	5	rg(y)
0	2.5	-2.5	-0.5	4	1	-3	-4	2.5	di
61	6.25	6.25	0.25	16	1	9	16	6.25	di^2

So:

$$rs = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6(61)}{8(8^2-1)} = 0.27$$

We note that the correlation coefficient is small, which indicates that the results of the two tests are not correlated, which makes us suggest not only the written test but also the oral test.

B. linear correlation coefficient (Pearson's coefficient):

It is a measure of the degree of linear correlation between two quantitative variables X and Y and is calculated by the following relationship:

$$r = \frac{cov(X,Y)}{\sigma_X \,.\sigma_Y}$$

where:

cov(X,Y) represents the covariance between variables X and Y and is calculated as follows:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}(x_i - \bar{X})(y_j - \bar{Y}).$$
$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{K} n_{ij}(x_i y_j) - \bar{X}\ \bar{Y} \quad .$$

Further:

σ_x : represents the standard deviation of the variable X i.e. $\sigma_x = (V(X))^{1/2}$.

σ_Y : represents the standard deviation of the variable Y i.e. $\sigma_Y = (V(Y))^{1/2}$.

Properties:

- The correlation coefficient is between -1 and +1. ($-1 \leq r \leq +1$)

* If r = 0, there is no linear relationship between X and Y.

* If r = +1, there is a perfect linear relationship between X and Y.

* If r = -1, there is an inverse strict linear relationship between X and Y.

3.2- The Regression:

By regression, we mean the attempt to find a linear relationship between two statistical variables, the first variable is called the explanatory

(independent) variable and is symbolized by the X, and the second is called the dependent variable and is usually symbolized by the Y.

The linear relationship is of the form

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

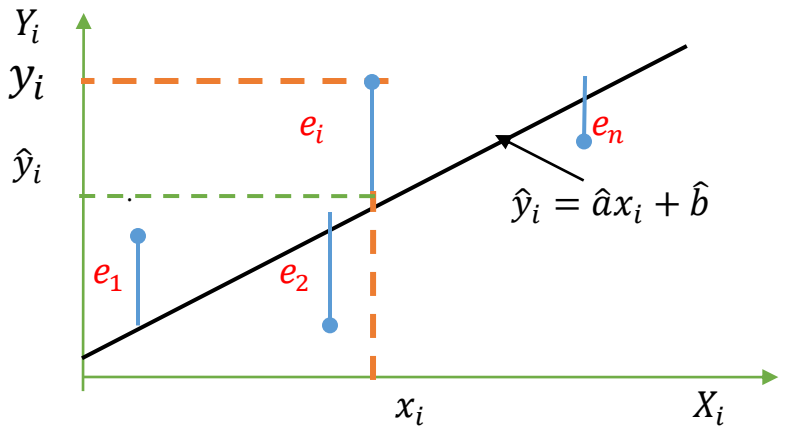
Where $\hat{Y}$ represents the estimated value of the actual value of Y resulting from the estimated linear relationship based on the actual data for both X and Y.

- **The Linear relationship estimation (ordinary least squares method):**

Let us have the actual data about the variables X and Y for a sample of statistical individuals of size n
represented by the pairs (x_i , y_i) where i=1...n. This data can be represented by a table as follows:

X	x_1	x_2	…… x_n
Y	y_1	y_2	…… y_n

If we were to graphically represent this data, we would have a set of points (Cloud of points), each point representing the couple (x_i , y_i) for individual number i as follows



If we want to represent this point cloud with a line that has an equation of the form aX+b The point cloud is the best representation to infer the true relationship between Y and X. We apply the least squares method as follows:

Let the estimated relationship be:

$$\hat{Y}_i = aX_i + b$$

The actual value of y_i is the estimated value of y_i plus the error caused by this estimate, which is e_i :

$$y_i = \hat{y}_i + e_i$$

$$y_i = ax_i + b + e_i$$
$$e_i = y_i - (ax_i + b)$$

The least squares method is all about minimizing the sum of the squares of the errors, i.e., we look for the line that makes the sum of the squares of the errors as small as possible.

