People's Democratic Republic of Algeria

The Ministry of higher education and scientific research

Mohammed ben Ahmed University Oran 02

Faculty of Social Sciences

Department of demography

A pedagogical publication for first-year students

Social Sciences

Scale of the first Hexagon

Descriptive statistics

Prepared by Prof. Dr

BOUDIA Leila

Approval of the scientific committee of the department	Approval of the Scientific Council of the faculty
جامعة وهران 2 محمد بن أحمد اللجنة العلمية قسم الديموغرافيا د. بن رشدي خضرة رئيس اللجنة العلمية لقسم [illegible] العلوم الاجتماعية	الأستاذ الدكتور : زاوي عمر رئيس المجلس العلمي لكلية العلوم الإجتماعية جامعة وهران 2 محمد بن أحمد المجلس العلمي كلية العلوم الاجتماعية

Academic year 2023-2024

People's Democratic Republic of Algeria

The Ministry of higher education and scientific research

Mohammed ben Ahmed University Oran 02

Faculty of Social Sciences

Department of demography

A pedagogical publication for first-year students

Social Sciences

Scale of the first Hexagon

Descriptive statistics

Prepared by Prof. Dr

BOUDIA Leila

Academic year: 2022 2023

Table of contents:

Introduction

Chapter 01:

The main concepts of statistics

1. Background of statistics
2. Concept of statistics
3. Statistical terms
4. Statistical variables
5. Sources of statistical data
6. Methods of statistical data collection
7. Types of samples and sampling methods
8. Types and steps of the statistical researches

Chapter 02:

Presentation of the statistical data:

1. Table presentation of the statistical data

1.1 Definition of the table presentation

1.2 Types of statistical tables

1.2.1 Simple tables of the frequency distribution

1.2.2 Two-way tables of the frequency distribution

2. Graphic representation of the frequency distributions.

2.1 Graphic representation in the case of a discrete quantitative variable.

2.2 Graphic representation in the case of a continuous quantitative variable.

2.3 Graphic representation in the case of a qualitative variable.

Activities of chapter 02.

Chapter 03:

Measures of the central tendency:

1. Concept of the central tendency

2. Measuring the central tendency

2.1 The arithmetic mean

2.2 The geometric mean

2.3 The harmonic mean

2.4 The quadratic mean

2.5 The median

2.6 The mode

3. The relation between the measures of the central tendency

Activities of chapter 03

Chapter 04:

Measures of dispersion

1. Definition of dispersion
2. Measuring data dispersion
2.1 The general range
2.2 Interquartile range
2.3 Mean deviation
2.4 Variation and standard deviation.
2.5 Measures of the relative dispersion
Activities of chapter 04

Introduction:

Statistics is one of the necessary sciences for any scientific research, experiment, theory, or empirical study that aims at reaching objective and valid results. Its importance lies within the fact that it is a tool used by people in their daily life, mainly by decision makers in administrations and companies because any economic or administrative decision must be built on exact and real data to reach exact and objective results about the phenomenon under study and analysis.

Statistics got scientific and academic interest in the world institutes and universities, and in all the fields of study. The descriptive statistics is the first introduction to statistics in general. It is taught to students because it relies on introductions on how to describe the economic, social, natural, and other phenomena using descriptive digital methods in the form of numbers and graphics.

In order to help students master this module, we sought to set handouts about the descriptive analysis to help overcome the difficulties they face and better understand this module. This handout is a series of lectures in the descriptive statistics for the students of 1st year LMD in economics, commercial sciences, and management. The content is provided in the form of brief and simple lessons supported by many examples, activities, and solutions.

In order to cover what must be studied in the descriptive statistics by the students of economics, we divided the content of this handout into 06 chapters. The first chapter includes the main concepts of statistics, the 2nd includes how to present, summarize, and present data in frequency tables and graphics, the 3rd covers the measures of the central tendency, the 4th revolves around the dispersion, the 5th is about the measures of the shape, while the latter is about the index numbers.

Chapter 01:

The main concepts of statistics

1. Background of statistics
2. Concept of statistics
3. Statistical terms
4. Statistical variables
5. Sources of statistical data
6. Methods of statistical data collection
7. Types of samples and sampling methods
8. Types and steps of the statistical researches

1. Background of statistics

Statistics emerged in the Medieval Ages because of the interest of the states in making a census about the number of the individuals to be able to form strong armies that can defend the state against the expansionist greed of the other states. Besides, it was the result of the interest of the states in knowing the wealth of the individuals to impose taxes and collect the necessary money to fund the army and manage the state. Then, the census expanded to include data about births, deaths, production, consumption, etc. After that, there was a need to organize and summarize the obtained data in tables to be easily read. These methods were called the science of the state, the science of the kings, and statistics, respectively.

Statistics is a Latin word derived from “state”. It was believed in the beginning that statistics is about the scientific methods for data organization and presentation in graphics or tables. Nevertheless, with the development of the probabilities in the 17th and 18th centuries, there was an increase in the use of the statistical analysis to reach beneficial data in decision-making, prediction, estimations, and deductions from a set of variables that is bigger than the set of variables that had been really observed.

2. Concept of statistics

One of the common concepts among people is that statistics is a set of numbers and data, such as the number of population, births, deaths, etc. Therefore, people believe that statistics is counting and expressing things in numbers. This is, in fact, a limited vision of statics. Here are some definitions of this concept:

- It is the science that investigates the suitable scientific methods and methods for data collection, tabulation, organization, analysis, and interpretation to reach the necessary results for increasing knowledge, decision-making, and generalization.

- It is the science that studies the various tools and methods of collecting quantitative data about economic, social, and other phenomena. Besides, it divides, analysis, interprets, and presents data in suitable graphics to facilitate the decision-making on a good basis.
- It is a science that investigates the method of collecting facts about the scientific and social phenomena that manifest in various cases or views. Besides, it investigates how to record these data in digital and metric ways and summarize them in a way that facilitates knowing the orientations and interrelations of the phenomena. In addition, it focuses on the study of these relations, orientations, and their use in understanding the reality of the phenomena and knowing the laws they follow.

Due to the confusion between the statistical data and the census, we present these definitions to help distinguish them:

- **The census**:

It is a process by specialized bodies to get data about one or more phenomena. Thus, it is the quantitative study of the phenomena.

- **The statistical data:**

This refers to a set of quantitative (digital) and descriptive information about the phenomenon under study. The information are collected by specialized bodies and are, then, presented in scientific methods in official and unofficial documents for a specific purpose.

Based on what was aid, we can divide statistics into two types:

- The descriptive statistics:

It is about the methods of collecting, dividing, and presenting the data. It summarizes data in one or more information in tables or graphics, and counts some statistical measures (means, percentages…).

- The inferential statistics:

It is related to inferring and taking the suitable decisions to the phenomenon under study, and calculating the degree of trust of the decisions and inferences. In this context, the rule of the part is applied on the whole.

3. The statistical terms:

3.1 Population:

It is a set of items or individuals focused on in a specific study, or it is a set of views and measures of a set of statistical units of a measurable phenomenon such as the students, the families, the companies, etc. The population can be divided into two types:

- The target population: It is a group of people or items targeted in a study to generalize the results on them, such as the university students, public officials in a given state, etc.

- The study population: It is a set of individuals from whom we get information and data about the study phenomenon or problem.

Generally, the population is big and, therefore, the study of all its items may be difficult. Therefore, we study part of the population that is called "the sample".

3.2 The sample:

It is a partial group of the items of the population. The items of the sample are generally chosen on the basis that all the items get equal chances in order to have a good representation of the population. The size of the sample differs according to the study and the material and human resources available for the study. In this context, sampling is one of the methods used in most of the field studies because it is impossible to collect statistical data from the units that represent the studied population.

3.3 The statistical unit:

It is the item on which the statistical study or sampling is carried out. The unit must be clearly and exactly defined.

3.4 The statistical phenomenon:

It is the quality of items that differ from one to another in shape, type, and quantity. This quality is called the variable. Examples include the height, the age, the weight, the production, the savings, the investment, the consumption, etc.

3.5 The variables:

They refer to the features and characteristics of the items of the sample. These features differ from one item to another, such as the heights of children of a certain age in a certain city, the price of a given product in a given market, etc.

3.6 The parameter:

It is the measure that describes some characteristics of the population. We get it from the analysis of the population data and the complete census. For instance, the average income of an individual in a given state is a parameter because it reflects the living standard of the inhabitants of that state. Besides, the parameter is referred to with one of the Latin letters; for example, the statistical population mean is referred to as μ and its variation as $\boldsymbol{\sigma}^2$.

4. The statistical parameter:

It is that quality or quantity that may change from one individual to another, or from one view to another, and allows distinguishing and classifying them. Its value is known as the value of the statistical variable. The statistical variable can be divided into:

4.1 The qualitative variable:

It is the variable or phenomenon that cannot be measured with digits because we can only measure its frequency. It is a set of non-numerical qualities and types that can be divided into:

- **Orderable qualitative data:** They can be ordered from up to bottom or from bottom to up, such as the economic growth levels, the educational level, the military grades, the success estimates, etc.
- **Unorderable qualitative data:** such as the nationality, types of diseases, familial status, etc.

4.2 Quantitative variables:

They are the variables that can be expressed and measured in numbers. They are the most spread variables because the statistics language is s language of numbers. Examples include the production, consumption, weight, height, investment, etc. These variables can be divided into:

- **Discrete variables:** They are the variables that take valid undividable values such as the number of children in a family, the number of students in the various educational levels, the number of rooms in a house, the number of goods produced, etc.
- **Continuous variables:** They are the variables that take the most possible values of the study. Due to the infinite number of these values, we divide the interval of the study into subintervals known as the classes. Generally speaking, the variable is continuous if it is related to time (speed, age, etc), mass ((weight, density, etc), money (incomes, wage, price, etc), or space (length, surface, etc). We refer to this parameter with X_i and to the corresponding value with x_i.

5. Sources of the statistical data:

The sources refer to where the statistician brings the study data from. Researchers rely on two main sources to get statistical data:

5.1 The direct sources:

The statistician collects his data through direct contact with the units of the statistical population, or through relying on the documents that contain raw information. In this case, the questionnaire and the interview are important sources to get information from primary sources.

5.2 The indirect sources:

The researcher gets the statistical information from the previous studies. These data are classified by previous researches or official or unofficial bodies. In addition, the data are published in special statements or periodicals, or may be maintained in the traditional or electronic archives.

6. Methods of collecting the statistical data:

When conducting a statistical study on a given phenomenon, we need to collect information and data about the units of the study population. These data are collected through:

6.1 The complete census:

Here, the data are collected from all the items of the population such as in the population census. This method is one of the best because it gives full and exact information about the problem if the requirements of the scientific research are available. Nevertheless, it requires high costs because it needs big material and human tools and long time. In addition, the bigger the population is, the higher the mistake potential is. This method is used if we need highly exact data such as in the population census, in agriculture, and in economy. Therefore, it is generally used by the governments.

6.2 The sampling:

It is used when it is difficult to conduct the study on all the population items. Therefore, we can use the information related to only part of the population rather than the information about the whole population. A part (sample) of the items of the study population is chosen based on a good scientific method. Then, with the statistical analysis of the data of the sample, we can generalize the results on the whole population. In this context, we must point that the bigger the sample is, the better and more valid the results are.

The causes of using the sampling are:

- Saving time and effort.
- The difficulty of making the complete census due to the nature of the society, which may be limited, very large, made up of precious or dangerous units, etc.

The sampling aims at reaching conclusions about the population. The 1st step of choosing the sample from the population is determining its size, looking for the frame of sampling from which we shall choose the sample, and, then, using one of the sampling methods.

7. Types of samples and sampling methods:

There are many types of samples. The researcher can choose the sample based on the circumstances of the study. Generally speaking, the samples can be divided into:

7.1 The probability samples: They are samples drawn from the population in a way that each item gets the same chance as the others. Therefore, this sample represents well its population, accepts the statistical analysis methods, and has generalizable results. The probability samples are divided into:

- The simple random sample: It is the sample drawn from the population in a way that all the items get equal chances. This type is beneficial and influential in case of the existence of harmony and common traits between the items of the study population. For example, a population of average employees represents a harmonious population regarding the incomes. To get a random sample, we generally use the draw or the random numbers table.

- The stratified random sample: This type is used with the inharmonious populations that are made up of many classes based on various considerations, such as the income, the expenditures, the education, etc. In this case, we divide the population into harmonious classes and determine the ratio of each population. Thus, the sizes of the classes are referred to with N_i...............N_1, N_2, etc, where $_i$ is the number of the classes of the population.

In order to draw a stratified sample, we:

1. Determine the ratio of each class for the population **Ni/N** .
2. Determine the size of the sample we want to draw ***n***.
3. Determine the number of the statistical units we must draw from each class: $\boldsymbol{n_i}$ according to the ratios determined in (1), where $\boldsymbol{n_i = n * \frac{Ni}{N}}$.
4. Draw $\boldsymbol{n_i}$ from $\boldsymbol{N_i}$ at random using the random numbers table. Then, we put all the drawn units together to form a stratified random sample.

- The systematic random sample: It is used to test a sample whose items are known or limited. In this case, we determine the group that we shall use to choose the sample. We start with choosing a random umber from the 1st group and, then, we add the size of the group we determined to the number we chose. For example, if we want to choose a sample of 100 people from a population of 2000 people, we start by dividing the number of the population by the size of the sample $\frac{2000}{100}$. Thus, we get the size of the group, which equals 20 individuals. Then, we choose the first number at random (09 for example) and add the size of the group to it (20) each time until we get 100 numbers. Thus, the chosen numbers will be (9, 29, 49, 69,…).

- **The cluster random sample:** It is a sample taken for the necessity more than for the choice. The statistical sample is divided into clear sub-groups known as the clusters. We draw a simple random sample from the clusters. For example, when we study the family budget, we divide each district to a group of buildings. Thus, we get a list of groups of buildings. These groups are the items of the population, and each group makes a cluster. Then, we choose the group to be studied at random.

- **The multi-stage random sample:** This method is used when we cannot directly reach all the items of the study society, and when it is difficult to make a sampling frame that includes all the items of the population. Hence, it is not necessary to get a sampling frame for all the items of the society, mainly in the last phase. This type of sampling is used in the agricultural statistics in general.

7.2 The non-probability samples:

There are many methods of choosing the non-probability samples based on the trends of the researchers. However, we shall discuss only two:

- **The accidental sample:** Its choice depends on the pure chance. It is characterized with saving time and costs, and allows getting reliable information if the study population is highly harmonious. On the other hand, if the population is not harmonious, there will be bias in choosing the items of the sample.

- **The quota sample:** The researcher, here, determines the quota of each group or stratum of the population. Then, he chooses the items of the sample by accident. This method reduces the probable bias, and is beneficial in case there are no sampling frames for the population strata. Nevertheless, it includes risks of bias when there is balance between the quotas of the strata of the sample.

8. Types and steps of statistical researches:

The statistical researches are divided into three classes:

8.1 Descriptive researches:

They aim at collecting data about a specific phenomenon, not for a given purpose, but for providing data that may be used in the future, such as the census of the population, the agricultural and industrial census, etc.

8.2. The analytical statistical researches:

They aim at collecting information for a given purpose, which help interpret a given problem noticed by the researcher.

8.3 The experimental statistical researches: This type of researches is used in different fields such as medicine, agriculture, and socio-economic aspects.

The methods and steps of the research differ from one field to another. However, they have common points. Thus, when conducting a statistical study, we must respect the following points:

1. Identifying the study phenomenon:

We identify the general frame of the study phenomenon that includes the aim, the population, the suitable time and space for data collection, the qualities that must be known, and the measurement units.

2. Collecting the statistical data:

It is one of the main points of the statistical work because the availability of the exact and good data about the study phenomenon gives reliable results and helps take good decisions based on these results. Besides, the researcher identifies the sources and the methods of data collection based on the aim of the study.

3. Dividing and presenting the data:

After the success in collecting the statistical data that may be big, we do not get a clear idea about the results. Therefore, the researcher divides and classifies the data by putting them in harmonious groups that have one or more common traits. Then, he presents these data in suitable tables or graphs so that the information becomes practical for the researcher per se, or for other researchers when published in special statements or general periodicals.

4. Analyzing the data and reading the results:

Data analysis is very important for any statistical research because it answers the problematic of the research. Therefore, the researcher makes the statistical analysis of the study phenomenon through the suitable statistical tools that help analyze data to get the study results, read their connotations, and take decisions.

Chapter 02:

Presentation of the statistical data:

1. Table presentation of the statistical data

1.1 Definition of the table presentation

1.2 Types of statistical tables

1.2.1 Simple tables of the frequency distribution

1.2.2 Two-way tables of the frequency distribution.

2. Graphic representation of the frequency distributions.

2.1 Graphic representation in the case of a discrete quantitative variable.

2.2 Graphic representation in the case of a continuous quantitative variable.

2.3 Graphic representation in the case of a qualitative variable.

Activities of chapter 02.

Preamble:

After the success of collecting the statistical data, we find a big set of unorganized facts. In this context, we cannot understand these facts or deduce any results because they are not organized. Therefore, it is necessary to organize them to facilitate their study. Thus, we classify and divide them into harmonious groups, and put them in tables. This division depends on the nature of the data and the aim of the research.

1. Table presentation of the statistical data:

1.1 Definition of the table presentation:

It refers to putting the primary data that we collected in final tables of two columns. The 1st shows the values of the phenomenon or the study variable. These values take the form of qualities, point values, or intervals (classes). The 2nd shows the frequencies of these qualities, values, or intervals.

1.2 Types of the table presentations:

The table presentations differ according to the type and aim of the study. The most important ones are:

1.2.1 The simple frequency distribution tables:

These tables represent the method of organizing the raw data about the phenomenon (the variable) and dividing them in tables that include qualities or values of the phenomenon, and their corresponding frequencies in order to study and analyze them. This type of tables is used to describe and summarize data related to one phenomenon, be they qualitative or quantitative. This table is an example

Table 02-01: the general shape of the simple frequency distribution table:

We write in these columns the number of items that correspond to each quality, value, or class of the statistical variable

Variable X_i	Frequency n_i
X_1	N_1
X_2	N_2
.	.
.	.
.	.
*X*k	n_k

In case we have a qualitative variable, we put the quality of the variable in these columns. If we have a discrete quantitative variable, we write the values of the variable from up to bottom, or vice versa. If we have a continuous quantitative variable, we write the values of the form of categories

The method of table presentation differs according to the type of variables. Thus, we find the following cases:

-Data of the qualitative variables: They are variables whose data are types and qualities, not numbers. To make a frequency distribution table for the qualitative data, we need a table of three columns. The 1st is for the qualities after ordering (if they are orderable), the 2nd is for inputting data, and the 3rd is for the frequencies. This example illustrates what we said:

Example (2-01): The following data show the blood types of 20 patients who underwent medical surgeries in a given week:

O, AB , O , B , A , B , O , A , B , O , A , O , A , B , O , B , O , O , AB , A.

Task: present them in a frequency distribution table?

Solution:

Table 2-02: distribution of the patients according to the blood type (qualitative variable)

Blood type ($X_{i)}$	**Grades**	**Number of patients (frequencies n_i)**
A	/////	**5**
B	/////	**5**
AB	//	**2**
O	///////	**8**
Total∑	**-**	**20**

Putting the data in this way makes it very clear and easy to know information that were not clear when they were raw. For example, it is easy to know the number of patients who have the same blood type, and to know the dominating type.

- Data of the discrete quantitative variables: They are the data that take the form of numbers, such as the number of university students, employees, etc. In order to divide and classify the discrete data, we classify them in similar groups and, then, put them in tables of three columns. The 1st is for the values of the phenomenon (the variables) after ordering them, the 2nd is for inputting the data, and the 3rd is for the frequencies. The following table is an example:

Example (2-02): The following data represent the number of individuals in a sample of 30 families:

2	**3**	**2**	**4**	**5**	**2**	**4**	**4**	**5**	**2**	**2**	**4**	**3**	**4**	**5**
3	**5**	**4**	**3**	**4**	**5**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**	**3**

Task: Put the data in a frequency distribution table?

Solution:

Table 2-02: distribution of the families according to the number of individuals (discrete quantitative variable)

Size of the family (X_i)	**Grades**	**Number of families (frequencies n_i)**
2	/////	**5**
3	/////	**7**
4	//	**10**
5	///////	**8**
Total∑	**-**	**30**

- Data of the continuous quantitative variables: They are the most used variables. The items may take the form of numbers of fractures. When studying a continuous quantitative variable, the study interval includes infinite values. Because of the inability to put all these values, we divide the interval into sub-intervals called classes. In this line, the number of the classes is identified based on the size of the sample and the distribution of the statistical units on the study interval. To make a frequency distribution data for a continuous quantitative variable, we follow these steps:

1. **Identify the range:** the range is the interval where the data are spread. It is the difference between the maximum and minimum values.

The range= the maximum value- the minimum value

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

2. **Identifying the number of the classes:** The number of classes required to make a frequency distribution table is identified using some mathematical equations as follows:

-Staugres equation:

$$K= 1 + 3.322 \log (n)$$

Where

K: is the number of the classes

n: is the number of the values

$$K= 2.5 \sqrt[4]{n}$$

3. **Identifying the length of the class:**

$L=\frac{R}{K}$ $\qquad$ L ength= $\frac{\text{Range}}{\text{Number of classes}}$

ng

The length of the class * the number of the classes ≥ the range

4. **Identifying the limits of the classes:** In this phase, we identify the start and end of each class. The start of the 1st class must be lower than, or

equal to, the minimum value in the data. In addition, the end of the last class must be higher than the maximum value in the data.

