University of El Oued

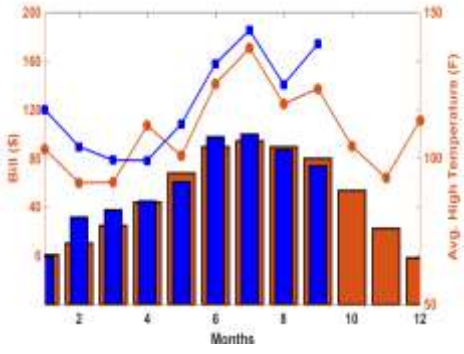
Faculty of Technology

Department of Mechanical Engineering

Lectures on Descriptive Statistics

Educational publication for LMD students

Dr. Ferhat Mohammed Said



2024 2025

Preface

This educational publication on descriptive statistics is designed for LMD students, specifically for exact sciences and technology students, aiming to provide them with a strong foundation in one of the most fundamental aspects of statistical analysis. Descriptive statistics plays a crucial role in organizing, summarizing, and interpreting raw data, which is essential in various fields. Throughout this publication, students will be guided through key concepts such as measures of central tendency, variability, and graphical data representation.

Our goal is to equip learners with the tools and understanding necessary to effectively analyze and present data in their academic and professional pursuits.

Based on the importance of this topic and its necessity for students and researchers, we have prepared these lessons in descriptive statistics according to the official decision of the Ministry of Higher Education and Scientific Research.

This educational publication contains six chapters:

- Chapter 1: Basic Concepts
- Chapter 2: Graphical Representations
- CHapter 3: Measures of Location
- Chapter 4: Measures of Dispersion
- Chapter5: Measures of Shape
- Chapter 6: Bivariate Descriptive Statistics

Contents

Chapter 1 Basic Concepts

1.1 Introduction

Statistics, as a discipline, has a deep and diverse history that stretches back centuries. Its roots can be traced to early civilizations in Egypt, Babylon, and China where people used rudimentary methods to collect and analyze data for various purposes. In the 17th century mathematicians like Blaise Pascal and Pierre de Fermat began studying games of chance, leading to the foundation of probability theory. The 18th century saw the emergence of statistical methods for data analysis and the birth of demographic statistics. During the 19th century, statistics became a formal discipline with the development of techniques for drawing inferences from data. The advent of computers in the mid-20th century enabled the development of complex statistical models and simulations. In the 21st century, statistics merged with technology, leading to the rise of data science. Big data, machine learning, and artificial intelligence became integral parts of statistical analysis.

1.2 Generalities

Definition 1.1

Statistics is the scientific discipline that involves collecting, organizing, interpreting, presenting and analyzing data on various phenomena in order to reach results that help in making appropriate decisions.

Branches of statistics

Statistics is divided into two parts:

- Descriptive statistics: This branch involves organizing and summarizing data to provide a clear picture of its main characteristics. It include measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and graphical representations like charts and graphs.
- Inferential statistics: It involves drawing conclusions or making predictions about a population based on a sample of data. It helps us make generalizations and test hypotheses.

Importance of statistics

Statistics is present in a wide range of fields such as

- Engineering: Various engineering disciplines utilize statistics to design experiments, optimize processes, and assess the reliability of materials and structures.
- Physics: Statistics plays a role in analyzing experimental data and validating theoretical models. It's used in fields like particle physics, astrophysics, and quantum mechanics.
- Economics: Statistics is used to analyze economic data, model market behaviors, estimate economic indicators, and forecast trends.
- Psychology: Psychological research uses statistics to analyze data from experiments, surveys, and observational studies.