The question is, what is the value of the parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ that minimize

$$\sum_{i=1}^{n} e_i^2$$

- The mathematical answer is as follows:

Our goal is:

$$Min\left[S = \sum_{i=1}^{n} e_i^2\right]$$

so

$$Min\left[\sum_{i=1}^{n} (y_i - ax_i - b)^2\right]$$

- The Necessary condition (first order condition):

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial a} = 0 \quad \ldots\ldots\ldots\ldots (1) \quad -$$
$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} = 0 \quad \ldots\ldots\ldots\ldots (2) \quad -$$

If we take equation (2), we find:

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial\, b} = 0 \quad \leftrightarrow \quad \frac{\partial \sum_{i=1}^{n} (y_i - a\, x_i - b)^2}{\partial\, b} = 0 \quad .$$

$$(-1)(2)\sum_{i=1}^{n}(y_i - a\, x_i - b) = 0$$

$$\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0$$

$$\boxed{b = \bar{Y} - a\, \bar{X}}$$

Substituting this result into equation (1), we find:

$$\frac{\partial \sum_{i=1}^{n}(y_i - a\, x_i - b)^2}{\partial\, b} = \frac{\partial \sum_{i=1}^{n}(y_i - a\, x_i - (\bar{Y} - a\, \bar{X}))^2}{\partial\, b} = 0$$

$$\left[\frac{\partial \sum_{i=1}^{n}\left((y_i - \bar{Y}) - a\,(x_i - \overline{X})\right)^2}{\partial\, b}\right] = 0$$

$$2\sum_{i=1}^{n}\left((y_i - \bar{Y}) - a\,(x_i - \overline{X})\right)(x_i - \bar{X}) = 0$$

$$\sum_{i=1}^{n}((y_i - \bar{Y})(x_i - \bar{X}) - a\,(x_i - \bar{X})^2) = 0 \quad \text{أ}$$

$$\boxed{a = \frac{\sum_{i=1}^{n}((y_i - \bar{Y})(x_i - \bar{X}))}{\sum_{i=1}^{n}((x_i - \bar{X})^2}}$$

Note that if we divide the numerator and denominator by the sample size n, we find*:*

$$\boxed{a = \frac{Cov\,(X,Y)}{V(X)}}$$

or*:*

$$\boxed{a = r\,\frac{\sigma_Y}{\sigma_X}}$$

Example:

Consider a statistical distribution of 20 students by math marks X and economics marks Y as follows:

y / x	4	6	8	12	Total	$x_i(\sum n_{ij}y_j)$
5	*2*	*0*	*1*	*1*	*4*	140
10	*1*	*2*	*0*	*3*	*6*	520
12	*2*	*3*	*0*	*2*	*7*	600
16	*0*	*2*	*1*	*0*	*3*	320
TOTAL	*5*	*7*	*2*	*6*	*20*	1580

1- Find the marginal distribution of x and then calculate $\bar{x}$ and v(x).
2- Find the marginal distribution of Y and then calculate $\bar{Y}$ and v(Y).
3- Calculate the conditional mean of the math marks for students with an mark of 12 in economics.
4- Calculate the linear correlation coefficient between X and Y.
5- Find the equation of the regression line y=àx+b by the least squares method.

Solution:

1- The Marginal distribution of X.

x_i	5	10	12	16	total
$\underline{\boldsymbol{n}}_{i.}$	4	6	7	3	20

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{l} n_{i.}x_i = \frac{1}{20}(5*4+10*6+12*7+16*3) = 10.6 \quad .$$

$$V(X) = \frac{\sum_{i=1}^{L} n_{i.}(x_i{}^2)}{n} - \bar{x}^2 = \frac{4*25+6*100+7*144+3*256}{20} - (10.6)^2$$

$$= 123 - 112.36 = 10.64$$

2- The Marginal distribution of Y:

y_j	4	6	8	12	total
$\underline{\boldsymbol{n}}_{.j.}$	5	7	2	6	20

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{l} n_{.j}y_j = \frac{1}{20}(4*5+6*7+8*2+12*6) = 7.5 \quad .$$

$$V(Y) = \frac{\sum_{j=1}^{k} n_{.j}(y_j{}^2)}{n} - \bar{y}^2 = \frac{5*16+7*36+2*64+6*144}{20} - (7.5)^2$$

$$= 66.2 - 56.25 = 9.95$$

3- The conditional mean of the math marks

$$\bar{X}_j = \frac{\sum_{i=1}^{L} n_{ij}\, x_i}{n_{.j}} \quad -$$

We have y=12 and j=4 then we have:

$$\bar{X}_4 = \frac{\sum_{i=1}^{L} n_{i4}\, x_i}{n_{.4}} = \frac{5*1+10*3+12*2+16*0}{6} = 9.83$$

4- The linear correlation coefficient between X and Y

$$r = \frac{cov(X,Y)}{\sigma_X \,.\sigma_Y} \qquad .$$

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{L} \sum_{j=1}^{K} n_{ij}(x_i y_j) - \bar{X}\ \bar{Y} \qquad .$$

$$cov(X,Y) = \frac{1}{n} \sum_{i=1}^{L} x_i \sum_{j=1}^{K} n_{ij} y_j - \bar{X}\ \bar{Y} \qquad . \qquad .$$

= 1580/20 - (10.6)(7.5)= - 0.5

then :

$$r = \frac{cov(X,Y)}{\sigma_X \,.\, \sigma_Y} = \frac{-0.5}{\sqrt{10.64}\sqrt{9.95}} = -0.048$$

The value of the correlation coefficient is very small, so there is no linear correlation between math rate and economics rate.