5. **Identifying the number of values or views:** The identification of the number of values or views in each class requires that each value have one class. This is called frequency (***n***$_i$). Moreover, we must make sure the number of frequencies equals that of the values.
6. **Identifying the centers of the classes:** When making frequency distribution tables including classes, the original statistical values of the items are lost and we no more know about them only that they belong to a certain class with two known limits. To overcome this problem, we extract the center of the class that refers to the middle of the class. We get it through:

$$\text{The center of the class}=\frac{\text{The minimum limit of the class + the maximum limit of the class}}{2}$$

$$C_i = \frac{Li + L\,i{+}1}{2}$$

Where L_i is the minimum limit of the class $_I$ and L_{i+1} is the maximum limit of the class $_i$

Example 2-03: The researcher wants to know the distribution of the degrees of the salaries (10^3 DA) that the employees of a given company get. He collected data about 42 employees as shown in the table:

66	52	56	40	55	23	26
49	42	38	44	51	64	14
26	30	63	67	16	42	38
14	15	53	35	42	**67**	45
60	56	57	50	45	50	40
24	46	53	39	35	52	49

Task: clarify the basic landmarks of these data in a frequency distribution table?

Solution: Despite that the number of values does not exceed 42 views, it is difficult to have a clear and rapid idea about these values. Therefore, we must order them in classes (the study variable: the monthly salaries that are a continuous quantitative variable) and, then, put them in a frequency distribution table that includes the classes and the frequency of the individuals in each class. Therefore, we follow these steps:

1. **Identifying the range**: The previous data show that the minimal salary is 14 x 10^3 DA while the maximum is 67 x 10^3. Thus, the range is:

$$R = X_{\max} - X_{\min} = 67 - 14 = 53$$

2. **Identifying the number of the classes:** The use of a low number of classes facilities the calculations and exactness. However, the increase of the number of the classes leads to a high number of calculations and raises the exactness. The number of the classes is identified by the circumstances of the study phenomenon and the attitude of the researcher. Generally speaking, it is better that the number of the classes be between 5 and 15. In addition, because of the differences in identifying the number of the classes, it is necessary to use one of the agreed upon equations that help identify the number of the classes that is related to the number of the views. In order to identify the number of classes in our example, we shall rely on Staurges equation that is the most used:
   The number of the values n= 42
   K= 1 + 3.322 log (n) = 1+ 3.222 log (42) = 1+ 3.322 (1.623)= 6.39 = 6.
   The number of the classes is K= 6.
3. **Identifying the length of the class:**
   $$L=\frac{R}{K} = \frac{53}{6} 8..83 =9$$
   Where:
   R: is the range.
   L: is the length of the class.
   K: is the number of the classes.
   Thus, the length of the class is L= 9.
   When identifying the length of the class, we must follows this inequality.

The length of the class * the number of the classes ≥ the range

Thus:

$$53 < 54 = 6 * 9$$

4. **Identifying the limits of the classes:** The class starts with the minimum value in the data and then we add to it the length of the class to identify its end and the beginning of the next class. Thus:
   - **1st class:** The minimum limit of the 1st class is 14. The maximum limit of the 1st class is 14 +9 = 23. Thus, the 1st class is]23- 14].
   - **2nd class:** The minimum limit of the 2nd class is23 The maximum limit of the 2nd class is 23 +9 = 32. Thus, the 2nd class is]32- 23].

   We carry on this way until making the 06 classes.
5. **Tabulation:** It is the process of inputting data in the frequency distribution table, taking into account that each value should have only one class and that the total of the frequencies equals the number of the values.

Salaries class X_i	Data input	Number of employees (frequency) n_i	Center of the class C_i
]23- 14]	////	4	18.5
]32- 23]	//////	5	27.5
]41- 32]	// /////	7	36.5
]50- 41]	// /////	7	45.5
]59- 50]	/ ///// //////	11	54.5
]68- 59]	//// ///////	8	63.5
	-	42	-

Inputting data in a frequency distribution table enables us to understand the facts and make deductions that we cannot make from the absolute data. After making the frequency distribution table, we should present data in a relative frequency distribution form to express the relative importance of each class for the total frequencies. In this context, the relative frequency is calculated as follows:

The relative frequency=$\frac{\text{The frequency of the quality (variable or class)}}{\text{The total frequencies}}$

$$fi=\frac{Ni}{\sum ni}$$

Knowing that the total frequencies equals the number of values or views ∑fi=1 and ∑ni=N

We can transform the relative frequency into a percentage relative frequency by multiplying it by 100. The percentage relative frequency is calculated as follows:

$$f_i\,\% = f_i \times 100$$

Knowing that: $\sum f_i\,\% = 100\%$

The relative frequency helps reduce the chart when the number of values is big. On the other hand, it helps show the chart when the number of the values is small. Moreover, we may need additional information about the data. For example, we may need to know the items whose value is less or more than a given limit. These information are obtained by finding the ascending and descending cumulative frequencies:

- **The ascending cumulative frequency:** It represents the number of individuals whose statistical value is less than the maximum limit of the corresponding class.

- **The descending cumulative frequency:** It represents the number of the individuals whose statistical value exceeds the minimum limit of the corresponding class.

Using the data of the previous example (2-03), we shall find the relative frequencies, the percentage relative frequencies, and ascending and descending cumulative frequencies.

Table 2-05: the relative frequencies, the percentage relative frequencies, and the ascending and descending cumulative frequencies

Salaries class X_i	frequency n_i	Relative frequency fi	Percentage relative frequency fi%	Ascending cumulative frequencies n_i	Descending cumulative frequencies n_i
]23- 14]	4	0.095	9.5	4	42
]32- 23]	5	0.12	12	4+5=9	42-4= 38
]41- 32]	7	0.17	17	9+7=16	38-5= 33
]50- 41]	7	0.17	17	16+7=23	33-7= 26
]59- 50]	11	0.26	26	23+11= 34	26-7= 19
]68- 59]	8	0.19	19	34+8= 42	19-11= 8
	42	1	100	-	-

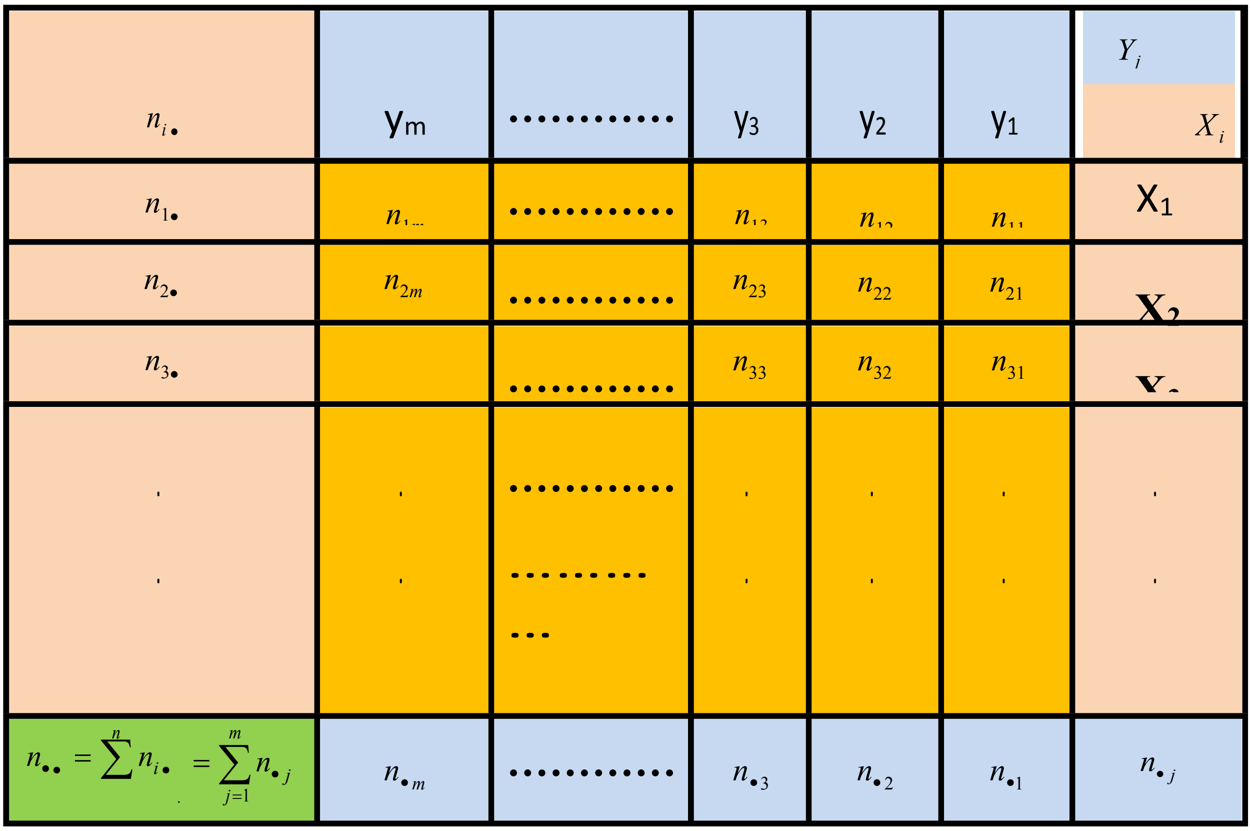
1.2.2 The two-way frequency distribution tables:

The two-way frequency distribution table is used when studying two phenomena (two features) of a given population simultaneously. The statistical data are put in the tables as follows:

- The lines include the data of the 1st feature while the columns include the data of the 2nd feature.

- We refer to the values of the 1st feature with X_i (where $_i$= 1, 2,…..n) and to the data of the second feature with Yj (where $_j$= 1, 2, ……, m).

Table 2-06: the general shape of the two-way frequency distribution table



$n_{i\bullet}$	y_m	....	y_3	y_2	y_1	Y_j X_i
$n_{1\bullet}$	n_{1m}	....	n_{13}	n_{12}	n_{11}	X_1
$n_{2\bullet}$	n_{2m}	....	n_{23}	n_{22}	n_{21}	X_2
$n_{3\bullet}$		....	n_{33}	n_{32}	n_{31}	X_3
. .	. .	....	. .	. .	. .	. .
$n_{\bullet\bullet}=\sum^{n} n_{i\bullet}=\sum_{j=1}^{m} n_{\bullet j}$	$n_{\bullet m}$	....	$n_{\bullet 3}$	$n_{\bullet 2}$	$n_{\bullet 1}$	$n_{\bullet j}$

Example 2-04: We draw a random sample from a population of 100 families to study the phenomena of expenditures of the family (10^3 DA) and the number of children in the family. Findings are shown in this table:

Number of children	0	1	2	3	$\sum$
Expenditures					
]30-20]	3	8	6	3	20
]40-30]	8	15	12	13	48
]50-40]	4	7	11	10	32
$\sum$	15	30	29	26	100

From this table, we can see that:

- From 100 families, the expenditures of 48 families are between 30x 10^3DA and 40x10^3DA.
- 30 families have one child, the expenditures of 07 families out of them are between 40x10^3DA and 50x10^3DA, the expenditures of 08 families

are between $20x10^3$DA and $30x10^3$DA, and the expenditures of the remaining 15 families are between $30x10^3$DA and $40x10^3$DA.

2. The graphic representation of the statistical data:

We can describe and summarize data using graphs and charts, which help make a fast analysis of the study phenomenon. The type of charts and graphs differs according to the study variable.

2.1 Graphic representation in the case of a discrete quantitative variable:

2.1.1 The graphic representation of the simple frequencies:

It is the use of simple bars whose lengths suit the frequency that corresponds to a given value of the study variable.

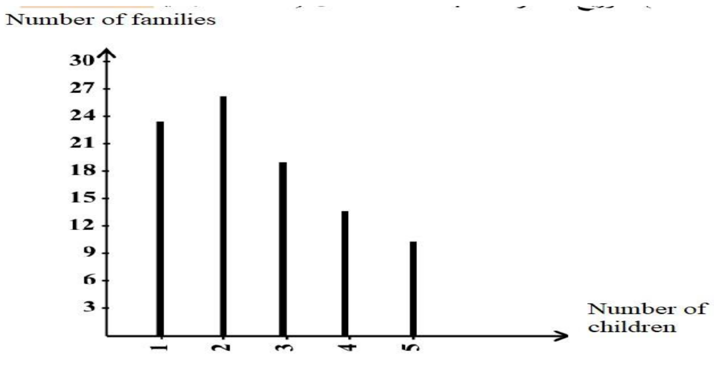
Example 2-05: The following table shows the number of children per family for a sample of 100 families.

number of children per family X_i	1	2	3	4	5	6
Frequency n_i	25	28	20	15	12	100

Task: present these data with the suitable method.

Solution: The best method is using the simple bars

Figure 2-01: the distribution of the families according to the number of children (simple bars)



2.1.2 The graphic representation of the ascending and descending cumulative frequencies:

- Ascending cumulative frequencies: It is a set of ascending straight parts according to the ascendance of the ascending and cumulative frequencies corresponding to each value of the studied statistic variable.

- Descending cumulative frequencies: It is a set of descending straight parts according to the descendance of the descending cumulative frequencies. The 1st straight part corresponds to the total frequencies and the minimum value of the study variable, and the 2nd part corresponds to the total frequencies minus the 1st simple frequency with the 2nd value of the statistical variable, and so on.

To illustrate how to draw the ascending and descending cumulative frequencies, we take example 2-05:

Table 2-08

Children per family X_i	Frequency n_i	n_i	n_i
1	25	25	100
2	28	53	75
3	20	73	47
4	15	88	27
5	12	100	12
	100	-	-

Figure 2-02: the chart of the ascending cumulative frequencies of the families' distribution:

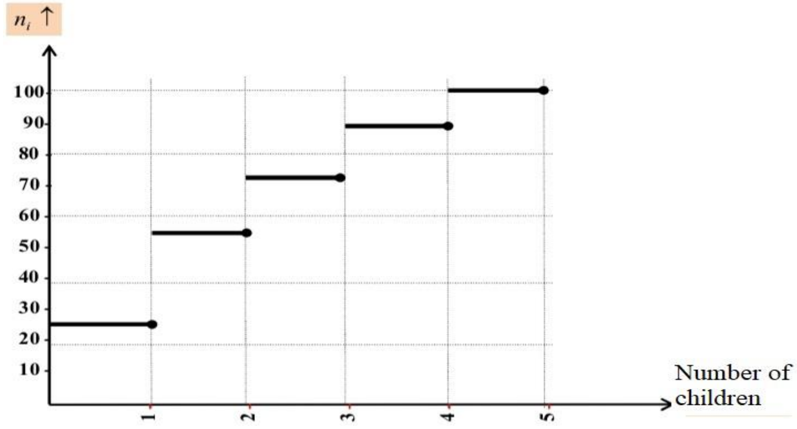
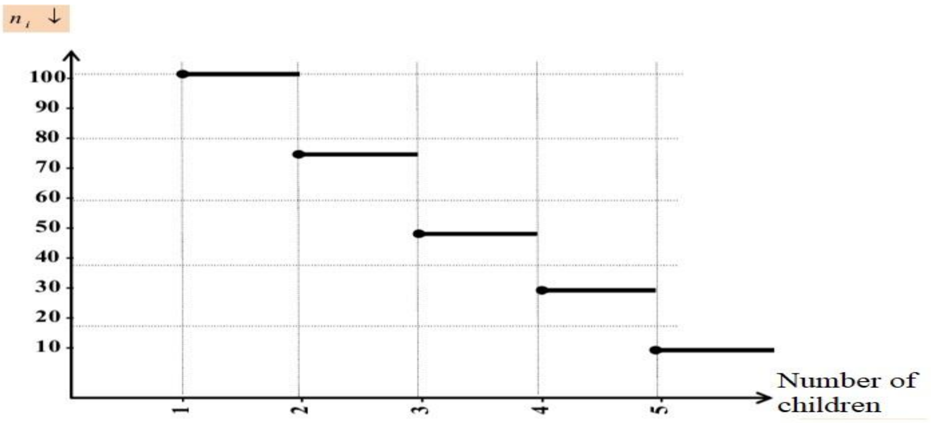


Figure 2-03: the chart of the descending cumulative frequencies of the families' distribution:



2.2 Graphic representation in the case of a continuous quantitative variable:

The graphic representations of the continuous quantitative variable are the most used, mainly:

2.2.1 The histogram:

It is a set of attached rectangles that represent the frequencies or values of each class. The length of each rectangle suits the corresponding frequency, and the basis of each one equals the length of the corresponding class. The classes are written on the X-axis while the frequencies are written on the Y-axis. In this context, it is important to notice if the lengths of the classes are equal before drawing the graph. Therefore, there are two cases when drawing it:

- In case the classes have equal lengths: In this case, the comparison basis is fixed and equal. Therefore, we directly draw the histogram on it. To illustrate this, we look at this example:

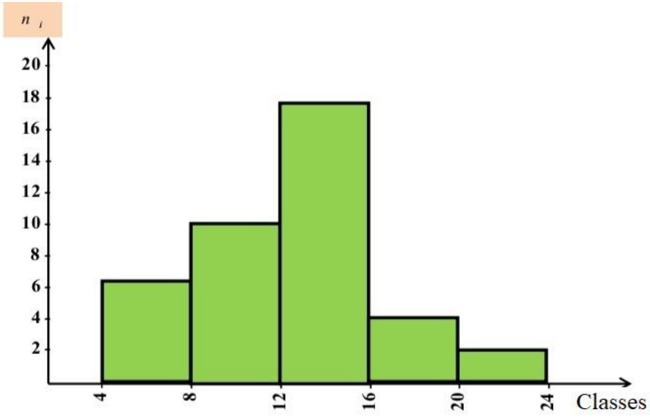
Example 2-06: The distribution of the daily expenditures (unit: 10 DA) of a sample of 40 students.

Expenditures classes X_{i}	]4-8]	]8-12]	]12-16]	]16-20]	]20-24]	
Frequency n_i	6	10	18	4	2	40

Task: Make a graphic representation for the distribution of the expenditures.

Solution: Because the classes have equal lengths, we directly draw the histogram.

Figure 2-04: Histogram of the expenditures distribution.



- In case the classes do not have equal lengths: In this case, we modify the frequencies (because the comparison basis is not fixed) to create suitability between the length of the class and the corresponding frequency. Therefore, we use this equation to modify the frequencies:

$$\text{The modified frequency}=\frac{\text{The frequency of the class}}{\text{The length of the class}} * \text{the chosen length of the class}$$

x

$$\text{ni*}=\frac{ni}{Li}$$

PS: the chosen length of the class (L^*) is the highest common denominator of the classes lengths.

PS: we modify the frequencies (in case the classes do not have equal lengths) for two purposes:

- Drawing the histogram.
- Identifying the mode class and calculating the mode.

Example 2-07: The following table shows the distribution of a sample of 100 employees in a given company according to the monthly salary (unit: 10^3 DA);

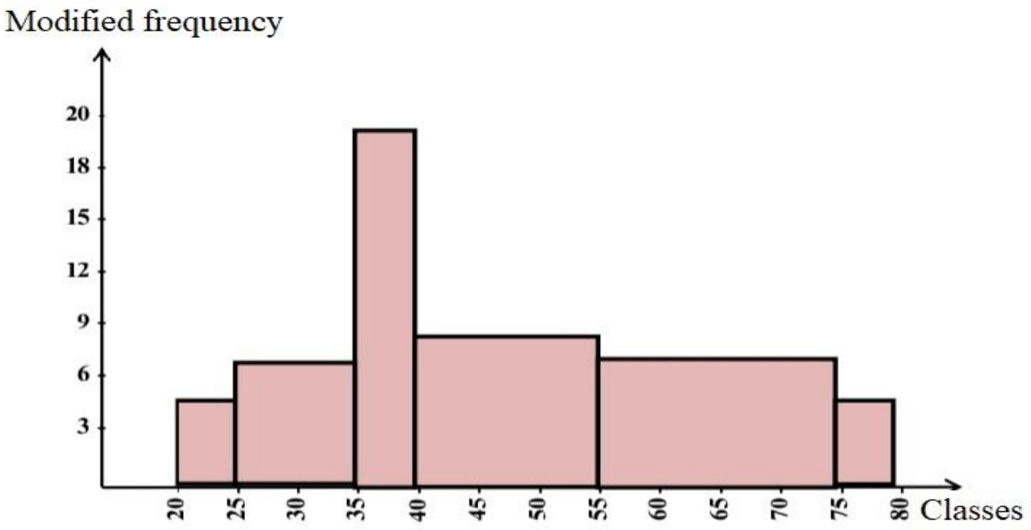
Salary classes X_i	]20-25]	]25-35]	]35-40]	]40-55]	]55-75]	]75-80]	
Frequency n_i	5	15	20	25	3	5	100

Task: Present these data using the histogram?

Solution: Because the distribution classes do not have equal lengths, we modify the frequencies and take the chosen length of the class that equals 5 as a basis for the modification (the chosen length of the class is the highest common denominator of the classes lengths or the lowest class length)

Salary classes X_i	Frequency n_i	Length of the class (L*)	Modified frequency n_i^*
]20-25]	5	5	$n^*_1 = \frac{5}{5} * 5 = 5$
]25-35]	15	10	$n^*_2 = \frac{15}{10} * 5 = 7{,}5$
]35-40]	20	5	$n^*_3 = \frac{20}{5} * 5 = 20$
]40-55]	25	15	$n^*_4 = \frac{25}{15} * 5 = 8{,}33$
]55-75]	30	20	$n^*_5 = \frac{30}{20} * 5 = 7{,}5$
]75-80]	5	5	$n^*_6 = \frac{5}{5} * 5 = 5$
$\sum$	100	-	-

Figure 2-05: The histogram of the salary distribution (modified frequencies)

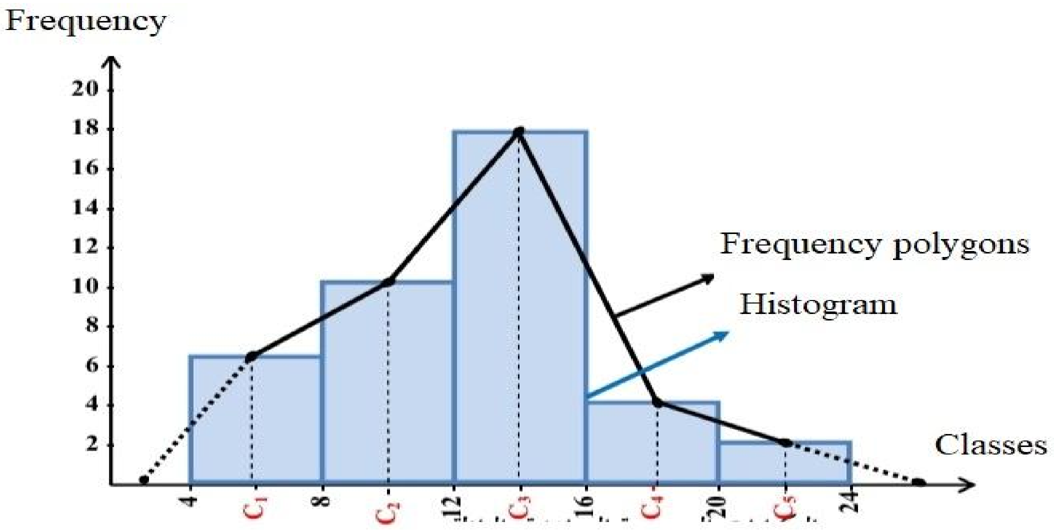


2.2.2

Frequency polygons:

It is a set of connected refracted straight parts determined according to their coordinates, the classes’ centers, and the corresponding frequencies. To illustrate how we can draw frequency polygons, we consider example (2-06) and, then, draw the histogram and the frequency polygons.

Figure 2-06: The histogram and the frequency polygons of the salary distribution.



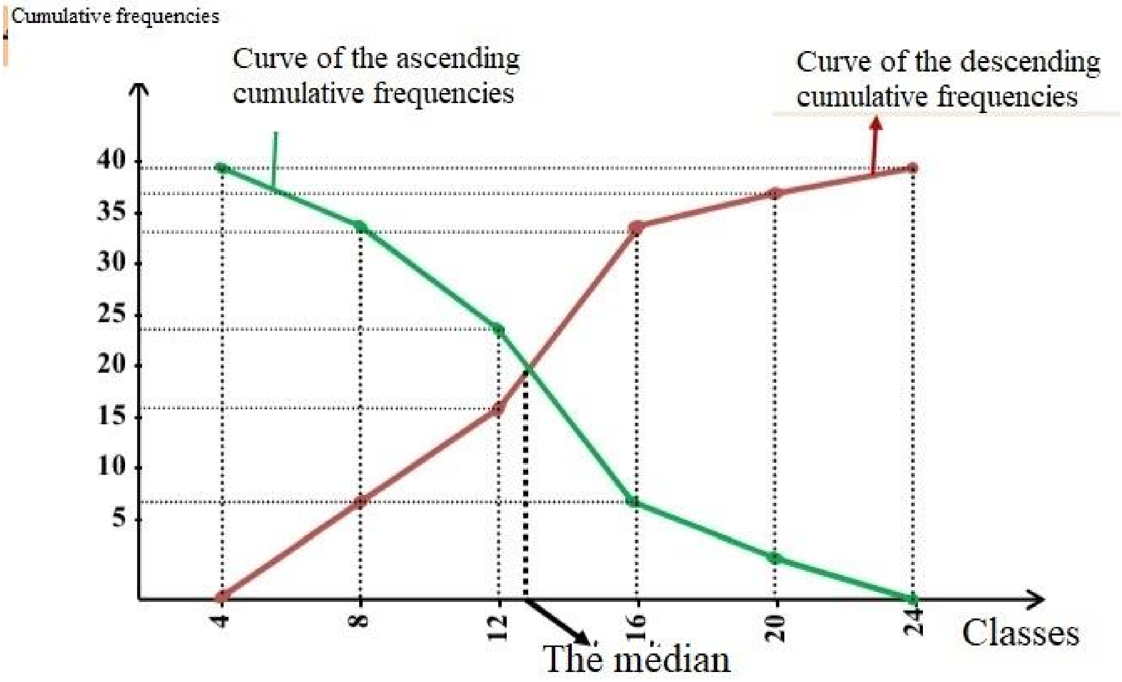
2.2.3 The curve of the ascending and descending cumulative frequencies:

We draw the curve of the ascending cumulative frequencies through connecting a set of points whose coordinates are the maximum limits of the classes and the corresponding ascending cumulative frequencies. As for the curve of the descending cumulative frequencies, we draw it through connecting a set of points whose coordinates are the minimum limits of the classes and the corresponding descending cumulative frequencies. The cross point between the two curves is called the median. We consider example 2-06 to draw an example:

Table 2-10: The descending and ascending cumulative frequencies

Classes of expenditures X_i	Frequency n_i	n_i	n_i
]4-8]	6	6	40
]8-12]	10	16	34
]13-16]	18	34	24
]16-20]	4	38	6
]20-24]	2	40	2
	40	-	-

Figure 2-07: curve of the ascending and descending cumulative frequencies



2.3 Graphic representation in the case of a qualitative variable

2.3.1 The pie chart:

It is a circle divided into many parts that correspond central angles that suit the frequencies of each studied feature. We add a column to the data table that contains the central angle that corresponds each frequency. We calculate the central angle as follows:

$$\text{The central angle}= \frac{\text{The frequency of the feature}}{\text{The total frequencies}} * 360$$

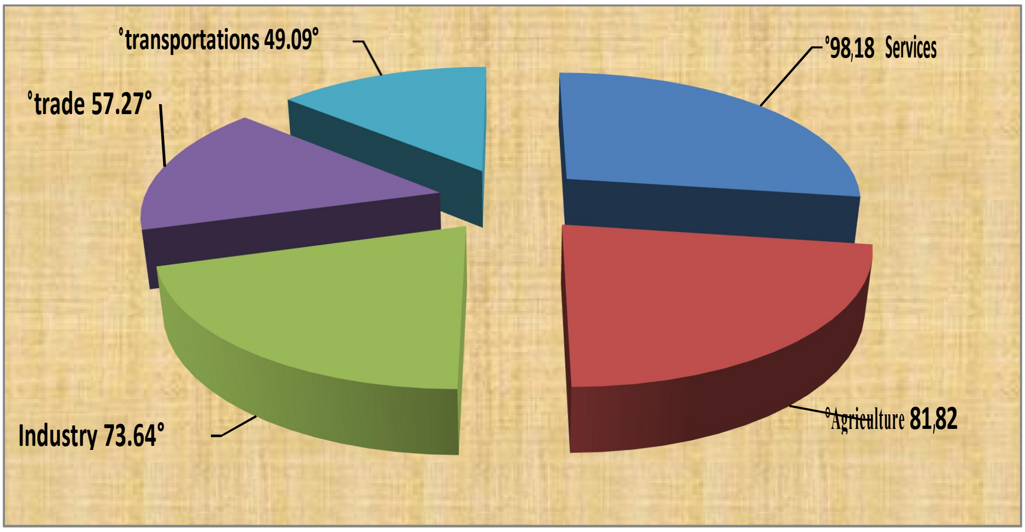
Example 2-08: The following table shows the national income from different sectors (million USD) of a given state in a given year.

Task: Make the pie chart.

Sector	National income	f_i %	Central angle
Services	1200	27%	98.18°
Agriculture	1000	23%	81.82°
Industry	900	20%	73.64°
Trade	700	16%	57.27°
Transportation	600	14%	49.09°
Total	4400	100%	360°

Figure 2-08: the distribution of the national income according to the sectors

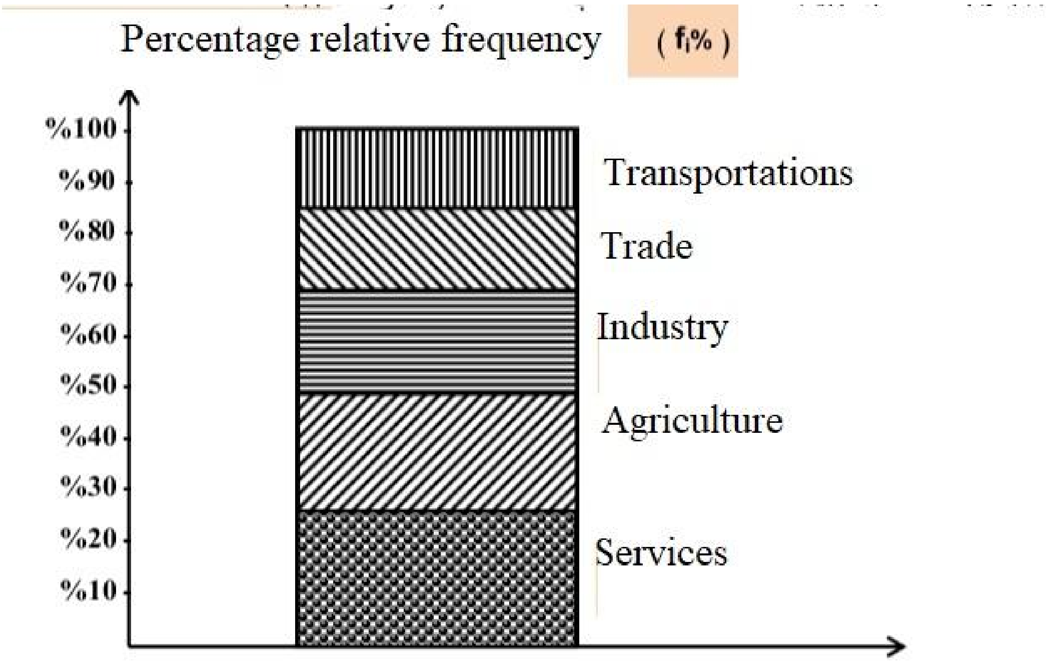
(pie chart)



2.3.2 The divided bar chart:

It is a rectangle divided into many parts that correspond to the frequencies of the studied features. It is better to use the percentages that correspond to each frequency, as the length of the straight line is 100%.

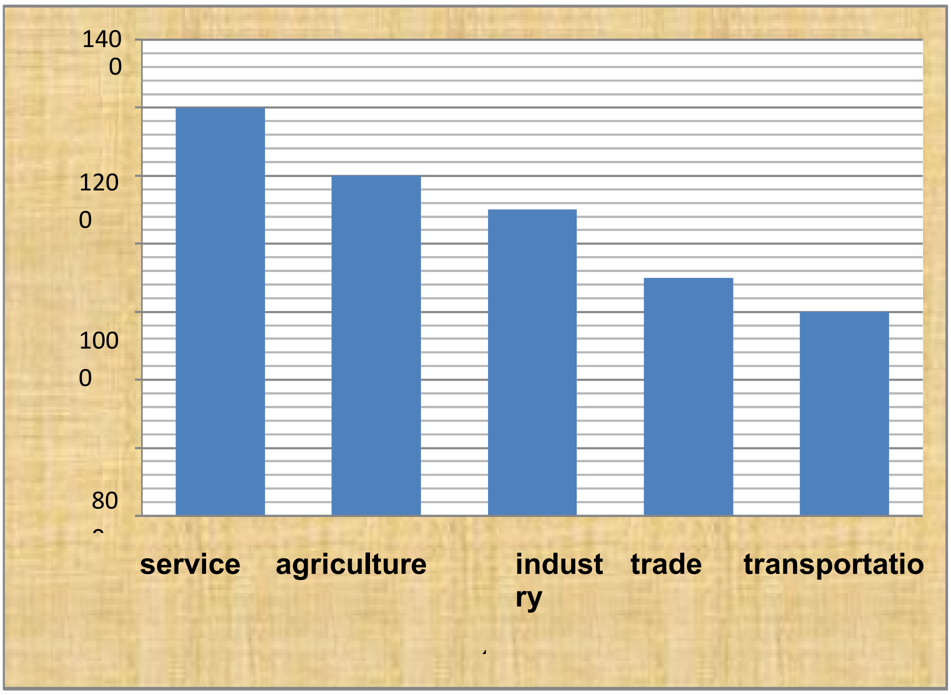
Figure 2-09: The distribution of the national income according to the sectors (divided bar chart)



2.3.3 Rectangular bars:

It is a set of rectangles separated by fixed distances, and having equal bases whose lengths suit the frequencies corresponding to the components of the studied feature. We consider the previous example to draw the rectangular bars:

Figure 2-10: the distribution of the national income according to the sectors (the rectangular bars)



Activities of chapter 02

Activity 01: The following table shows the number of male and female students enrolled at 1st year economics in a given university from 2010 to 2015.

Year	Males	Females
2010-2011	250	450
2011-2012	350	500
2012-2013	400	520
2013-2014	410	490
2014-2015	390	610

Task: present the data of the table using two different graphic representations.

Solution: We can present the data using the rectangular bars.

Figure 2-11: distribution of students according to the gender.

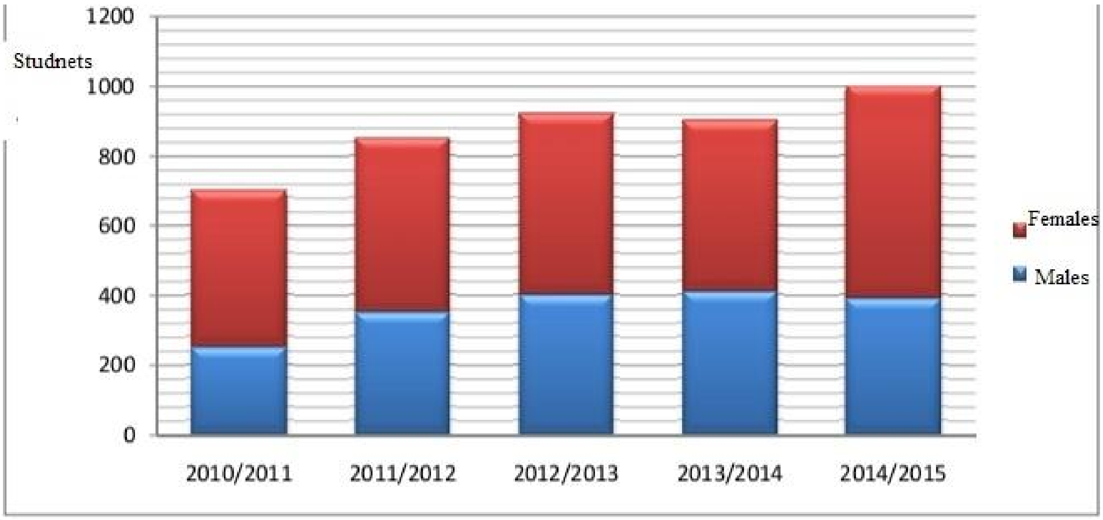
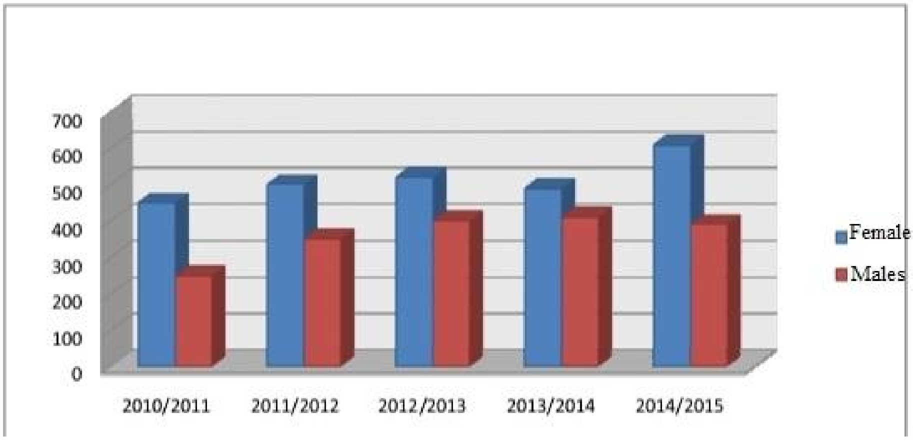


Figure 2-12: Distribution of the students according to the gender



Activity 02: The following data represent the expenditures of 75 people in a week (unit 10^2 DA)

62	72	68	53	73	82	68	78	66	62	65	74	73	67	73
69	74	81	63	63	83	60	79	75	71	79	62	69	97	78
83	75	61	76	65	82	78	75	73	66	75	82	73	84	77
93	73	57	90	60	96	78	79	71	85	75	60	90	71	79
62	88	68	76	83	65	75	87	74	85	91	80	79	89	76

Task:

1. Identify the study variable and its nature.
2. Put these data in a frequency distribution table using Yule method.
3. Draw the histogram and frequency polygons.
4. Find the relative frequency and the percentage frequency distribution.
5. Find the ascending and descending cumulative frequencies.
6. Identify the rate of people whose weekly expenditures exceed 77x10^2 DA.
7. Identify the people whose expenditures are between 65 x10^2 DA and 80 x10^2 DA.

Solution:

1- The study variable: the weekly expenditures

Its nature: continuous quantitative variable.

2. Putting these data in a frequency distribution table:

- Identifying the range: $R= X_{\max} - X_{\min}= 97 - 53 = 44$

- Identifying the number of classes using Yule equation:

$K= 2{,}5\sqrt[4]{n}= 2{,}5\sqrt[4]{75} = 2{,}5 * 2{,}9428 = 7, 357 = 8$

$L= \frac{R}{K} = \frac{44}{7{,}357} = 5{,}98 = 6$

We take the number of classes (K= 8) to achieve the inequality:

The length of the class * the number of the classes ≥ the range

$8 * 6= 48 > 44$

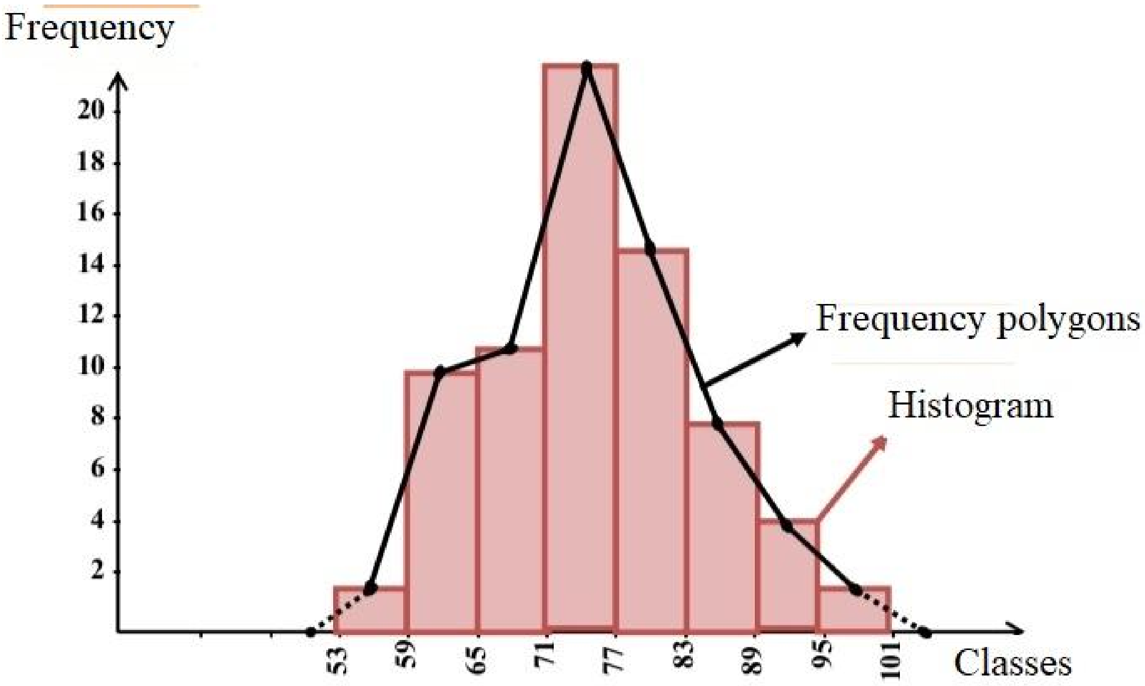
- Tabulation:

Table 2-11: Distribution of people according to their weekly expenditures:

Salary classes X_i	Inputting data	Number of employees n_i	f_i	$f_i\%$	n_i	n_i
]53-59]	//	2	0.0267	2.67	2	75
]59-65]	////// ////	10	0.133	13.3	12	73
]65-71]	////// /// ///	11	0.147	14.7	23	63
]71-77]	////////////////////////	22	0.293	29.3	45	52
]77-83]	/////////////	15	0.2	20	60	30
]83-89]	/////////	8	0.107	10.7	68	15
]89-95]	////	5	0.067	6.7	73	7
]95-101]	//	2	0.0267	2.67	75	2
	-	75	1	100	-	-

3- Drawing the histogram and the frequency polygons:

Figure 2-13: the histogram and the frequency polygons of the expenditures distribution



6. The rate of people whose weekly expenditures exceed 77x10^2 DA:

We start by determining the interval of expenditures:]77-101]. It is divided into:

Interval]53-59] with 15 people.

Interval]83-89] with 8 people.

Interval]89-95] with 5 people.

Interval]95-101] with 02 people.

Thus, the number of people is 15+8+5+2= 30. Or, we can find it from the descending cumulative frequencies that correspond to the class]77-83].

Thus, the rate of people whose weekly expenditures exceed 77x10^2 DA is $\frac{30}{75}$= 0.4= 40%.

7. The number of people whose expenditures are between 65 x10^2 DA and 80 x10^2 DA:

We identify the interval of the classes:]65-80]. It is divided into:

Interval]65-71] with 11 people.

Interval]71-77] with 22 people.

Interval]77-3] : The length of class 6 → 15 people

Interval]77-80]: The length of class 6 → x

Thus, x= $\frac{3*15}{6}$ = 7.5=8. The number of people is 11+22+8= 41

Thus, the number of people whose expenditures are between 65 x10^2 DA and 80 x10^2 DA is 41.