5- the equation of the regression:

$$a = r\,\frac{\sigma_Y}{\sigma_X} = (-0.048)\frac{\sqrt{9.95}}{\sqrt{10.64}} = -0.049$$

$$b = \bar{Y} - a\,\bar{X} = 7.5 - (-0.046)(10.6) = 7.98$$

then: $$Y = -0.049\,X + 7.98$$

Exercises.

Exercise 01:

In order to monitor the development of the level of activity in the food industries sector, the ministry of industry departments prepared a table showing the distribution of institutions of this sector in the year 2022 according to revenues as follows:

Revenues(10^9DZD)	2 – 8	8 - 10	10 – 18	18 – 28	28 – 36	total
Number of institutions (n_j)	36	18	22	16	8	100

1- What is the statistical population studied? – What is a statistical individual?

2- - What is the studied phenomenon (objective of this study) ?What is the variable used to measure it?- what type of this variable ? – What is the measurement scale for it?

3- Give the graphical representation of this distribution.

4- Calculate the arithmetic mean of this variable.

5- Calculate the Mode, interpret it ?

6- The authority of this sector wants to provide subsidy to 60% of institutions with low revenues. Determine the maximum revenue for the institutions concerned with this subsidy.

7-Measure the spread (dispersion) of this distribution.

(We **give** $\sum_{j=1}^{5} n_j x_j^2 = 23326$)

7- If we assume that the global tax T on revenues X , is given by the relationship : **T= 20000 + 0.02 X** . Calculate the value of the average global tax and its standard deviation.

Exercise 2.

In order to an have idea of the workers standard of living in productive companies, we prepared a table showing the distribution of workers according to wage categories in the industrial zone of Blida in the year 2019 which as follows:

Wage (thousand DZD)	60-70	70-80	80-90	90-100	100-110	110-120
Number of workers (n_j)	4	6	10	18	10	12

1- Identify each of the following: the studied statistical population, the studied characteristic, the variable used to measure it, and the appropriate measurement scale for it.

2- Find the following:

2.1. Calculate the average wage, the wage of the majority of workers, and the median. Conclude whether the distribution is symmetrical or not. Justify your answer.

2.2. If you knew that 10% of highly paid workers quit their job,

Calculate the maximum wage that the industrial zone will have to pay it.

2.3. Another industrial zone employs 80 workers with an average wage of 85 thousand DZD with a standard deviation of 20 thousand DZD:

a- Calculate the standard deviation of wages for workers in the industrial zone of Blida if you know that $\sum_{j=1}^{6} n_j x_j^2 = 554300$

b- Compare the two regions in terms of dispersion, which region offer higher wages, and which region offer lower wages than the other.

c- What is the average wage for workers in both regions together?

2.4. We assume that the tax revenue T on the wage (X) in the industrial zone in Blida is shown with the relationship **T=2000+0.02X.**

- Calculate the average value of the tax revenue and its standard deviation.

Exercise 3 .

We consider a discrete quantitative statistical variable X defined by the cumulative distribution function F(x) as follows:

$$F(x) = \begin{cases} 0 & ; \ x < 1 \\ 0.20 & ; \ 1 \leq x < 2 \\ 0.55 & ; \ 2 \leq x < 3 \\ 0.85 & ; \ 3 \leq x < 4 \\ 1 & ; \ x \geq 4 \end{cases}$$

- Represent graphically the cumulative distribution function (CDF)
- Calculate the mean, mode and median.

Exercise 4:

During a study of a statistical sample of size "n" (where n is an even number) on a discrete quantitative variable (X), it turns out that half of the sample members have the same value which is $x_1=a$; and the second half have the same value which is $x_2=b$.

1- Find in terms of "a" and "b" both the arithmetic mean and the median.

2- Find the variance expression V(x) in terms of the range E.

exercise 5.

We consider the following information about the annual development rate of the price of commodity « A » over three consecutive years as follows:

year	2020	2021	2022
rate (%)	5	8	4

1- Calculate the average growth rate of the commodity "A" during the three years.

2- Calculate the simple price index number for commodity "A" for the years 2021 and 2022 based on 2020 (2020=100).