Activity 03:

The researcher wants to study the monthly salaries of 100 employees of a given company. After the study, he found the data shown in this table (unit: 10^3 DA)

Salary classes X_i	]30-34]	]34-42]	]42-46]	]46-54]	]54-62]	]62-70]	
Frequency n_i	4	16	20	24	28	8	100

Task:

1. Identify the study variable and its nature.
2. Present the data with the suitable graphic representation.

3. Find the ascending and descending cumulative frequencies.
4. Identify the rate of employees whose monthly salaries are less than $46x10^3$ DA.

Solution:

1. The study variable: the monthly salaries.

Its nature: continuous quantitative variable**.**

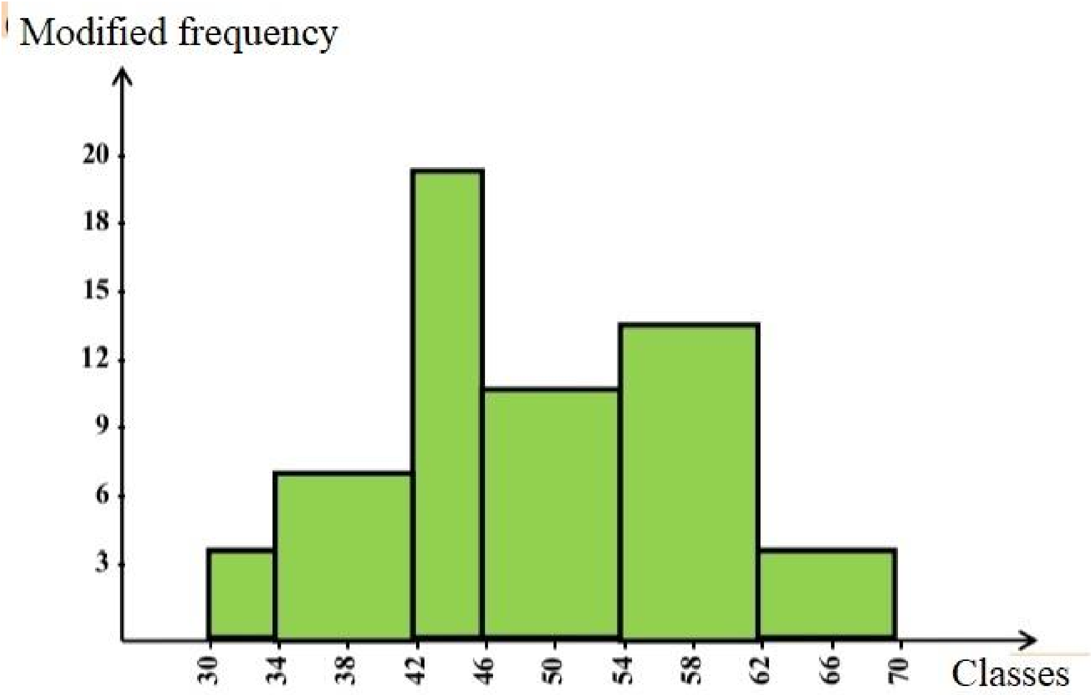
2. Graphic representation of data:

Because the variable is continuous quantitative, it is better to use the histogram. Nevertheless, we notice that the distribution classes do not have equal lengths. Therefore, we must modify the frequencies and take the chosen length of the class that equals 4 as a basis for modifying the frequencies.

Table 2-12: the ascending and descending cumulative frequencies.

Salary classes X_i	Frequency n_i	Length of the class L_i	Modified frequency n_i^*	n_i	n_i
]30-34]	4	4	$n^*_1 = \frac{4}{4} * 4 = 4$	4	100
]34-42]	16	8	$n^*_2 = \frac{16}{8} * 4 = 8$	20	96
]42-46]	20	4	$n^*_3 = \frac{20}{4} * 4 = 4$	40	80
]46-54]	24	8	$n^*_4 = \frac{24}{8} * 4 = 8$	64	60
]54-62]	28	8	$n^*_5 = \frac{28}{8} * 4 = 8$	92	36
]62-70]	8	8	$n^*_6 = \frac{8}{8} * 4 = 8$	100	8
	100	-	-	-	-

Figure 2-14: The histogram of the salaries distribution



3. The rate of employees whose monthly salaries are less than 46x10^3 DA:

The interval of the salary is]30-46]. It is divided into:

Interval]30-34] with 04 people.

Interval]34-42] with 16 people.

Interval]42-46] with 20 people.

Thus, the number of employees is 4+16+20=40. Or, we can find it from the ascending cumulative frequency that corresponds to the class]42-46].

Thus, the rate of employees whose monthly salaries are less than 46x10^3 DA is $\frac{40}{100}$ = 0.4=40%

Further activities:

Activity 01:

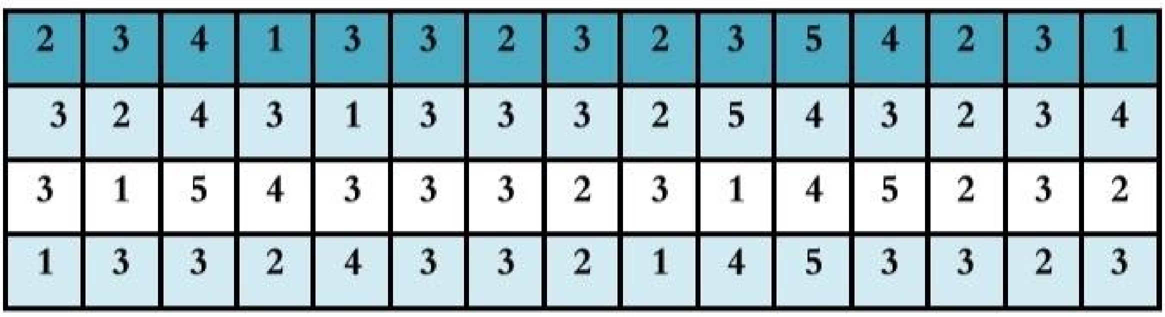
In order to study the familial status of a given company, the researcher administered questionnaires to the employees. Findings are shown in the table:

Married	Single	Single	Divorced	Divorced	Married	Married	Divorced	Married	Married
Single	Married		Divorced	Married	Single	Single	Married	Divorced	widowed
Widowed	Divorced	Married	Single	Divorced	Married	Married	Married	Single	Single
Married	Married	Divorced	Widowed	Married	Single	Married	Single	Married	Widowed

Task: what is the type of these data? Put them in a suitable table.

Activity 02:

A random sample of 60 families in a given district was chosen to study the number of rooms in their houses. Findings are shown in the table:



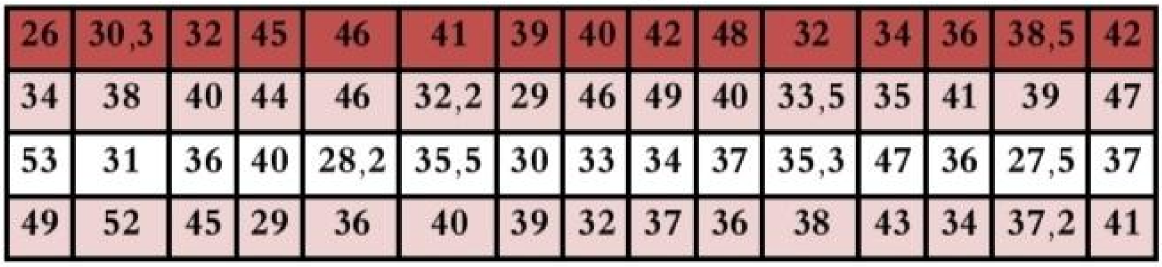
2	3	4	1	3	3	2	3	2	3	5	4	2	3	1
3	2	4	3	1	3	3	3	2	5	4	3	2	3	4
3	1	5	4	3	3	3	2	3	1	4	5	2	3	2
1	3	3	2	4	3	3	2	1	4	5	3	3	2	3

Task:

1. Identify the variable and its nature.
2. Present graphically the data after putting them in a suitable table.
3. Find the relative frequencies and the percentage relative frequencies.
4. Present graphically the ascending and descending cumulative frequencies.

Activity 03:

A research unit conducted a study and collected data about the weights of a group of students at a given middle school (unit: Kg). Findings are as follows:



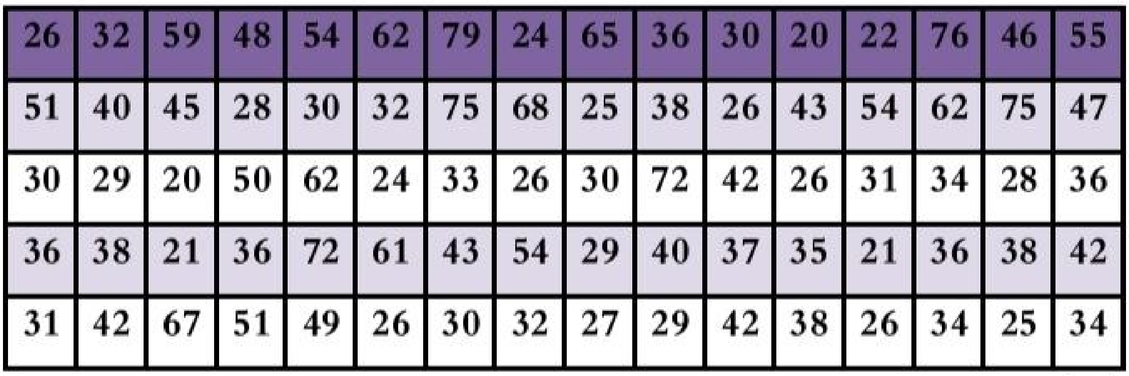
26	30,3	32	45	46	41	39	40	42	48	32	34	36	38,5	42
34	38	40	44	46	32,2	29	46	49	40	33,5	35	41	39	47
53	31	36	40	28,2	35,5	30	33	34	37	35,3	47	36	27,5	37
49	52	45	29	36	40	39	32	37	36	38	43	34	37,2	41

Task:

1. Identify the variable and its nature.
2. Put the data in a table using the classes of length 04.
3. Present graphically the data after putting them in a suitable table.
4. Find the relative frequencies and the percentage relative frequencies.
5. Present graphically the ascending and descending cumulative frequencies.

Activity 04:

In a survey made by a researcher on a sample of 80 families in a given city to study their monthly expenditures (unit: 10^3 DA), he found these findings:



26	32	59	48	54	62	79	24	65	36	30	20	22	76	46	55
51	40	45	28	30	32	75	68	25	38	26	43	54	62	75	47
30	29	20	50	62	24	33	26	30	72	42	26	31	34	28	36
36	38	21	36	72	61	43	54	29	40	37	35	21	36	38	42
31	42	67	51	49	26	30	32	27	29	42	38	26	34	25	34

Task:

1. Identify the variable and its nature.
2. Put the data in a frequency distribution table using Staugres method.
3. Draw the histogram and the frequency polygons.

4. Present graphically the ascending and descending cumulative frequencies.
5. Identify the rate of families whose expenditures exceed 60 x10^3 DA.
6. Identify the number of families whose expenditures are between 65 x10^3 DA and 75 x10^3 DA.

Activity 05:

The table shows the distribution of a group of employees according to their monthly salary (10^3 DA)

Salary classes X_i	]18-24]	]24-36]	]36-42]	]42-48]	]48-60]	]60-72]	
Frequency n_i	12	24	36	24	30	24	150

Task:

1. Identify the study variable and its nature.
2. Draw the histogram.
3. Find the ascending and descending cumulative frequencies and present them with graphic representations.
4. Find the number of employees whose salaries are between 36x10^3 DA and 54 x10^3 DA.
5. Identify the rate of employees whose salaries are less than 42 x10^3 DA.

Activity 06:

The following table shows the sales of the unit of an electronics company in 2014.

Device	TV	Radio	Refrigerator	Oven	Washing machine
Number of devices	1000	750	600	550	300

Task: present the data with a pie chart and rectangular bars.

Chapter 03:

Measures of the central tendency:

1. Concept of the central tendency

2. Measuring the central tendency

2.1 The arithmetic mean

2.2 The geometric mean

2.3 The harmonic mean

2.4 The quadratic mean

2.5 The median

2.6 The mode

3. The relation between the measures of the central tendency

Activities of chapter 03

Preamble:

The descriptive statistics uses various measures to describe the phenomenon, mainly the measures of the central tendency that prepare a basis of necessary values or parameters for analysis, prediction, or decision taking. By observing the data of any phenomenon, either in their primary image or after summary and classification in frequency distribution tables, we find that most of the items are centered around a specific value that represents the distribution center. Therefore, it is necessary in studying the characteristics of the distribution, comparison, and frequency distributions of the same phenomenon. Thus, after data collection, the researchers summarize and tabulate them in frequency distribution tables and, then, graphically represent them. Next, the phase of using the statistical measures on the data of the phenomenon starts to express them in one or more values that serve in comparison or prediction. These measures are known as the central tendency measures.

1. Concept of the central tendency:

It refers to the measures that estimate a value around which most of the values centralize. This mean or central value is one number that represents all the data of the group. Besides, these measures give a clear idea about the phenomenon, not substitute the detailing data.

2. Measuring the central tendency:

It expresses the center of the statistical distribution. To conduct it, we use these tools:

2.1 The arithmetic mean:

It is one of the most used central tendency measures. It is the balance center of any phenomenon. We can define it as the outcome of dividing the total data values **i** by their number **ii** in case of a sample, and on **N** in case of a population. It is referred to with $\overline{X}$.

2.1.1 Calculating the arithmetic mean:

There are various methods, and it can be calculated from the primary data (untabulated) or the tabulated data.

2.1.1.1 Calculating the arithmetic mean from the primary data:

The direct method:

It is used when the data are not tabulated, i.e., when the measures of the study variable have the same importance. If we have X_1, X_2…….X_n, representing the values of a specific phenomenon, the arithmetic mean is the total of the values divided by their number. The formula of the arithmetic mean is given as follows:

$$\overline{X}=\frac{x1+x2+……….xn}{n}=\frac{\sum_{i=1}^{n} xi}{n}$$

$$\overline{X}=\frac{\sum_{i=1}^{n} xi}{n}$$

Example 3-01: This statistical series is the expenditures of 08 customers in their breakfast in a coffee shop:

40	45	35	50	55	3	25	60

Task: Find the expenditure mean.

$$\overline{X}=\frac{\sum_{i=1}^{n} xi}{n}=\frac{\sum_{i=1}^{8} xi}{8}=\frac{40+45+35+50+55+30+25+60}{8}=\frac{370}{8}=42.5$$

Remark:

The arithmetic mean of a sequenced series of values is the average value if the number of data is odd, and the average of the two middle values if the number of the values is even.

- Odd number of data (n): $\bar{x}$= x $\frac{n+1}{2}$
- Even number of data (n): $\bar{x} = \frac{x\frac{n}{2} + x\frac{n}{2}+1}{2}$

The simple deviations method (the hypothetical mean):

If we have X_1, X_2…….X_n representing the values of a given phenomenon, the number X_0 is the hypothetical mean of the data (it is better that the value of the hypothetical mean X_0 equals one of the values of the phenomenon and that this value is in the middle), and (X_i – X_0) represents the deviation of the values from the hypothetical mean, the arithmetic mean is given as follows:

$$\overline{x} = x_0 + \frac{\sum_{i=1}^{n}(xi - x0)}{n}$$

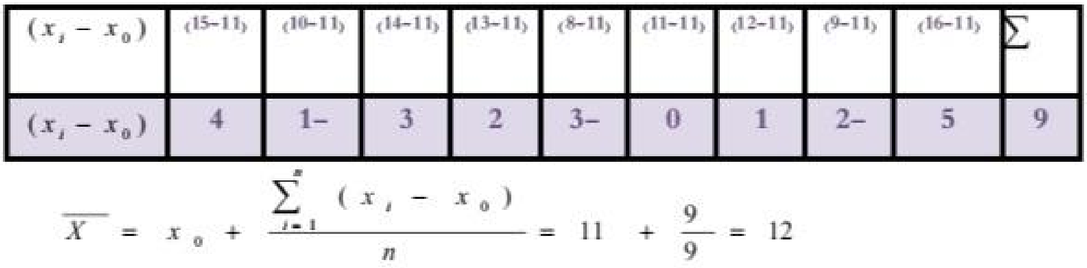
Example 3-02:

Find the arithmetic mean using the hypothetical mean:

15	10	14	13	8	11	12	9	16

Solution:

We suppose that the hypothetical mean is X_0= 11.



$(x_i - x_0)$	(15–11)	(10–11)	(14–11)	(13–11)	(8–11)	(11–11)	(12–11)	(9–11)	(16–11)	$\sum$
$(x_i - x_0)$	4	1–	3	2	3–	0	1	2–	5	9

$$\overline{X} = x_0 + \frac{\sum_{i=1}^{n}(x_i - x_0)}{n} = 11 + \frac{9}{9} = 12$$

2.1.1.2 Calculating the arithmetic mean from the tabulated data:

The direct method:

In some cases, the values do not have the same importance as they have relative importance that differs with the weighting factor. Thus, we need the weighted arithmetic mean. In this context, we have two cases:

- In case of discrete quantitative variable:

If X_1, X_2……..X_k represent values of a given phenomenon X, and n_1, n_2……..n_n represent the corresponding values, the arithmetic mean is given as follows:

$$\overline{x} = \frac{x1.n1 + x2.n2 + \dots\dots\dots + xk.nk}{n1 + n2 + \dots\dots\dots + nk} = \frac{\sum_{i=1}^{k} xi.ni}{\sum_{i=1}^{k} ni}$$

$$\overline{x} = \frac{\sum_{i=1}^{k} xi.ni}{\sum_{i=1}^{k} ni}$$

Where:

X_i is the value of the item i (i= 1, 2,…..k)

n_i is the frequency of the item i.

k: is the number of the different values.

Example 3-03:

The following statistical data show the distribution of families in a given district according to the number of the rooms:

Table 3-01: The distribution of families according to the number of the rooms:

Number of the rooms X_i	**1**	**2**	**3**	**4**	
Number of families n_i	**5**	**10**	**20**	**30**	**65**

Task:

Find the mean of the rooms:

Solution:

We have $\bar{x}=\frac{\sum_{i=1}^{k} xi.ni}{\sum_{i=1}^{k} ni}$

We add a third column to the table and, thus, the result of multiplying the value of the statistical variable by the corresponding frequency of this value is (x_i . n_i).

Table 3-02: The arithmetic mean

Number of rooms x_i	**Frequency n_i**	x_i n_i
1	**5**	**5**
2	**10**	**20**
3	**20**	**60**
4	**30**	**12**
	65	**205**

Thus, the mean of the rooms is:

$$\bar{x}=\frac{\sum_{i=1}^{k} xi.ni}{\sum_{i=1}^{k} ni}=\frac{\sum_{i=1}^{4} xi.ni}{\sum_{i=1}^{4} ni}=\frac{205}{65}=3{,}15 \sim 3.$$

- In case of a continuous quantitative variable:

In such case, the frequency distribution table takes the form of classes whose number is k. If c_1, c_2…………, c_k represents the class centers, and n_1, n_2…………, n_k represents the corresponding frequencies, we consider that the frequencies centralize around the class centers. Thus, we use the following formula:

$$\overline{x}=\frac{c1.n1+c2.n2\ldots\ldots\ldots ck.nk}{n1+n2\ldots\ldots\ldots\ldots nk}=\frac{\sum_{i=1}^{k} ci.ni}{\sum_{i=1}^{k} ni}$$

$$\overline{x}=\frac{\sum_{i=1}^{k} ci.ni}{\sum_{i=1}^{k} ni}$$

Where:

c_i is the center of the class i (i= 1, 2,…..k)

n_i is the frequency of the class i.

k: is the number of the classes.

Example 3-04:

The following statistical table shows the distribution of the employees of a given company based on their monthly salaries (unit: 10^3 DA)

Salary X_i	]10-20]	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]
Number of employees n_i	18	30	25	17	12	8

Task:

Find the mean of the monthly salaries of the employees.

Solution:

Finding the mean of the salaries:

1. We find the total frequencies $\sum_{i=1}^{k} ni$
2. We calculate the classes' centers c_i.
3. We multiply the center of each class by the corresponding frequency (c_i . n_i) and then find the total.

4. We calculate the arithmetic mean using $\overline{x}=\frac{\sum_{i=1}^{k} ci.ni}{\sum_{i=1}^{k} ni}$

Table 3-04: The arithmetic mean:

Salary classes *X*$_i$	**Frequency n$_1$**	**Center of the class c$_1$**	**c$_i$ x n$_i$**
]10-20]	**18**	**15**	**270**
]20-30]	**30**	**25**	**750**
]30-40]	**25**	**35**	**875**
]40-50]	**17**	**45**	**765**
]50-60]	**12**	**55**	**660**
]60-70]	**8**	**65**	**520**
	110	**-**	**3840**

Thus, the salaries mean is:

$$\overline{x}=\frac{\sum_{i=1}^{k} ci.ni}{\sum_{i=1}^{k} ni}=\overline{x}=\frac{\sum_{i=1}^{6} ci.ni}{\sum_{i=1}^{6} ni}=\frac{3840}{110}=34{,}9.$$

- The simple deviations (the hypothetical mean):

In this case, we calculate the arithmetic mean as follows:

$$\overline{x}= x0+\frac{\sum_{i=1}^{k}(xi-x0)ni}{\sum_{i=1}^{k} ni}$$

Where:

X_0 is the hypothetical mean.

(X_i - X_0) is the deviations of the values (items) or classes’ centers from their hypothetical mean

Remark:

It is better that the value of the hypothetical mean is the value of the most frequent item in case of a discrete quantitative variable, and the center of the class that is in the middle of the classes in case of a continuous quantitative variable. Besides, it is better that this class has the maximum frequency if possible.

Example 3-5:

We consider the data of the example 3-04 and calculate the arithmetic mean using the simple deviations.

To calculate the mean of the monthly salaries, we:

1. Determine the classes' centers c_i and x_i.
2. Determine the value of the hypothetical mean; it is the center of the class that corresponds to the maximum frequency (x_0= 25).
3. Calculate the differences between the classes centers x_i and the value of the hypothetical mean (x_0= 25).
4. Find the result of multiplying the frequency of each class by the difference between the class center and the value of the hypothetical mean. Then, we find the total of the multiplication results to get$\sum (xi - x0)ni$
5. Calculate the arithmetic mean through $\overline{x}$= x_0+$\frac{\sum_{i=1}^{k}(xi-x0)ni}{\sum_{i=1}^{k}ni}$

Table 3-05: The arithmetic mean using the hypothetical mean

Salary classes X_i	**Frequency n_1**	**Center of the class c_1**	(x_i x x_0)	(x_i x x_0) n
]10-20]	**18**	**15**	**-10**	**-180**
]20-30]	**30**	**25**	**0**	**0**
]30-40]	**25**	**35**	**10**	**250**
]40-50]	**17**	**45**	**20**	**340**
]50-60]	**12**	**55**	**30**	**360**
]60-70]	**8**	**65**	**40**	**320**
	110	**-**	**-**	**1090**

Thus, the mean salary is:

$$\bar{x}= x_0+\frac{\sum_{i=1}^{k}(xi-x0)ni}{\sum_{i=1}^{k}ni} = 25+\frac{\sum_{i=1}^{6}(xi-x0)ni}{\sum_{i=1}^{6}ni} = 25+\frac{1090}{110} =34,9.$$

Comment: The mean of the salaries is one value $\bar{x}$ = 34,9 despite the different calculation methods.

2.1.1.3 The weighted arithmetic mean of arithmetic means:

The weighted arithmetic mean is used to find the arithmetic mean of two or more data sets, whose arithmetic mean is known in case they are integrated in one group. If we have the group n whose arithmetic mean is $\bar{x}_2$and a second group m whose arithmetic mean is $\bar{x}_2$ the arithmetic mean of the two groups is

$$\bar{x}=\frac{n.\bar{x}1+m.\bar{x}2}{n+m}$$

Example 3-6:

The table shows the number of employees and the mean salary of the employees in a given company that has three productive branches:

Company branches	1st branch	2nd branch	3rd branch
Number of employees	150	120	80
Mean salary (DA)	35000	32000	4000

Task: find the mean of the salaries.

$$\bar{x}=\frac{n1.x1+n2.x2+n3.x3}{n1+n2+n3}=\frac{150*35000+120*32000+80.40000}{150+120+80}=\frac{12290000}{350}=\mathbf{35114\ da}$$

2.1.2 The characteristics of the arithmetic mean:

- It is the most used measure of the central tendency.
- It accepts the algebraic calculations and cannot be graphically calculated.
- It gets affected by the extreme values (the values in the limits of the study range).
- It cannot be calculated from the frequency distribution tables that are open from the beginning to the end because it depends on the classes centers.

2.2 The geometric mean:

In many cases, the values of the study phenomenon are rates or averages. Here, we study the average of a given phenomenon. Besides, the arithmetic mean neither properly describes the phenomenon nor gives correct idea. Therefore, another mean was found, known as the geometric mean G. It is a set of values or numbers X_1, X_2…………, X_n whose number is n with the nth root of the result of multiplying these values.

2.2.1 Calculating the geometric mean:

2.2.1.1 The geometric mean of the primary data:

If X_1, X_2…………, X_n represent items (values) of a given phenomenon X, the geometric mean of these values (n value) is the nth root of the multiplication of these value, given as follows:

$$G=\sqrt[n]{x1.x2\ldots\ldots.xn}$$

In order to simplify the calculation in case the data are big, we use one of the data logarithms to have the formula of the geometric mean as follows:

$$G=\sqrt[n]{x1.x2\ldots\ldots.xn}=(x_1.x_2\ldots\ldots x_n)^{1/n}$$

By inserting the logarithm to the two extremes, we get

$$\text{In } G=\frac{1}{n}\,\text{In}(x_1.x_2\ldots x_n)==\frac{1}{n}\,(\text{In } x_1.\text{In } x_2\ldots\text{In } x_n)=\frac{1}{n}\sum_{i=1}^{n} In\ xi$$

Thus, the relation of the geometric mean becomes

$$G=\ l^{\frac{1}{n}\sum_{i=1}^{n} In\ xi}$$

Example 3-07:

Consider this statistical series:

9	5	12	6	7

Task: find the geometric mean.

Solution:

Finding the geometric mean:

$$\boldsymbol{G=\sqrt[n]{x1\ .x2\ldots\ldots xn}=\ \sqrt[5]{9*5*12*6*7}=7{,}43.}$$

2.2.1.2 The geometric mean of the tabulated data:

If X_1, X_2…………, X_n represent items (values) of a given phenomenon X, and n_1, n_2…………, n_k represent the corresponding frequencies, the geometric mean is as follows:

$$\mathbf{G=\sqrt[N]{x1^{n1}\,.\,x2^{n2}\ldots\ldots xk^{nk}}}$$

Where N=$\sum_{i=1}^{k} ni$, and x_i is the values of the statistical variable or the classes centers in case of the tabulated data of discrete or continuous quantitative variables. In case the data are big, we simplify the calculation using:

$$G=\sqrt[N]{x1^{n1}.\,x2^{n2}\ldots..\,xk^{nk}}.= (x_1^{n1}.x_2^{n2}\ldots\ldots x_k^{nk}\ )^{1/n}$$

By inserting the logarithm, we get:

$$\text{In } G=\frac{1}{N}(x1^{n1}.\,x2^{n2}\ldots..\,xk^{nk})==$$
$$\frac{1}{N}(n1\ In\ x1+n2\ In\ x2\ldots..+nk\ in\ xk)=\frac{1}{N}\sum_{i=1}^{k} ni\ In\ xi$$

Thus, the relation of the geometric mean is as follows:

$$G= 1\ l^{\frac{1}{N}\sum_{i=1}^{k} ni\ In\ xi}$$

Where:

X_1 is the values of the variable or the classes' centers.

n_i is the frequencies corresponding to the values of the variable or the classes' centers.

Example 3-08:

We take the data of example 3-03 and calculate the geometric mean.

Solution:

Calculating the geometric mean:

Since the data are tabulated in a frequency distribution table, we use the following relation:

$$G= 1\, l^{\frac{1}{N}\sum_{i=1}^{k} ni\, In\, xi}$$

Number of rooms x_i	Frequency n_i	In x_i	n_i In x_i
1	5	0	0
2	10	0.693	6.93
3	20	1.09	21.97
4	30	1.38	41.58
	65	-	70.49

Thus, the geometric mean is

$$G= 1\, l^{\frac{1}{N}\sum_{i=1}^{k} ni\, In\, xi} = 1^{\,1,08} = 2,94 \sim 3$$

Example 3-09:

To calculate the geometric mean in case of tabulated data in a frequency distribution table (continuous quantitative variable), we take the data of example 3-04:

Solution:

Calculating the geometric mean:

We have this relation:

$$G= 1\, l^{\frac{1}{N}\sum_{i=1}^{k} ni\, In\, xi}$$

Table 3-07: Calculation of the geometric mean

Salary classes X_i	Frequency n_1	Center of the class c_1	In x_i	n_i In x_i
]10-20]	18	15	2.708	48.74
]20-30]	30	25	3.218	96.57
]30-40]	25	35	3.55	88.88
]40-50]	17	45	3.806	64.71
]50-60]	12	55	4.007	48.08
]60-70]	8	65	4.17	33.39

	110	**-**	**-**	**380.37**

The geometric mean of the salaries is G= $1\, l^{\frac{1}{N}\sum_{i=1}^{k} ni\ In\ xi}$ = $1^{3,46}$ =31,75

2.2.2 The uses of the geometric mean:

It is widely used in the economic life because the focus is generally on finding the mean of the change rates of some phenomena, such as the rate of GDP growth, rate of salaries increase, the demographic growth, etc. To show how to use the geometric mean in the economic life, we take example 3-10.

Example 3-10:

The table shows the development of the imports of a given state during 2011-2014 (unit: billion USD):

Year	2011	2012	2013	2014
Imports	10	11.2	13.5	16

Task: find the mean of the imports' increase rate.

Solution:

First, we find the rate of the imports increase from one year to another using

ti=$\frac{x_i - x_{i-1}}{x_{i-1}}$ $*100$

-The rate of the imports increase in 211-2012: We consider 2011 as the basis year. Thus,

t_1=$\frac{11,2-10}{10}$ $*100$ = 12%. We do the same with the rest of the periods and get the following table:

Year	2011	2012	2013	2014
Imports	10	11.2	13.5	16
Rate of the increase (t)	-	12%	20.5%	18.5%

We refer to the rate of the imports increase during 2011-2014 with (t), and to the increase rate during the 1st, 2nd, and 3rd years with (t_1), (t_2), and (t_3). So that (t) really represents the mean of the imports increase rate during 2011-2014, we must make sure that:

(1+t) (1+t) (1+t) = (1+t_1) (1+t_2) (1+t_3)

Thus, $(1+t)^3$ = (1+t_1) (1+t_2) (1+t_3)

We take the cube root to the equation limits and get 1+t $=\sqrt[3]{(1+t_1)(1+t_2)(1+t_3)}$

Thus, t=$\sqrt[3]{(1+t_1)(1+t_2)(1+t_3)}$ -1

Generally, if we have an increase rate of (n) period, the mean of the increase rate (rates mean) is given through:

t=$\sqrt[n]{(1+t_1)(1+t_2)....(1+t_n)}$ -1

By applying this relation on the data of example 3-10, we get:

t=$\sqrt[3]{(1.12 * 1.205 * 1{,}185}$ -1 =1,1694 -1 =0,1694 = 16,94%.

The mean of the imports increase rate during 2011-2014 is t= 0.169 = 16.94%.

2.2.3 The characteristics of the geometric mean:

- Its calculation involves all the values; however, it is less affected by the extreme values than the arithmetic mean.
- It cannot be calculated from the open frequency distribution tables.
- It is meaningless if one of the values is negative or equals 0.
- It is used with more realism when describing the relative phenomena.
- The value of the geometric mean of any phenomenon is always lower than the value of the arithmetic mean (G < $\bar{x}$)
- It is used in calculating the index numbers and when calculating the rates (interest rate, economic growth rate, democratic growth rate, etc).

2.3 The harmonic mean:

It is the reciprocal of the arithmetic mean of reciprocals of the values. It is one of the measures that determine the speed, price, and population rates.

2.3.1 Calculating the harmonic mean:

2.3.1.1 The harmonic mean of the primary data:

If X_1, X_2…………, X_n represent items (values) of a given phenomenon X, the reciprocals of these values are $\frac{1}{x_1}$, $\frac{1}{x_2}$,….. $\frac{1}{x_n}$, the arithmetic mean of the values reciprocals is $\frac{\frac{1}{x_1}+\frac{1}{x_2},+.....\frac{1}{x_n}}{n}$. On the other hand, the reciprocal of the arithmetic mean of these values reciprocals is the harmonic mean given as follows H = $\frac{n}{\frac{1}{x_1}+\frac{1}{x_2},+\dots\dots\frac{1}{xn}}$ It is abbreviated as follows H = $\frac{n}{\sum_{i=1}^{n}\frac{1}{xi}}$

Example 3-11:

We have the following statistical series:

9	5	12	6	7

Task: find the harmonic mean.

Solution: We have

$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2},+\dots\dots\frac{1}{xn}} = H = \frac{6}{\frac{1}{9}+\frac{1}{5},+\frac{1}{12}+\frac{1}{6}+\frac{1}{7}} = \frac{6}{0{,}703} = 8{,}52.$$

2.3.1.2 The harmonic mean of the tabulated data:

If X_1, X_2…………, X_n represent items (values) of a given phenomenon X, and n_1, n_2…………, n_k represent the corresponding frequencies, the harmonic mean is defined as follows:

$$H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k}\frac{n_i}{x_i}}$$

Where:

X_1 is the values of the variable or the classes' centers.

n_i is the frequencies corresponding to the values of the variable or the classes' centers.

Example 3-12:

To calculate the harmonic mean in case of tabulated data in a frequency distribution table (continuous quantitative table), we take the data of example 3-04.

Solution:

Calculating the harmonic mean:

We use the following relation $H=\frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Table 3-08: The calculation of the harmonic mean

Salary classes X_i	**Frequency n_1**	**Center of the class c_1**	$\frac{n_i}{x_i}$
]10-20]	**18**	**15**	**1.2**
]20-30]	**30**	**25**	**1.2**
]30-40]	**25**	**35**	**0.71**
]40-50]	**17**	**45**	**0.38**
]50-60]	**12**	**55**	**0.22**
]60-70]	**8**	**65**	**0.12**
	110	**-**	**3.83**

The harmonic mean of the salaries is $H=\frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}=\frac{110}{3.83}=$ 28.72.

Example 3-13:

A driver drove the distance between two cities in 4 equal phases. The total distance is 100 Km. He drives the 1st phase with a speed of 100km/h, the 2nd phase with a speed of 120 km/h, the 3rd phase with a speed of 150 km/h, and the 04th phase with a speed of 80 km/h.

Task: Find the mean of the speed all along the journey.

Solution:

In this case, the arithmetic mean does not give a correct idea. Therefore, we shall use the harmonic mean because it is the best measure of the speed average.

The speed average of the driver all along the journey is:

$H=\frac{100+100+100+100}{\frac{100}{100}+\frac{100}{120}+\frac{100}{150}+\frac{100}{80}}=\frac{400}{1+0,83+0,67+1,25}=\frac{400}{3,75}=$ 106,67 km.

Thus, the average speed of the driver is 106.67 km/h.

2.3.2 The characteristics of the harmonic mean:

- It takes into consideration all the values.
- It is less affected by the extreme values than the arithmetic mean.
- It cannot be calculated from the open frequency distribution tables, and in case of null data.
- It gives more realistic data regarding the speeds and prices.
- Its value is always less than that of the geometric mean. Thus, $\bar{x}$ > G > H.

2.4 The quadratic mean:

It is the quadratic root of the arithmetic mean of the squares of those values. The quadratic mean of the primary and tabulated data is calculated as follows:

2.4.1 Calculating the quadratic mean of the primary data:

If X_1, X_2..........., X_n represent items (values) of a given phenomenon X, the relation of the quadratic mean in case of primary data is given as follows Q= $\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$

Example 3-14:

Find the quadratic mean of the data of this series:

9	5	12	6	7

Solution:

$$Q=\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}=\sqrt{\frac{9^2+5^2+12^2+6^2+7^2}{5}}=\sqrt{\frac{335}{5}}=\sqrt{67}= 8,18$$

2.4.2 Calculating the quadratic mean of the tabulated data:

If X_1, X_2..........., X_n represent items (values) of a given phenomenon X, and n_1, n_2..........., n_k represent the corresponding frequencies, the quadratic mean is defined as follows:

$$Q=\sqrt{\frac{\sum_{i=1}^{n} x_{i.n_i}^{2}}{\sum_{i=1}^{k} n_i}}$$

Where:

X_1 is the values of the variable or the classes' centers.

n_i is the frequencies corresponding to the values of the variable or the classes' centers.

Example 3-15:

To calculate the quadratic mean in case of tabulated data in a frequency distribution table (continuous quantitative table), we take the data of example 3-04.

Solution:

Calculating the harmonic mean:

Table 3-09: The calculation of the quadratic mean

Salary classes X_i	**Frequency n_1**	**Center of the class c_1**	x_i^2	$x_i^2 n_i$
]10-20]	**18**	**15**	**225**	**4050**
]20-30]	**30**	**25**	**625**	**18750**
]30-40]	**25**	**35**	**1225**	**3625**
]40-50]	**17**	**45**	**2025**	**34425**
]50-60]	**12**	**55**	**325**	**36300**
]60-70]	**8**	**65**	**4225**	**33800**
	110	**-**	**-**	**157.950**

The quadratic mean of the salaries is

$$Q=\sqrt{\frac{\sum_{i=1}^{n} x_{i.n_i}^{2}}{\sum_{i=1}^{k} n_i}}=\sqrt{\frac{157950}{110}}=\sqrt{1,435,9}\ =37,89\text{da}.$$

Remark: When calculating the previous means, it is necessary to make sure of this relation:

$$H < G < \overline{X} < Q$$

From the data of example 3-04, and after calculating the previous means, we found out that:

Q= 37 x 10^3 DA, $\overline{X}$= 34.9 x 10^3 DA, G= 31.75 x 10^3 DA, and H= 28.72 x 10^3 DA. Thus, the relation H < G < $\overline{X}$ < Q is correct.

2.5 The median:

It is the value that can be divided by the data that are descendingly or ascendingly ordered into two equal classes; where the number of the values that are bigger than the median equals the number of the values that are smaller than it. It is refereed to with M_e.

2.5.1 Calculating the median:

2.5.1.1 The median of the primary data (untabulated):

To calculate the median of the primary data, we must order the values from up to bottom or from bottom to up. Thus, we have two cases:

- The median of the odd untabulated data:

If the number of the values n is odd, the median is the value whose order is $\frac{N+1}{2}$. Thus, the value of the median is the value whose order is the order of the median M_e=$X_{\frac{n+1}{2}}$

Example 3-16: Find the median of the following data:

8	6	3	10	12	15	9

Solution:

We order the data (values) from bottom to up:

X_1	X_2	X_3	X_4	X_5	X_6	X_7
3	6	8	9	10	12	15

$M_e = X_4 = 9$

The number of the values n= 7 is odd. Thus, the order of the median is $\frac{n+1}{2} = \frac{7+1}{2}$ = 4 and the value of the median is the item (value) whose order corresponds to the order of the median M_e= $X_{\frac{n+1}{2}}$= $X_{\frac{7+1}{2}}$= X_4= 9.

- The median of the even untabulated data:

If the number of the values n is even, the order of the median is $\frac{n}{2},\frac{n}{2}+1$. Thus, the value of the median is the value of the arithmetic mean of the two items whose orders correspond the order of the median, i.e., $M_e = \frac{X_{\frac{n}{2}}+X_{\frac{n}{2}+1}}{2}$

Example 3-17: Find the median of the following data:

8	6	8	10	3	1	13	7

Solution:

We order the data (values) from bottom to up or from up to bottom (it is better from bottom to up):

X1	X2	X3	X4	X5	X6	X7	X8
16	13	10	8	7	6	3	1

$M_e = \frac{7+8}{2} = 7,5$

The number of the values n= 8 is odd. Thus, the order of the median is ($\frac{8}{2},\frac{8}{2}+1 = 4,5$ and the value of the median is the item (value) whose order corresponds to the order of the median $M_e = \frac{X_{\frac{n}{2}}+X_{\frac{n}{2}+1}}{2} = \frac{X_{\frac{8}{2}}+X_{\frac{8}{2}+1}}{2} = \frac{X_4+X_5}{2} = \frac{7+8}{2} = 7,5$

2.5.1.2 The median of the tabulated data:

To calculate the median of the tabulated data, we have two cases:

- In the case of a discrete quantitative variable:

To calculate the median of data tabulated in a frequency distribution table (discrete quantitative variable), we follow these steps:

1. We calculate the ascending cumulative frequency.
2. We identify the order of the median $\frac{N}{2}$where N= $\sum_{i=1}^{k} ni$
3. We look in the line of the ascending cumulative frequency for the value whose ascending cumulative frequency equals the order of the median $\frac{N}{2}$, or is directly higher. Thus, the corresponding value is the median.

Example 3-18:

The following statistical data represent the distribution of the families of a given district according to the number of the rooms.

Number of rooms X_i	1	2	3	4	
Number of families n_i	5	10	20	30	65

Task: find the median of these data.

To find the median of these data, we follow these steps:

1. We calculate the ascending cumulative frequency.

Table 3-10: How to determine the median

Number of rooms X_i	Number of families n_i	n_i
1	**5**	**5**
2	**10**	**15**
3	**20**	**35**
4	**3**	**65**
	65	**-**

The median

The order of the median $\frac{N}{2}$

2. We identify the order of the median $\frac{N}{2} = \frac{65}{2} = 32,5$, where N=$\sum_{i=1}^{k} ni$.

3. We look in the line of the ascending cumulative frequency for the value whose ascending cumulative frequency equals the order of the median, or is directly higher (35 is directly higher than 32.5, i.e., $(35 > \frac{N}{2} = 32,5 > 15$, Thus, the value of the median is the value of X_i that corresponds to the value 35. Therefore, M_e= 3.

- In the case of a continuous quantitative variable:

To calculate the median of data tabulated in a frequency distribution table (continuous quantitative variable), we follow these steps:

1. We calculate the ascending cumulative frequency.

2. We identify the order of the median $\frac{N}{2}$, where N=$\sum_{i=1}^{k} ni$
3. We identify the median class (the class of the median), which is the class that corresponds to the ascending cumulative frequency that equals the order of the median, or is directly higher.

M_e= L$_1$ +$\frac{\frac{N}{2}-N0}{Ne}$ *k

Where:

L$_i$: the minimum limit of the median class.

N: the total frequencies N= $\sum_{i=1}^{k} ni$.

N$_0$: the ascending cumulative frequency of the pre-median class.

n$_e$: the frequency of the median class.

k: the length of the median class.

Example 3-19:

The following table shows the distribution of a group of employees based on their monthly salaries (Unit: 10^3 DA):

Salary X_i	]10-20]	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]
Number of employees n$_i$	18	30	25	17	12	8

Task:

Find the value of the median.