3- If you know that the simple index number for the price of commodity "B" for the year 2021 based on 2020 is 115 , and the price index number for the year 2022 based on 2021 is 120 , and that the budget allocated for spending on the two commodities in the year 2022 was distributed equally between the two commodities. Calculate the price index number for the two commodities for the year 2022 based on 2020.

The price of commodity A increased by 6% in 2023 compared to 2022, and the price of commodity B decreased by 3% in 2023 compared to 2022. Calculate the price index number for the two commodities for the year 2023 based on 2020.

Exercise 6.

We consider the following information about the annual development of a worker's wage over three consecutive years:

Years	2020	2021	2022
Rate of development of the price of commodity A (%)	2	8	10
Rate of development of the price of commodity B (%)	4	-3	8
Wage development rate (%)	5	8	6

1. Calculate the average annual rate of development of the worker's wage during the three years.
2. If you know that in the year 2022, this worker has allocated 40% of his budget to spending on commodity A and 60% to spending on commodity B.
- Calculate the purchasing power index for this Worker for the years 2021 and 2022, based on 2020, and comment the results.

Exercise7.

A researcher wanted to explain the level of students in statistics by the effort expended in reviewing, so he took a sample of 8 students and studied the relationship between student's exam marks (variable Y) and the numbers of hours of review (variable X) based on the following data:

student(i)	1	2	3	4	5	6	7	8
X_i	3	5	1	2	4	1	6	3
Y_i	11	18	7	9	15	4	18	8

Where:

$$\sum x_i = 25 \quad , \quad \sum y_i = 90 \quad , \quad \sum (x_i - \bar{X})^2 = 22.875$$

$$\sum (y_i - \bar{Y})^2 = 191.5 \quad , \quad \sum (y_i - \bar{Y})\ (x_i - \bar{X}) = 62.75$$

1- Calculate and interpret the linear correlation coefficient (r).

2- Estimate the relationship **Y=a X + b** (by the OLS method)

3- Interpret the regression coefficient (a).

4- Find the relationship between the regression coefficient and the linear correlation coefficient.

. **Exercise 8.**

Firstly:

A researcher wanted to explain the level of grain production by the amount of rainfall in an agricultural area, so he recorded the amounts of grain production (Yi) and the amounts of rainfall (Xi) for the year (i) during the period 2000-2023, so you get the totals for all observations as follows:

$$\sum X_i = 480\,;\ \sum Y_i = 720.;\ \sum X_i Y_i = 16600\ ;\ \sum X_i^2 = 11544\ ;\ \sum Y_i^2 = 24504$$

1- Calculate the linear correlation coefficient between the two variables. Comment

2- If we assume that the quantities of grain production are linearly related to rainfall amounts, estimate this relationship based on the ordinary least squares method.

3- Give an explanation of the model parameters.

Secondly:

It was revealed from a study of the two quantitative statistical variables X and Y on a sample which number of individuals is equal to $\boldsymbol{n}$ .The data related to the individual (i) represented by the pair $(x_i\,, y_i\,)$ is defined as follows:

$$\begin{cases} x_i = i & \forall i = 1 \ldots n \\ y_i = 3i + 8 & \forall i = 1 \ldots n \end{cases}$$

- Calculate the covariance Cov(X,Y) in terms of n.
- Deduce the linear correlation coefficient between X and Y.

*- **We note that**:

$$\sum_{k=1}^{m} k = \frac{m(m+1)}{2} \quad , \quad \sum_{k=1}^{m} k^2 = \frac{m(m+1)(2m+1)}{6}$$

References.

1- Fabrice Mazeroue, “ Statistique descriptive” , 2005
2- J.J.Droesbeke, « Elements de Statistique »., 2015
3- K.M.Ramachandran, Chris.P, “Mathematical Statistics with Applications”., 2009
4- L.Leboucher, MJ.Voisin, « Introduction a la Statistique Descriptive. Cours et exercices avec tableur. »France 2011.
5- M.Jmbu, “ Exploration informatique et statistique des Données”, Dunod, 1989
6- Prems MANN , Christophen Jay Lake, “ Introductory Statistics”, 2010
7- Saporta.G, « Theories et méthode de la Statistique », Tschnip, 1978.
8- Stephen Bernstein, « Theory and Problems of Elementary of Statistics”., 1999
9- Z.Holcomb, « Fundamentals of Descriptive Statistics ». 1st edition, usa, 1998.

10- شرف الدين خليل، " الإحصاء الوصفي" . شبكة الأبحاث والدراسات الاقتصادية WWW.RR4EE.net.

11- محمد مفيد القوصي، " الإحصاء الوصفي والاستدلالي " . مركز الكتاب الاكاديمي. دمشق.2011