Solution:

To determine the value of the median, we follow these steps:

1. We calculate the ascending cumulative frequency.

The median class

Salary classes X_i	**Frequency n1**	n_i
]10-20]	**18**	**18**
]20-30]	**30**	**48**
]30-40]	**25**	**73**
]40-50]	**17**	**90**
]50-60]	**12**	**102**
]60-70]	**8**	**110**
	110	**-**

The order of the median $\frac{N}{2}$=55

2. We determine the order of the median $\frac{N}{2}$,=$\frac{110}{2}$ =55 where N=$\sum_{i=1}^{k} ni$

3. We identify the median class (the class of the media), which is the class that corresponds to the ascending cumulative frequency that equals the order of the median, or is directly higher (73 is directly higher than the order of the median 55 (37>$\frac{N}{2} = 55 > 48$). Thus, the median class is]30-40].

4. We determine and calculate the median by applying the statistical relation of the median:

$$\boldsymbol{M_e\text{= L1 +}\frac{\frac{N}{2}-N0}{ne}\text{*k = 30 +}\frac{\frac{110}{2}-48}{25}\text{*10 = 32,8 DA.}}$$

2.5.2 The graphical method to find the median:

The graphical median is the cross point in the curves of the ascending and descending cumulative frequencies. Besides, it can be determined using one of the curves of the ascending or descending cumulative frequencies. In this context, we follow these steps:

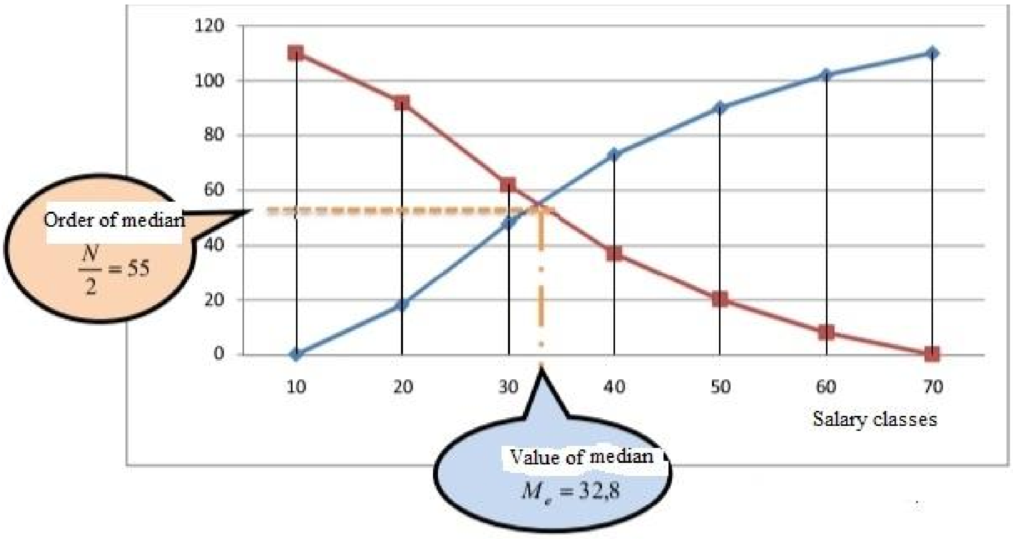
1. We draw the curve of the ascending and descending cumulative frequencies.
2. We determine the order of the median $\frac{N}{2}$ on the Y-axis and draw a horizontal line that starts from the order of the median on the Y-axis until it touches the curve of the ascending or descending cumulative frequencies.

3. We draw a vertical line that starts from the previous cross point and ends at the X-axis. Thus, it the cross point with the X-axis gives the value of the median.

Example 3-20:

To determine the median graphically, we take the data of example 3-04.

Figure 3-01: the curve of the ascending and descending cumulative frequencies to determine the value of the median



2.5.3 The characteristics of the median:

- It neither gets affected by the extreme vales nor accepts the algebraic calculations.
- It can be calculated from the open frequency distribution tables.
- It can be calculated graphically.
- Its calculation does not involve all the values.
- It is determined with the number of the data, not their values.

2.5.4 Quasi medians:

The median is the value that divides the data into two equal classes. Since we can do this, we can deal with the values that divide these data with the same way we deal with the median. The quasi medians are:

2.5.4.1 Quartiles:

They are the values that divide the data into 04 equal parts. The quartile i, where (i= 1, 2, 3) is defined as the value or the item that is preceded by (25% i) of the ascendingly ordered data.

Calculation of the quartiles:

The methods of calculating the quartiles differ than those of the frequency distribution tables. In this context, we have two cases:

- In the case of discrete quantitative variable:

We follow these steps:

1. We order the values of the statistical variables in the table ascendingly, and calculate the ascending cumulative frequencies.
2. We determine the order of the quartile i that is given with the relation $\frac{N_{i\cdot}}{4}$
3. We determine the order of the quartile i from the column of the ascending cumulative frequency. If the order of the quartile i is among the values of the column of the ascending cumulative frequency, the value of the quartile Q_i is the value that corresponds to that frequency. On the other hand, if the value of the quartile i is not among the values of the column of the ascending cumulative frequency, the value of the quartile Q_i is the value that corresponds to the ascending cumulative frequency that is directly higher than the order of the quartile i.

Example 3-21: The following data represent the number of children of a sample of 60 families:

Number of children X_i	0	1	2	3	4	5	
Number of families n_i	8	10	12	14	12	4	60

Task: find the 1st, 2nd, and 3rd quartile.

Solution

Table 3-10: How to determine the quartile

Number of children X_i	Number of families n_i	n_i
0	8	
1	**10**	**18**
2	**12**	**30**
3	**14**	**44**
4	**12**	**56**
5	**4**	**60**
	60	**-**

The 1st quartile Q_1

The 2nd quartile Q_2

The 3rd quartile Q_3

The order of the 1st quartile is 15

The order of the 2nd quartile is 30

The order of the 3rd quartile is 45

The 1st quartile: it is $\frac{iN}{4} = \frac{1*60}{4} = 15$

The table above shows that the value of the 1st quartile does not exist among the values of the ascending cumulative frequencies column. Therefore, we shall look for the value that is directly higher (18 is directly higher than the value of the 1st quartile 15). Therefore, the value of the 1st quartile is $Q_1 = 1$.

The 2nd quartile: it is $\frac{iN}{4} = \frac{2*60}{4} = 30$

The table above shows that the value of the 2ndt quartile is among the values of the ascending cumulative frequencies column. Therefore, the value of the 2nd quartile is the one that corresponds to the order of the 2nd quartile $Q_2 = M_e = 2$.

The 3rd quartile: it is $\frac{iN}{4} = \frac{3*60}{4} = 45$

The table above shows that the value of the 3rd quartile does not exist among the values of the ascending cumulative frequencies column. Therefore, we shall look for the value that is directly higher (56 is directly higher than the value of the 3rd quartile 45). Therefore, the value of the 3rd quartile is $Q_3 = 4$.

- In the case of a continuous quantitative variable:

We follow the same steps of the case of the discrete quantitative variable and use the following equation:

$$Q_{i=}L_1 + \frac{\frac{iN}{4} - N0}{nqi} *k$$

Where:

i= 1, 2, 3

2.5.4.2 The deciles:

They are the values that divide the data into ten equal parts. The decile i is defined, where (i= 1, 2, ….9), as the value or item that is preceded by (10 i%) of the ascendingly ordered data.

Calculating the deciles:

We have two cases:

- In the case of discrete quantitative variable:

We follow these steps:

1. We order the values of the statistical variables in the table ascendingly, and calculate the ascending cumulative frequencies.
2. We determine the order of the decile i that is given with the relation $\frac{iN}{10}$
3. We determine the order of the decile i from the column of the ascending cumulative frequency. If the order of the decile i is among the values of the column of the ascending cumulative frequency, the value of the decile D_i is the value that corresponds to that frequency. On the other hand, if the value of the decile i is not among the values of the column of the ascending cumulative frequency, the value of the decile D_i is the value that corresponds to the ascending cumulative frequency that is directly higher than the order of the decile i.

- In the case of a continuous quantitative variable:

We follow the same steps of the case of the discrete quantitative variable and use the following equation:

$$D_{i=}L_1+\frac{\frac{iN}{4}-N0}{ndi}*k$$

Where:

i= 1, 2,……9

2.5.4.3 The percentiles:

They are the values that divide the data into 100 equal parts. The percentile i is defined, where (i= 1, 2, ….9), as the value or item that is preceded by (i%) of the ascendingly ordered data.

Calculating the percentiles:

We have two cases:

- In the case of discrete quantitative variable:

We follow these steps:

1. We order the values of the statistical variables in the table ascendingly, and calculate the ascending cumulative frequencies.
2. We determine the order of the percentile i that is given with the relation $\frac{iN}{100}$
3. We determine the order of the percentile i from the column of the ascending cumulative frequency. If the order of the percentile i is among the values of the column of the ascending cumulative frequency, the value of the percentile P_i is the value that corresponds to that frequency. On the other hand, if the value of the percentile i is not among the values of the column of the ascending cumulative frequency, the value of the percentile P_i is the value that corresponds to the ascending cumulative frequency that is directly higher than the order of the percentile i.

- In the case of a continuous quantitative variable:

We follow the same steps of the case of the discrete quantitative variable and use the following equation:

$$P_{i=}\ C_i = L_1 + \frac{\frac{iN}{100} - N0}{npi} * k$$

Where:

i= 1, 2,……9

Example 3-22:

The following table shows the distribution of a group of employees based on their monthly salaries (Unit: 10^3 DA):

Salary X_i	]10-20]	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]
Number of employees n_i	18	30	25	17	12	8

Task:

Find the 1st quartile, the 06th decile, and the 85th percentile using the calculations and graphics.

Solution:

To calculate the Quartiles, the decile, and the percentiles, we follow these steps:

Table 3-13: How to determine the class of the quartiles and the percentiles

Salary classes X_i	**Frequency n_1**	n_i
]10-20]	**18**	**18**
]20-30]	**30**	**48**
]30-40]	**25**	**73**
]40-50]	**17**	**90**
]50-60]	**12**	**102**
]60-70]	**8**	**110**
	110	**-**

The 1st quartile class

The 85th percentile class

The order of the 1st quartile

The order of the 85th percentile

Calculating the 1st quartile Q_1:

- We determine the order of the 1st quartile $Q_1 \frac{N}{4} = \frac{110}{4} = 27{,}5$, where N=$\sum_{i=1}^{k} n_i$= 110.
- We determine the class of the 1st quartile Q_1 (the class where it belongs) that is the class that corresponds to the ascending cumulative frequency that equals the order of the 1st quartile, or is directly higher (48 is directly higher than the order of the 1st quartile 27.5 (48>$\frac{N}{4} = 27{,}5 > 18$). Thus, the 1st quartile class is]20-30].

We determine and calculate the 1st quartile using the same statistical relation of the 1st quartile:

$$Q_{i=}\text{L}_1 + \frac{\frac{iN}{4} - N0}{nqi} *\text{k} = 20 + \frac{27{,}5-18}{30} *10 = 23{,}17.$$

- **Calculating the 06th decile:**

We follow the previous steps and find

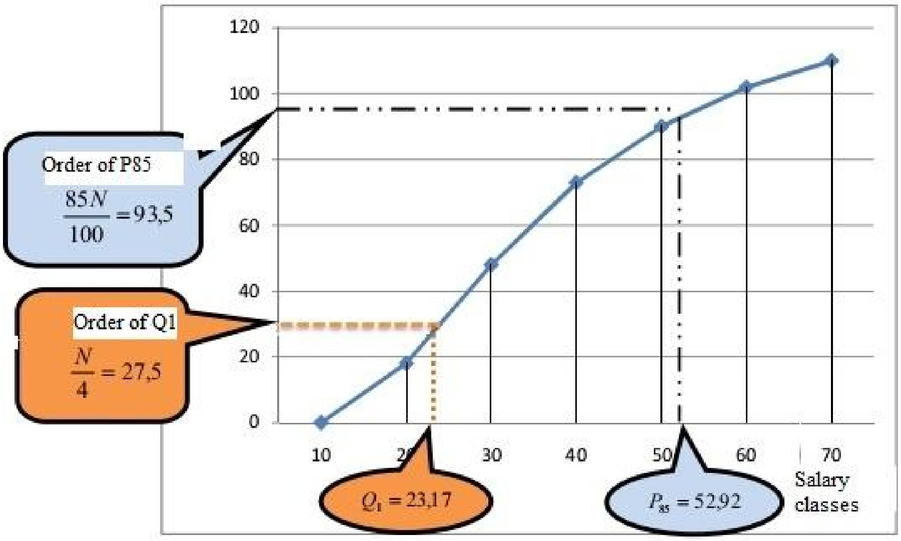
$$D_{6=}\text{L}_1 + \frac{\frac{6N}{4} - N0}{nd6} *\text{k} = 30 + \frac{66-48}{25} *10 = 37{,}2.$$

Calculating the 85th percentile:

We follow the previous steps and find

$$P_{85=}\ \text{L}_1 + \frac{\frac{85N}{100} - N0}{np85} *\text{k} = 50 + \frac{93{,}5-90}{12} *10 = 52{,}92.$$

Finding the previous values graphically:



2.6 The mode:

The value of the mode (M0) expresses the most frequent view. Thus, it is the most common value. The data may have one or more modes, or no mode. It is the best measure to describe the quantitative data.

2.6.1 Methods of calculating the mode:

The methods differ in the case of the primary data than in the case of the tabulated data.

2.6.1.1 The mode of the primary data:

The mode is the most frequent value. Therefore, the mode of a set of data may be more than one value or quality.

Example 3-23:

The following data show the grades of 10 students:

Excellent, good, very good, good, average, above the average, good, weak, very good, good.

Task: find the mode of these data.

Solution:

We see that the most frequent grade is “good”. Therefore, the mode is the grade “good” (M_0= good).

Example 3-24:

X is a statistical variable that represents the number of rooms of houses in a given district. Find the mode

3	4	2	5	4	3	3	5	2	3

Solution:

The mode is the most frequent value. Here, 3 is the most frequent (it is repeated 04 time). Therefore, the mode is (M_0 =3).

2.6.1.2 The mode of the tabulated data:

The mode is the most frequent value. In this context, we have two cases:

- In case of a discrete quantitative variable:

The mode is the value of the variable x_i that corresponds to the highest frequency in the statistical distribution table.

Example 3-25:

The table shows the number of the absences of employees in a given company.

M_0 = 3

Number of days X_i	0	1	2	3	4	5
Number of employees n_i	22	18	15	25	12	8

The most frequent

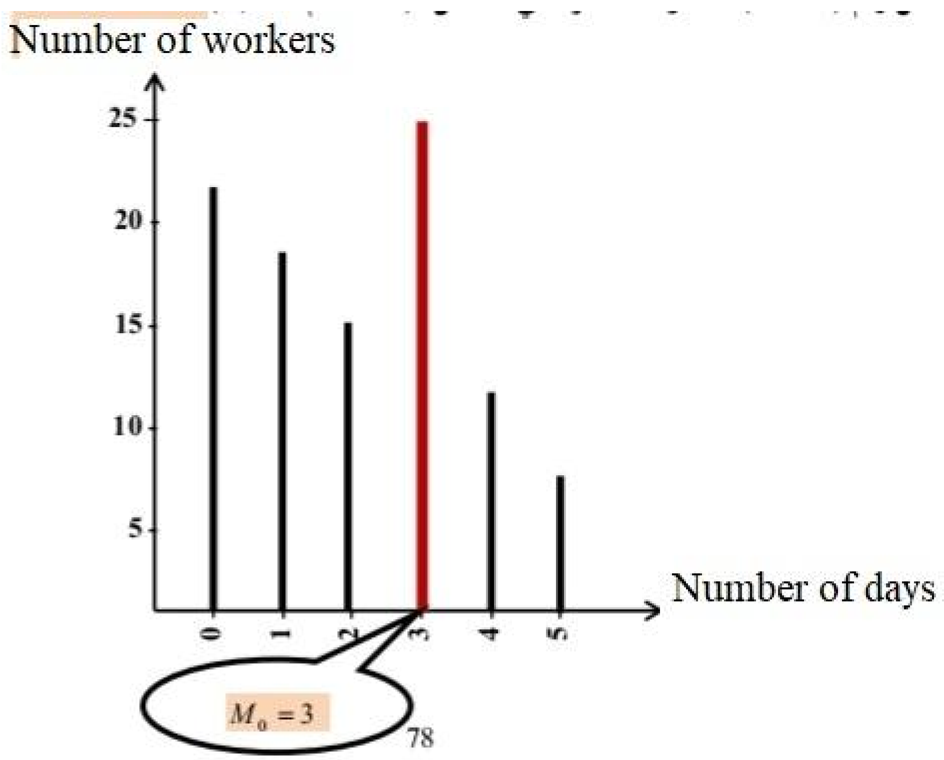
Task: find the mode with calculation and with graphics.

The mode with calculations:

The table shows that the biggest frequency is 25. Thus, the value of the mode is $M_0 = 3$. This means that the most common absence among the employees is 03 days.

The mode with graphics:

Figure 3-03: The mode of a discrete quantitative variable (days of absence):



- In case of a continuous quantitative variable:

To find the mode of data tabulated in a frequency distribution table (continuous quantitative variable), we determine the mode class and then calculate the mode. Here, we must focus on the lengths of the classes. If they are not equal, we modify the frequencies so that the classes become consistent. In this context, we have two cases:

- Classes with equal lengths: In this case, the mode class is the one that corresponds to the highest frequency n_i. The mode is calculated using this relation:

$$M_o = L1 + \frac{\Delta 1}{\Delta 1 + \Delta 2} * k$$

Where:

L_1: is the minimum limit of the mode class.

Δ1 is the difference between the frequency of the mode class and the frequency of the previous class.

Δ2 is the difference between the frequency of the mode class and the frequency of the next class.

k: is the length of the mode class

- Classes with unequal lengths: Here, we modify the frequencies. The mode class is the class that corresponds to the biggest modified frequency (n_i*). The mode is calculated using the same previous relation (the modified frequency is used instead of the absolute frequency).

M_o= L1 +$\frac{\Delta 1}{\Delta 1+\Delta 2}$* k

Where:

L_1: is the minimum limit of the mode class.

Δ1 is the difference between the modified frequency of the mode class and the modified frequency of the previous class.

Δ2 is the difference between the modified frequency of the mode class and the modified frequency of the next class.

k: is the length of the mode class

2.6.2 Determining the mode graphically:

We can determine the mode by drawing the frequency polygons of the mode class and the previous and the following classes. We link the start of the rectangle of the mode class with the start of the rectangle of the following class, and the end of the rectangle of the mode class with the end of the rectangle of the previous class. Then, from the cross point, we draw a bar that goes down to the X-axis. The cross point with this axis gives the value of the mode.

Example 3-26: To find the mode with calculations and graphics, we consider the data of example 3-04:

Table 3-14: How to determine the mode class

Salary classes X_i	**Frequency n_1**
]10-20]	**18**
]20-30]	**30**
]30-40]	**25**
]40-50]	**17**

The mode class

The highest frequency

]50-60]	**12**
]60-70]	**8**
	110

Solution:

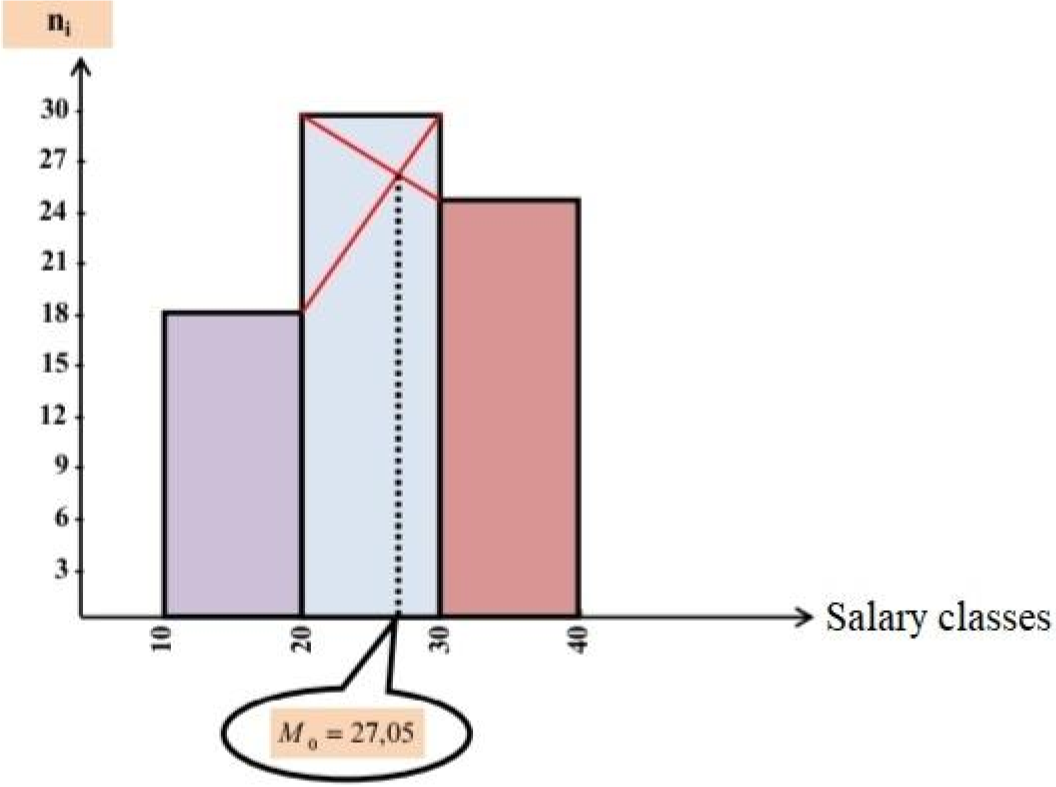
The table shows that the classes have equal lengths. Therefore, we follow these steps to calculate the mode:

- **We determine the mode class**: It is the class that corresponds to the biggest frequency (the biggest frequency is 30; therefore, the mode class is] 20-30]).

- **Calculating the mode:** We calculate the mode using this relation:

$$M_o = \text{L1} + \frac{\Delta 1}{\Delta 1 + \Delta 2} * \text{k} = 20 + \frac{(30-18)}{(30-18)+(30-25)} * 10 = 27{,}05.$$

Figure 3-04: A graphic representation that determines how to find the mode



2.6.3 The characteristics of the mode:

- It is the easiest measure of the central tendency.
- It neither gets affected by the extreme values nor accepts the algebraic calculations.
- It can be graphically calculated.
- It can be calculated from the open frequency distribution tables.
- It is the best mean to describe the qualitative phenomena.

3. The relation between the measures of the central tendency (the arithmetic mean, the median, and the mode):

If a set of data have one mode; the mean, the median, and the mode have one of these relations:

1. If the frequency distribution is symmetric:

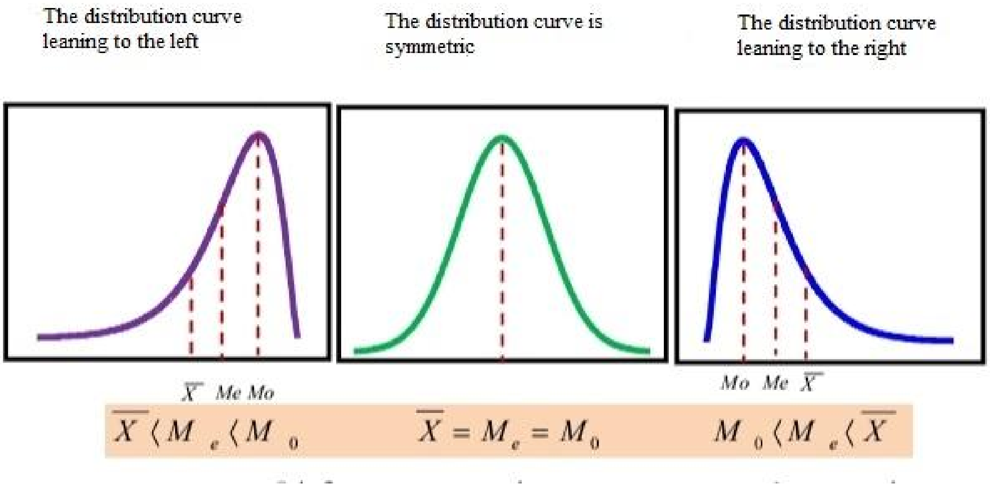
Here, the curve is symmetric with one peak and a shape that resembles the bell. In this case, $\overline{X} = M_e = M_0$. If the curve of the frequency distribution is a bit skewed, the following relation is correct $\overline{X} - M_0 = 3\ (X - M_e)$. This relation helps find the arithmetic mean if one of the sides of the frequency distribution table is open.

2. If the mean is more than the median and the mode:

Here, the curve of the distribution is skewed towards the right, and $M_0 < M_e < \overline{X}$

3. If the mean is less than the median and the mode:

Here, the curve of the distribution is skewed towards the left, and $M_0 > M_e > \overline{X}$



The distribution curve leaning to the left

The distribution curve is symmetric

The distribution curve leaning to the right

$\overline{X} \langle M_e \langle M_0$ $\qquad$ $\overline{X} = M_e = M_0$ $\qquad$ $M_0 \langle M_e \langle \overline{X}$

In order to know the shape of the frequency distribution practically, we take the data of example 3-04 whose measures of central tendency were calculated. Thus, we have:

M_0= 27.05 DA $\qquad$ M_e= 32.8 DA $\qquad$ $\overline{X}$= 34.9 DA

Since $M_0 < M_e < \overline{X}$, the distribution curve leans to the right.

Activities of Chapter 3:

Solved activities:

Activity 01:

This table shows the working days of the workers of a construction site in January:

Number of days X_i	10	15	18	20	22	
Number of workers n_i	10	12	8	6	4	40

Task:

1. Find the study variable and its nature.
2. Determine the mean of working days of the workers in January.
3. Find the median and the mode.

Solution:

1. The study variable is the number of the working days of the workers in January. Is nature is discrete qualitative.

2. The mean of working days of the workers:

Number of days X_i	Frequency n_i	$x_i n_i$	n_i	n_i
10	10	100	**10**	40
15	**12**	180	**22**	3
18	8	144	30	18
20	6	120	36	10
22	4	88	40	4
	40	632	-	-

The mean of the working days is:

$$\bar{x}=\frac{\sum_{i=1}^{k} xi*ni}{\sum_{i=1}^{k} n_i}=\frac{\sum_{i=1}^{5} xi*ni}{\sum_{i=1}^{5} n_i}=\frac{632}{40}=15{,}8 \sim 16.$$

Thus, the workers worked 16 days in average in January.

3. Calculating the median and the mode:

The median:

Since the variable is discrete quantitative, we follow these steps:

- We determine the order of the median $\frac{N}{2}=\frac{40}{2}=20$ where N= $\sum_{i=1}^{k} ni$= 40.

- We notice that the value 20 (order of the median) is not in the column of the ascending cumulative distribution. Thus, we look for the direct higher value, that is 22 (22>$\frac{N}{2}$=20>10). Therefore, the value of the median is the value X_i that corresponds to the value 22. As a result, the value of the median is M_e =15.

- Finding the mode:

Since the biggest frequency is 12, the value of the mode is the value that corresponds to it. Therefore, M_0 = 15.

Activity 02:

The table shows the distribution of a sample of 100 families based on their monthly expenditures (Unit: 10^3 DA):

Expenditure classes X_i	]20-25]	]25-35]	]35-40]	]40-55]	]55-75]	]75-80]	
Frequency n_i	5	15	20	25	30	5	100

Task:

1. Determine the study variable and its nature.
2. Determine the mean expenditure of the families.
3. Find the median and the mode with calculations and graphics.
4. What is the shape of the statistical distribution using the measures of the central tendency?
5. Prove with calculations that M_e and D_5 equal Q_3 and C_{75}.

Solution of activity 02:

1. The study variable is the monthly expenditures. Is nature is discrete qualitative.

2. The mean of the monthly expenditures:

Expenditure classes X_i	Frequency n_i	Center of the class x_i	$x_i n_i$	n_i	Length of the class	Modified frequency n_i*
]20-25]	5	22.5	112.5	5	5	$n_1^*=\frac{5}{5}*5$ =5
]25-35]	15	30	450	20	10	$n_2^*=\frac{15}{10}*5$ =7,5
]35-40]	20	37.5	750	40	5	$n_3^*=\frac{20}{5}*5$ =20
]40-55]	25	47.5	1187.5	65	15	$n_4^*=\frac{25}{15}*5$ =8,33
]55-75]	30	65	1950	95	20	$n_5^*=\frac{30}{20}*5$ =7,5
]75-80]	5	77.5	387.5	100	5	$n_6^*=\frac{5}{5}*5$ =5
	100	-	4837.5	-	-	—

The mean expenditure of the families is:

$$\overline{x}=\frac{\sum_{i=1}^{k} xi*ni}{\sum_{i=1}^{k} n_i}=\frac{\sum_{i=1}^{6} xi*ni}{\sum_{i=1}^{6} n_i}=\frac{4837,5}{100}=48,375.$$

3. Finding the median and the mode with calculations and graphics:

- With calculations:

We follow these steps:

- We determine the order of the median $\frac{N}{2}$,=$\frac{100}{2}$ =50 where N= $\sum_{i=1}^{k} n_i$= 100.

- We determine the median class: The order of the median is 50. It does not exist among the values of the ascending cumulative frequency column. Therefore, we determine the directly higher value whose corresponding value shall be the median class (65> $\frac{N}{2}$ =50> 40)

Thus, the median class is 40-55.

- We determine and calculate the median using this statistical relation

$$\boldsymbol{M_e}= \mathbf{L_1} +\frac{\frac{N}{2}-N0}{ne}*\mathbf{k} = \mathbf{40} +\frac{\frac{100}{2}-40}{25}*\mathbf{15 = 46.}$$

- The mode with calculations:

Since the classes do not have equal lengths, we modify the frequencies. We take the chosen length of the class that equals 05 as a basis for modifying the frequencies (the smallest class length).

- Determining the mode class:

It is the class that corresponds to the biggest modified frequency (the biggest modified frequency is 20). Thus, the mode class is]35-40]

M_o= L1+$\frac{\Delta 1}{\Delta 1+\Delta 2}$*k =35+$\frac{(20-7{,}5)}{(20-7{,}5)+(20-8{,}33)}$*5= 37,59

Where L_i = 35 is the minimum limit of the mode class.

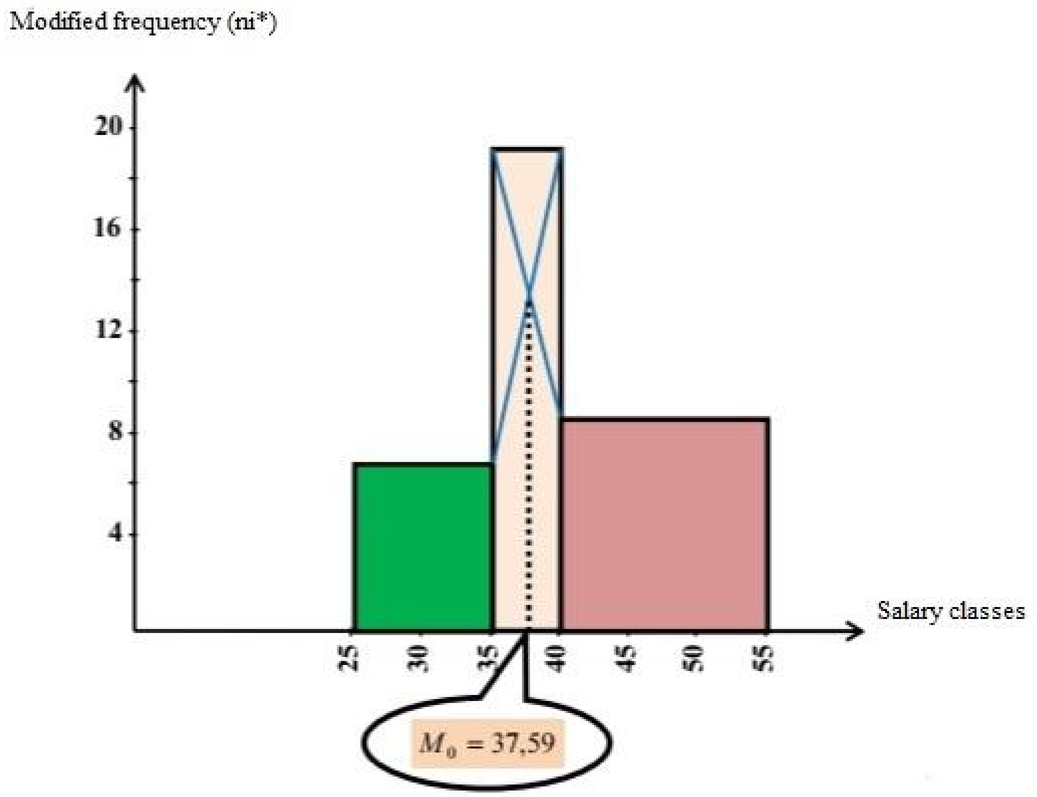
Δ1= (20- 7.5) = 12.5 is the difference between the modified frequency of the mode class and the modified frequency of the previous class.

Δ2 = (20 – 8.33) = 11.67 is the difference between the modified frequency of the mode class and the modified frequency of the next class.

k: is the length of the mode class

- The mode with graphics:

Figure 3-05: Graphic representation that shows how to find the mode:



4. Determining the shape of the statistical distribution:

Because the mean is higher than the median and the mode $M_0 < M_e < \overline{X}$, the curve of the statistical distribution leans to the right.

5. Calculating C_{75}, Q_3, D_5, M_e:

$M_e = 46$.

-Calculating the 5th decile D_5:

We follow these steps:

We determine the order of the 5th decile $\frac{5N}{10} = \frac{5*100}{10} = 50$, where $N = \sum_{i=1}^{k} ni$.

We determine the class of the 5th decile: The order of D_5 is 50. It does not exist among the values of the ascending cumulative frequency column. Therefore, we determine the directly higher value whose corresponding value shall be the D_5 class $(65 > \frac{5N}{10} = 50 > 40)$

Thus, the D_5 class is 40-55.

- We determine and calculate the D_5 using this statistical relation

D_5= L1 +$\frac{\frac{5N}{10}-N0}{n_{D5}}$ *k = 40 +$\frac{\frac{500}{10}-40}{25}$ *15 = 46. DA.

As a result: $M_e = D_5$.

-Calculating the 3rd quartile Q_3:

We determine the order of the 3rd quartile $\frac{3N}{4}$,= $\frac{3*100}{4}$ = 75 where N= $\sum_{i=1}^{k} ni.$=100

We determine the class of the 3rd quartile: The order of Q_3 is 75, i.e. (95> $\frac{3N}{4} = 75 > 65$). Thus, the class of Q3 is]55-75].

-We calculate the 3rd quartile with the following statistical relation:

Q_3= L1 +$\frac{\frac{3N}{4}-N0}{n_{Q3}}$ *k = 55 +$\frac{\frac{300}{4}-65}{30}$ *20 = 61,67. DA.

Calculating C_{75}: We follow the same steps and get:

C_{75}= L1 +$\frac{\frac{75N}{100}-N0}{n_{c75}}$ *k = 55 +$\frac{\frac{7500}{100}-65}{30}$ *20 = 61,67 DA. Thus, $Q_3 = C_{75}$.

Suggested activities:

Activity 01:

Consider this statistical series: 32, 25, 237, 32, 28, 33, 32, 3, 29, 33, 26.

1- Find the arithmetic mean (with the direct method and the hypothetical mean method), the geometric mean, the quadratic mean, and the harmonic mean. What do you conclude?

2. Find the median and the mode.

Activity 2:

Consider these tabulated values:

Value X_i	3	4	5	6	7
Frequency n_i	13	16	20	17	14

Task:

1. Find the mean, the median, and the mode.

2. Find the 1st, 2nd, and 3rd quartile; the 5th decile, and the 25th and the 50th percentiles. What do you conclude?

Activity 03:

The table shows the production of shoes of various sizes.

Size X_i	40	41	42	43	44	45	46
Number of pairs n_i	300	800	500	300	200	100	10

Task:

1. Find the mean.

2. Find the median and the mode with calculation and with graphics.

Task 04:

The table shows the distribution of a sample of families according to their expenditures in a given month (Unit: 10^3 DA):

classes X_i	]10-15]	]15-20]	]20-25]	]25-30]	]30-35]	]35-40]	
Number of families n_i	10	15	20	15	10	5	75

Task:

1. Determine the study variable and its nature.
2. Determine the mean expenditure of the families.
3. Find the median and the mode with calculations and graphics.
4. What is the shape of the statistical distribution using the measures of the central tendency?
5. Find D_5 and C_{50}.

Activity 05:

The table shows the monthly salaries in a given company (Unit: 10^3 DA):

classes X_i	]60-70]	]70-80]	]80-e_4]	] e_4-100]	]100-110]	]110-120]	
n_i	12	10	18	6	10	4	60

Task:

1. Determine the study variable and its nature.
2. What is the value of the unknown limit e_4 if the median salary is M_e= 84.445.
3. Find the arithmetic mean.
4. Find the common salary and present it graphically.
5. If the administration decided to fire 10% of the employees who get high salaries, what is the maximum limit of the salaries the company will be paying?

Activity 06:

The table shows the size of the exports of a set of companies in a given year (Unit: million USD):

Size of exports 10^6 USD	Less than 20	]20-35]	]35-55]	]55-75]	]75-1000]	+100
Number	10	14	18	20	10	8

of companies						

- What is the suitable measure of the central tendency for this distribution? Determine it with calculation and with graphics.

Activity 07:

An industrial company employs 150 workers. 90 of them get an average monthly salary of 33500 DA, 40 of them get an average monthly salary of 39000 DA, while the rest of them get an average monthly salary of 51000.

Task:

1. Find the mean of the salary.
2. If the company decides to hire more 20 workers with a salary mean that equals the mean of the previous salaries, what shall be the mean of the monthly salaries in the company?

Chapter 04:

Measures of dispersion

1. Definition of dispersion

2. Measuring data dispersion

2.1 The general range

2.2 Quadratic range

2.3 Mean deviation

2.4 Variation and standard deviation.

2.5 Measures of the relative dispersion

Activities of chapter 04

Introduction:

The measures of the central tendency are not enough to give a full image about the relation between the data because we may find that two different series have the same arithmetic mean and a different range. This difference makes it necessary to use other measures, namely the dispersion measures that complement the first measures. They are statistical measures that measure the dispersion and distance between the data. They are important because we cannot imagine equal production in all the industrial companies, equal services in the service departments, or equal lengths of people, etc. Therefore, the use of one value to describe the frequency distribution may be misleading sometimes.

Example: Consider these two statistical series:

The 1st series	0	14	7
The 2nd series	7	6	8

The arithmetic mean of the two series is 7. If we stop at this measure, we shall say that the two series are similar. However, the values of the 1st series are farther than those of the 2nd in reality. Here, the role of the dispersion measures appears.

1. Meaning of dispersion:

The data dispersion is the degree of variance or difference between the items of the phenomenon. The data of the phenomenon are harmonious when their values are close to each other. In this case, we say that the data are not dispersed. On the other hand, if the data are not harmonious, the values are not centralized and are dispersed.

2. Measuring the data dispersion:

There are some measures that measure the convergence or distance between data such as the range and the quadratic deviation. Besides, other measures, namely the mean deviation and the standard deviation, measure the convergence or distance between the values and a given value such as the arithmetic mean.

2.1 The range:

It is used when we want to have a fast measure to the dispersion of the values, without much focus on exactness, or when the extreme items (values) have special importance. The range of a set of data is the difference between the maximum value in the data and the minimum value. It is referred to with R and is given with this relation:

The range = the maxim value – the minimum value

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

Exception: The range in case of a frequency distribution with classes can be measured with various methods.

The range = the position of the last class – the position of the first class

$R = C_k - C_1$

The range = the maximum limit of the last class – the minimum limit of the first class

$R = U_k - L_1$

Example 4-01:

Find the range of these data:

12	18	28	22	30

Solution:

The range = the maxim value – the minimum value

$R = X_{\max} - X_{\min} = 30 - 12 = 8$

- The characteristics of the range:

- It is easy to calculate and understand, and relies on two values only.
- It is much affected by the extreme values.
- It cannot be calculated from the open frequency distribution tables.

2.2 The interquartile range:

To get rid of some defects of the range, mainly its influence by the extreme values and the impossibility of calculating it in case of open frequency distribution tables, we use the interquartile range that neglects the 1st and last quarters of the ascendingly ordered data. In this case, the maximum value of the data is the third quartile and the minimum is the 1st quartile. The difference between them gives the interquartile range. This measure is used if the median is the suitable measure for the central tendency, or when there are very extreme values. It is calculated from the primary data or the frequency distribution tables using this relation:

The quadratic range = $I_q = Q_3 - Q_1$

Example 4-02: Find the quadratic range of these data:

Salary X_i	]10-20]	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]
Number of workers n_i	18	30	25	17	12	8

Solution:

Table 4-01: The method of determining the class of the 1st and 3rd quartiles

Salary classes X_i	**Frequency n_1**	n_i
]10-20]	**18**	**18**
]20-30]	**30**	**48**
]30-40]	**25**	**73**
]40-50]	**17**	**90**
]50-60]	**12**	**102**
]60-70]	**8**	**110**
	110	**-**

The order of the 1st quartile

The order of the 3rd quartile

- Calculating the 1st quartile Q_1:

$$Q_1 = L1 + \frac{\frac{N}{4} - N0}{n_{Q1}} * k = 20 + \frac{27{,}5 - 18}{30} * 10 = 23{,}17 \text{ DA.}$$

- Calculating the 2nd quartile Q_2:

$$Q_3 = L1 + \frac{\frac{3N}{4} - N0}{n_{Q3}} * k = 40 + \frac{82{,}5 - 73}{17} * 10 = 45{,}59 \text{ DA.}$$

The interquartile range:

I_q=Q_3-Q_1= 45,59 - 23,17 = 22,42.

- The characteristics of the interquartile range:

- It is not affected by the extreme values and can be calculated graphically.
- It can be calculated from the open frequency distribution tables.
- It is determined by the number of data, not their values.
- It is used as a measure of dispersion in the much skewed frequency distributions.
- It is a period that includes 50% of the data.

2.3 The mean deviation:

The dispersion measures are measures of a force that gathers the data around themselves. The gathering may be around an average value. If the difference (deviation) between the values and their means is big, the data are not harmonic, and vice versa.

2.3.1 Definition of the mean deviation:

It is the arithmetic mean of the extreme values of the deviation of the values from their arithmetic means. We refer to it with E_x.

2.3.2 Calculating the mean deviation:

2.3.2.1 Calculating the mean deviation from the primary data:

If X_1, X_2,…..X_n are items (values) of a given phenomenon X, the mean deviation is given with this relation:

$$E_{\bar{x}}=\frac{|x_1-\bar{x}|+|x_2-\bar{x}|+..........+|x_n-\bar{x}|}{n}=\frac{\sum_{i=1}^{n}|x_{i-\bar{x}}|}{n}$$

Example 4-03: Find the mean deviation of these data:

2	3	5	6	9

Solution:

The arithmetic mean of these data is $\bar{x}=\frac{\sum_{i=1}^{n}xi}{n}=\frac{25}{5}=5$

X_i	2	3	5	6	9	

$\lvert x_i - \bar{x} \rvert$	3	2	0	1	4	10

Thus, the mean deviation of these data is $E_{\bar{x}} = \frac{\sum_{i=1}^{n} \lvert x_{i-\bar{x}} \rvert}{n} = \frac{10}{5} = 2$

2.3.2.2 Calculating the mean deviation from the tabulated data:

If X_1, X_2,…….X_k are items (values) or classes centers of a given phenomenon X, and if n_1, n_2,…..,n_k represent the corresponding frequencies:

$$E_{\bar{x}} = \frac{\lvert x_1 - \bar{x} \rvert\, n_1 + \lvert x_2 - \bar{x} \rvert n_2 + \text{……….} + \lvert x_n - \bar{x} \rvert n_k}{\sum_{i=1}^{k} n_i}$$

$$E_{\bar{x}} = \frac{\sum_{i=1}^{n} \lvert x_{i-\bar{x}} \rvert n_i}{\sum_{i=1}^{k} n_i}$$

Where:

X_1 is the values of the variable or the classes' centers.

n_i is the frequencies corresponding to the values of the variable or the classes' centers.

Example 4-04:

Find the dispersion of the data tabulated in this table using the mean deviation:

Salaries X_i	]10-20]	]20-30]	]30-40]	]40-50]	]50-60]	]60-70
Number of workers n_i	18	30	25	17	12	8

Solution:

Table 4-02: Finding the mean deviation

Salary	**Frequency**	**Center of**	$x_i n_i$	$\lvert x_{i-\bar{x}} \rvert$	$\lvert x_{i-\bar{x}} \rvert n_i$

classes X_i	**n_i**	**the class c_i**			
]10-20]	**18**	**15**	**270**	**19.9**	**358.2**
]20-30]	**30**	**25**	**750**	**9.9**	**297**
]30-40]	**25**	**35**	**875**	**0.1**	**2.5**
]40-50]	**17**	**45**	**765**	**10.1**	**171.1**
]50-60]	**12**	**55**	**660**	**20.1**	**241.2**
]60-70]	**8**	**65**	**520**	**30.1**	**240.8**
	110	**-**	**3840**	**-**	**1310.8**

The arithmetic mean

$$\overline{x}=\frac{\sum_{i=1}^{k} x_i * n_i}{\sum_{i=1}^{k} n_i}=\frac{\sum_{i=1}^{4} x_i * n_i}{\sum_{i=1}^{4} n_i}=\frac{3840}{110}=34,9 \text{ DA.}$$

The arithmetic deviation

$$E_{\overline{x}} = \frac{\sum_{i=1}^{n} |x_{i-\overline{x}}| n_i}{\sum_{i=1}^{k} n_i} = \frac{1310,8}{110} = 11,92.$$

2.3.3 Characteristics of the mean deviation:

- It gets affected by the extreme values and relies on all the values in its calculation.
- It cannot be calculated in case of the open frequency distribution tables.
- It can be calculated through the deviations from the median, with remark that the mean deviation from the median is less than the mean deviation from the arithmetic mean.

2.4 The standard deviation and variance:

2.4.1 The variance:

It is the arithmetic mean of the squares of the differences between the values of the statistical variable and their arithmetic mean. We use the squares of the differences to avoid using the extreme values as in the mean deviation. The variance is referred to with σ^2 or *v(x)* in case of the population data, and s^{2-} in case of a sample. The relation of the variance for the population is $\sigma^2 = \frac{\sum (x_i-\mu)2}{N}$

Sometimes, the number 01 is subtracted from the items (n-1) or the total of the frequencies when estimating the variance measure or the standard deviation of the small samples. The number 01 represents the freedom degree (the sample is big if its size exceeds 30 views). To calculate the variance, we have two cases:

2.4.1.1 Calculating the variance from the primary data:

- The variance of the population:

If the values X_1, X_2,…….X_k represent data of a given population, the variance of the society is given as follows:

$$\sigma^2 = \frac{\sum (x_i-\mu)2}{N} = \frac{\sum (x_i 2 - N\mu 2)}{N} = \frac{\sum x_i\ 2}{N} - \mu^{\ 2}$$

Where μ is the arithmetic mean of the society; i.e., $\mu = \frac{\sum x_i}{N}$

- The variance of the sample:

If the values X_1, X_2,…….X_k represent data of a given sample and $\overline{X}$ is the arithmetic mean, the variance of the sample is given as follows:

$$S^2 = \frac{\sum x_i - \bar{x}^2}{n-1} = \frac{1}{n-1}\left(\sum x_i^2 - n\overline{x^2}\right) = \frac{1}{n-1}\left(\sum x_i^2 - \frac{\sum xi^2}{n}\right)$$

Example 4-05: Find the variance of the following data:

2	3	5	6	9

Solution: The arithmetic mean of these data is $\overline{X}$= 5.

X_i	2	3	5	6	9	
$(X_i-\overline{X})^2$	9	4	0	1	16	30

Thus, the variance of the data is $S^2 = \frac{\sum (x_i-\bar{x})2}{n-1} .= \frac{30}{4} = 7,5$.

2.4.1.2 Calculating the variance from the tabulated data:

- The variance of the population:

If the values X_1, X_2,…….X_k represent items (values) or classes centers of the data of a given population (whose size is N item), and n_1, n_2,………,n_k are the corresponding frequencies, the variance of the society is given as follows:

$$\sigma^2 = \frac{\sum (x_i-\mu)2n_i}{N} = \frac{\sum (x_i 2 - n_i)}{N} - \mu^{\ 2}$$

- The variance of the sample:

If the values X$_1$, X$_2$,…….X$_k$ represent items (values) or classes centers of the data of a given population (whose size is N item), n$_1$, n$_2$,………,n$_k$ are the corresponding frequencies, and $\overline{X}$ is the arithmetic mean, the variance of the sample is given as follows:

$$S^2 = \frac{\sum x_i - \bar{x})^2 n_i}{n-1} = \frac{1}{n-1}\left(\sum x_i^2 - n\overline{x^2}\right) = \frac{1}{n-1}\left(\sum x_i^2 n_i - \frac{\sum (xi * n_i)^2}{n}\right)$$

2.4.2 The standard deviation:

It is the square root of the variance. It is among the most important statistical measures of the dispersion and the most used in the statistical laws and theories. It is referred to with σ in the case of a population and S is case of a sample.

To show how we calculate the standard deviation and the variance, we take this example:

Salaries *X*$_i$	]10-20]	]20-30]	]30-40]	]40-50]	]50-60]	]60-70
Number of workers *n$_i$*	18	30	25	17	12	8

Task: Find the dispersion of the workers' salaries.

Solution:

The best measure for the dispersion is the standard deviation. To calculate the standard deviation, we calculate the arithmetic mean and, then, the variance. Finally, we find the standard deviation.

Table 4-02: Finding the variance

Salary classes *X*$_i$	**Frequency n$_1$**	**Center of the class c$_1$**	***x*$_i$n$_i$**	***x*$_i$^{2}n$_i$**	**$(x_i$ -$\bar{x})$**	**$(x_i$ - $\bar{x})n_i$**
]10-20]	**18**	**15**	**270**	**4.050**	**-19.9**	**7182.18**
]20-30]	**30**	**25**	**750**	**18.750**	**-9.9**	**2940.3**
]30-40]	**25**	**35**	**875**	**3.0625**	**0.1**	**0.25**
]40-50]	**17**	**45**	**765**	**34.425**	**10.1**	**1734.17**
]50-60]	**12**	**55**	**660**	**36.300**	**20.1**	**4848.12**
]60-70]	**8**	**65**	**520**	**33.800**	**30.1**	**7248.08**
	110	**-**	**3840**	**157.950**	**-**	**23899.1**

The arithmetic mean

$$\overline{\boldsymbol{x}} = \frac{\sum_{i=1}^{k} x_i * n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{6} x_i * n_i}{\sum_{i=1}^{6} n_i} = \frac{3840}{110} = \textbf{34,91 DA.}$$

Finding the variance:

- Method 01:

$$S^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})2}{n-1} = \frac{23899{,}1}{109} = \textbf{219,26.}$$

- Method 02:

$$S^2 = \frac{1}{n-1} (\sum x_i^2\ n_i - n\overline{x}^{\,2} = \frac{1}{109} (157{,}950 - 110\ (34{,}91)^2 = 219{,}26.$$

Finding the standard deviation: S $=\sqrt{S^2} = \sqrt{219{,}26} =$14,81.

- Characteristics of the standard deviation:

- It cannot be calculated from the open frequency distribution tables.
- It gets affected by the extreme values.
- It accepts the algebraic calculations. Therefore, it is much used in the statistical laws and theories.
- It takes the same measure unit of the original variable.
- It can be used in comparing the dispersion of two statistical samples of the same quality and arithmetic mean.

2.5 Measures of the relative dispersion:

They are measures used to compare two different sets of data in measurement units. They are free of the measurement units and can be described as relative measures. The most important are:

2.5.1 The relative difference coefficient:

It is the percentage of the standard deviation attributed to the arithmetic mean. The more the value of the coefficient is big, the bigger is the dispersion between the distribution items; and vice versa. It is referred to with (C.V) and is calculated as follows:

The data of the population: C.V=$\frac{\sigma}{\mu}$*100

The data of the sample: C.V=$\frac{S}{\overline{x}}$*100

Example 4-07:

Find the difference coefficient of two groups of students' grades in statistics, as shown in the table:

	Group A	Group B
Arithmetic mean	40	33
Standard deviation	24	16

Solution:

- We determine the difference coefficient of group A: C.V$_A$ $=\frac{S}{\bar{x}}$ *100 $==\frac{24}{40}$ *100 = 0,6 *100 =60%

- We determine the difference coefficient of group B: C.V$_B$= $\frac{S}{\bar{x}}$ *100 $==\frac{16}{33}$ *100 = 0,48 *100 =48%

Through the comparison, we find that the dispersion in group A is bigger than the dispersion in group B that is more harmonious and less farther from its mean value.

- The characteristics of the relative difference:

- It measures the relative difference without a distinction unit.
- It does not have a meaning if the arithmetic means are null.
- It is used to compare two or more sets of data regarding the dispersion, mainly if their arithmetic means differ.

2.5.2 The coefficient of the quartile difference:

It is used in case of an open frequency distribution. It is referred to with C.Q.V. It is given with this relation

C.Q.V= $\frac{Q_3 - Q_1}{Q_2} * 100$

Example 8-04:

We take the data of example 4-02 and calculate the coefficient of the quartile difference.

Solution:

After calculating the 1st, 2nd, and 3rd quartiles, we find that:

Q_3= 45.59 . DA Q_2= 32.8 . DA Q_1= 23.17 . DA

Thus, we calculate the quartile difference coefficient through this relation:

$$C.Q.V= \frac{Q_3 - Q_1}{Q_2} * 100 = \frac{45,59-23,17}{32,8} *100 = 68,35\%$$

Conclusion: The dispersion of these data is big because the value of the quartile difference coefficient equals 68.35%.

Activities of chapter 04:

Solved activities

Activity 01:

If X_1, X_2,…….X_k represent values of a given phenomenon X, prove that:

- $V(x) = \frac{1}{n} \sum (x_i - \overline{x})^2 = \frac{\sum x_i^2}{n} - \overline{x}^2$
- $S^2 = \frac{\sum (x_i - \overline{x})2}{n-1} = \frac{\mathbf{1}}{\mathbf{n-1}} (\sum \boldsymbol{x_i^2}\ \boldsymbol{n_i} - \mathbf{n}\boldsymbol{\overline{x}}\ ^{\mathbf{2}}) = \frac{\mathbf{1}}{\mathbf{n-1}} (\sum \boldsymbol{x_i^2}\ - \frac{\sum \boldsymbol{x_i}\ \mathbf{2}}{\boldsymbol{n}}$

Solution:

We have $V(x) = \frac{1}{n} \sum (x_i - \overline{x})^2 = \frac{1}{n} \sum (x_i 2 - 2x_i \overline{x} + \sum \overline{x}\ ^2$

$= \frac{\sum x_i^2}{n} - 2\overline{x} * \frac{\sum x_i}{n} + \frac{\sum \overline{x}_i^2}{n} = \frac{\sum x_i^2}{n} - 2\ \overline{x} * \overline{x} + \frac{n\overline{x}2}{n} = \frac{\sum x_i^2}{n} - 2\ \overline{x}^2 + \overline{x}^2 = \frac{x_i^2}{n} - \overline{x}^2$

$S^2 = \frac{\sum (x_{i-\overline{x})\ 2}}{n-1}$

We have $S^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2 = \frac{1}{n-1} \sum (x_i^2 - 2x_i\ \overline{x} + \overline{x}^2) = \frac{1}{n-1} \sum x_i^2 - 2\overline{x}\ \sum x_i + \sum \overline{x}^2)$

$= \frac{1}{n-1} (\sum x_i^2 - 2\overline{x}(n\overline{x}) + n\overline{x}^2\ ) = \frac{1}{n-1} (\sum x_i^2 - 2n\overline{x}^2 + n\overline{x}^2 = \frac{1}{n-1} (\sum x_i^2 - n\overline{x}^2)$

Or $S^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2 == \frac{1}{n-1} (\sum x_i^2 - n\overline{x}^2) = \frac{1}{n-1} (\sum x_i^2 - n * (\frac{\sum x_i}{n})^2 = \frac{1}{n-1} (\sum x_i^2 - n * \frac{\sum x_i}{n^2} = \frac{1}{n-1} (\sum x_i^2 - \frac{(\sum x_i 2}{n}$

Activity 02:

The following table shows the distribution of a sample of workers of a company (B) according to their monthly salaries (Unit: 10^3 DA):

Salary X_i	Less than 2	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]	70 and more	
Number of workers n_i	5	12	22	34	20	12	5	110

1. Find the suitable dispersion measure of this distribution, and then determine its value.
2. Let us consider that the classes have equal lengths. Find the standard deviation and the difference coefficient.
3. Compare the dispersion of this distribution with the dispersion of example 4-02.

Solution:

1. The suitable dispersion measure: since the frequency distribution table is open, the best dispersion measure is the quadratic difference coefficient.

Table 4-04: The distribution of the workers according to their monthly salaries:

Monthly salary X_i	**Number of employees n_i**	n_i	**Monthly salary X_i**	**Class center x_i**	$x_i n_i$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 n_i$
Less than 20	**5**	**5**	]10-20]	**15**	**75**	**889.23**	**4446.16**
]20-30]	**12**	**17**	]20-30]	**25**	**300**	**392.83**	**4714**
]30-40]	**22**	**39**	]30-40]	**35**	**770**	**96.43**	**2121.51**
]40-50]	**34**	**73**	]40-50]	**45**	**1530**	**0.0324**	**1.1016**
]50-60]	**20**	**93**	]50-60]	**55**	**110**	**103.63**	**2072.65**
]60-70]	**12**	**105**	]60-70]	**65**	**780**	**407.23**	**4886.79**
70 and more	**5**	**110**	]70-80]	**75**	**375**	**910.83**	**4554.16**
	110	**-**		**-**	**4930**	**-**	**22796.36**

Finding the quartile difference coefficient: C.Q.V=$\frac{Q_{3-Q_1}}{Q_2}$ $*100$

First, we calculate:

- The first quartile: Q_1= L1 +$\frac{\frac{N}{4}-N0}{nQ_1}$ *k = 30+$\frac{27,5-17}{22}$*10 =34,77 DA.
- The second quartile: Q_2= L1 +$\frac{\frac{2N}{4}-N0}{nQ_2}$ *k = 40+$\frac{55-39}{34}$*10 =44,7 DA.
- The third quartile: Q_3= L1 +$\frac{\frac{3N}{4}-N0}{nQ_3}$ *k = 50+$\frac{82,5-73}{20}$*10 =54,75 DA.

Thus, we find that the quartile difference coefficient is:

C.Q.V= $\frac{Q_3-Q_1}{Q_2}$*100 =$\frac{54,75+34,77}{44,7}$ *100 = 44,63%

2. Calculating the standard deviation and the difference coefficient S= $\sqrt{\boldsymbol{S^2}}$, where

$$S^2=\frac{\sum_{I=1}^{K}(x_i-\bar{x})2\,n_i}{n-1}$$

We follow these steps:

- **Calculating the standard deviation:** $\boldsymbol{\bar{x}=\frac{\sum_{i=1}^{k}x_i*n_i}{\sum_{i=1}^{k}n_i}=\frac{\sum_{i=1}^{7}x_i*n_i}{\sum_{i=1}^{7}n_i}=\frac{4930}{110}}$ **=44,82 DA**
- **Calculating the variance:**

  S^2=$\frac{\sum_{i=1}^{k}(x_i-\bar{x})2\,n_i}{n-1}$=$\frac{22796,36}{109}$= 209,14.

  Thus, S= $\sqrt{209,14}$=14,46.

- **Calculating the difference coefficient: C.V=$\boldsymbol{\frac{S}{\bar{x}}}$*100 =$\boldsymbol{\frac{14,46}{44,82}}$*100= 0,3226 *100= 32,26%**

  Thu, we say that thee dispersion of this distribution is low.

3. Comparing the dispersions of the two distributions:

From the data of example 4-02, we found out that:

- The arithmetic mean:
  $\boldsymbol{\overline{x}}=\frac{\sum_{i=1}^{k} x_i*n_i}{\sum_{i=1}^{k} n_i}=\frac{\sum_{i=1}^{6} x_i*n_i}{\sum_{i=1}^{6} n_i}=\frac{\mathbf{3840}}{\mathbf{110}}$ **=34,9 DA**

- Calculating the standard deviation: we have S^2= 217.26. Thus, S= $\sqrt{219{,}26}$=14,81
  Therefore, the difference coefficient is:
  $C.V_A=\frac{S}{\overline{x}}$***100 =**$\frac{\mathbf{14{,}81}}{\mathbf{34{,}9}}$***100= 0,4243 *100= 42,43%**

  The previous difference distribution coefficient is:
  $C.V_B=\frac{S}{\overline{x}}$***100 =**$\frac{\mathbf{14{,}46}}{\mathbf{44{,}82}}$***100= 0,3226 *100= 32,26%**
  By comparing the difference coefficients of the two distributions, we conclude that the 2nd distribution is less dispersed (more homogenous and harmonic) than the 1st distribution.

Suggested activities:

Activity 01:

Suppose we have these two series:

Series A: 10, 15, 12, 9, 15, 36, 14, 6.

Series B: 6, 24, 18, 25, 20, 28, 30, 36.

1. Find the range. What do you notice?
2. How can we distinguish these two statistical series?

Activity 02:

Consider this statistical distribution:

X_i	3	5	7	9	11
n_i	15	28	8	40	9

1. Find the range, the mean deviation, and the difference coefficient.
2. Find the quadratic deviation and the coefficient of the quadratic deviation.

Activity:

A company produces two types of the agricultural machines. The 1st type works in average 6000 hours, with a standard deviation of 900 hours. The second types works in average 8000 hours, with a standard deviation of 1020 hours.

- What is the best type?

Activity 04:

The following table shows the monthly expenditures of a sample of families in cities A and B.

- Compare the degree of homogeneity in the expenditures of the two cities.

expenditure	]20-25]	]25-30]	]30-35]	]35-40]	]40-45]	]45-50]	
City A	2	15	30	24	19	10	100
City B	32	35	15	10	6	2	100

Activity 05:

The following data show the distribution of the workers of two factories that produce the same goods, according to the monthly salaries. However, the 01st factory is in Algeria while the 2nd is in Europe.

The distribution of the monthly salaries of the 1st factory (Unit: 10^3 DA):

Salary X_i	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]	]70-80]	
Number of workers n_i	15	20	25	15	10	5	90

The distribution of the monthly salaries of the 2nd factory (Unit: 10^2 Euro):

Salary X_i	]20-30]	]30-40]	]40-50]	]50-60]	]60-70]	]70-80]	
Number of workers n_i	15	12	22	18	15	8	90

1. Find the mean deviation and the standard deviation of the two factories.
2. Compare the dispersion of the salaries of the workers of the two factories.

Activity 06:

The following table shows the distribution of families according to their monthly incomes (Unit: 10^3 DA):

Income X_i	Less than 30	]3-45]	]45-60]	]60-75]	]75-90]	90 and more	
Number of families n_i	12	25	30	18	15	10	90

1. Find the suitable dispersion measure of this distribution, and then determine its value.

2. Let us consider that the classes have equal lengths. Find the standard deviation and the difference coefficient.

3. If Y is another distribution where $S_Y = 25$ and $\bar{y} = 70$, compare the two dispersions.

4. Find the quadratic difference coefficient.

References

1. Brahim Murad Daima and Mazen Hassan Pasha: "fundamentals of Statistics with application. Curriculum, Jordan, first edition2013

2. Djellali glato: "statistics with exercises and solved problems", university 2002 publications Bureau, Algeria.

3. Khalid Ahmed Farhan al-Mashhadani and Raed Abdul Khaleq Abdullah - Al-Obeidi: "principles of Statistics ", Dar Al-Ayyam. Jordan,2013

4. Abdelrazak Azzouz: "the complete Statistics" part I, university publications Bureau, Algeria,2010

5. Abdelrazak Azzouz: "the complete statistics" Part II, university publications 2011Bureau, Algeria,

6. Ali Abdulsalam al-Amari and Ali Hussein al-ajili: "statistics, theoretical probabilities and application", publications. Malta, , ELGA2000.

7. Fathi Hamdan and Kamel Fleifel: "principles of Statistics", Dar Al-Manah, Jordan, first edition2006

8. Mohamed bouhza: "lectures on descriptive statistics", Mohammedia Public House, Algeria,2011

9. Mohamed ratoul: "descriptive statistics", Diwan of university publications, Algeria, second edition2006

10. Mohammed Sobhi Abu Saleh and Adnan Mohammed Awad: "introduction Jordan,1997to statistics", Jordan Book Center.

11. Walid Ismail al-Seifu: "fundamentals of statistical methods for business", Zamzam publications, Jordan, first edition2010

12. Bernard Py : « Statistique Descriptive », 4 ème Edition Economica,Paris, 1996.

13. Françoise couty et autres: « Probabilités et Statistiques », Dunod,Paris,1999.

14. Michel Janvier : « Statistique Descriptive », Dunod, Paris,1999