- Medicine and Healthcare: Statistics is crucial for clinical trials, epidemiological studies, analyzing patient data, drug testing, and disease modeling.
- Biology: Statistics is crucial in genetics, ecology, and molecular biology...etc

1.3 Examples of statistical studies

Example 1.1

Consider the following data on the gender distribution of students in a school

Boys	Girls
540	900

Example 1.2

A survey of 200 university students was conducted to find out their age:

Age (years)	18	19	20	21	22	23
Nomber of students	40	48	30	22	36	24

Example 1.3

In a statistical study of the evolution of the monthly income in dinars, for 100 families, the results were as follows

Income($\times 1000$ D)	[20, 40[	[40, 60[	[60, 80[	[80, 100[	[100, 120[
Number of families	8	42	25	15	10

1.4 Statistical terms

Population

The statistical population is the set of elements on which the study relates. The elements of the population are called " statistical individuals " or " statistical units ".
If the population has N individuals, we denote the set of individuals $\Omega = \{w_1, w_2, \ldots, w_N\}$

Example 1.4

- The population studied in example 1 1.3 is students in a school with $N = 540 + 900 = 1440$
- The population studied in example 2 1.3 is the set of university students with $N = 200.$
- The population studied in example 3 1.3 is the set of families with $N = 100$

Sample

The sample is a subset of the study population. Each of its components is called a statistical unit. The sample size is the number of its statistical units, and is denoted by n

Variable (Character)

A variable is any characteristic, number, or quantity that can be observed or measured in a population or sample. It may also be called a data item.

Example 1.5

- The variable studied in example 1 1.3 is: the gender of the students.
- The variable studied in example 2 1.3 is: the age of the students.
- The variable studied in example 3 1.3 is: the monthly income in dinars of families.

Modality

A modality is any aspect or value that a variable can take, we call M the set of modalities.
If the number of modalities is denoted by r, the set of modalities of the statistical variable X will be noted by $M = \{m_1, m_2, ..., m_r\}$.

Example 1.6

- In example 1: $M = \{Boys, Girls\}$
- In example 2: $M = \{18, 19, 20, 21, 22, 23\}$
- In example 3: $M = \{[20, 40[, [40, 60[, [60, 80[, [80, 100[, [100, 120[\}$

Types of Variables

There are two types of statistical variables:

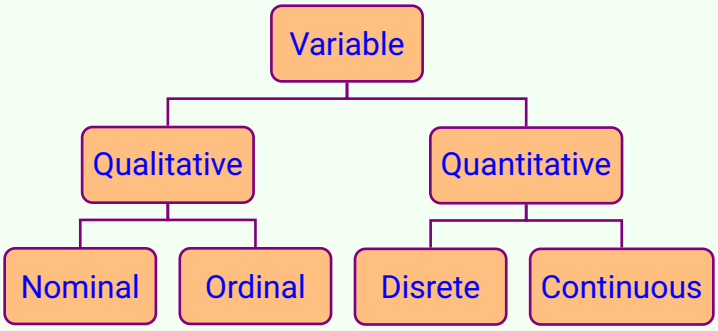
1) **Qualitative variables:** A variable is qualitative when its modalities are not measurable.
It can be of the type:

- Ordinal: The modalities are ordered. Observations can take a value that can be logically ordered or ranked.
Examples of ordinal variables include academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra large) and attitudes (i.e. strongly agree, agree, disagree, strongly disagree).
- Nominal: The modalities are not ordered. Observations can take a value that is not able to be organised in a logical sequence.
Examples of nominal variables include sex, business type, eye colour and brand.

2) **Quantitative variables:** A variable is quanitative when its modalities are measurable.
It can be of the type:

- Discrete: The modalities are countable (take isolated values).
Example: the statistical variable: (number of children per family) has finite modalities: $0, 1, 2, 3, ...$
- Continuous: The modalities take all values for the field of study. Since the number of these values is infinite, we divide them into intervals $[a_k, a_{k+1}[$ called classes.

Example: height, age, weight,... others.



The variable 'gender' in Example 1 is qualitative nominal.
The variable " age of students "in Example 2 is quantitative discrete .
The variable " monthly income " in Example 3 is quantitative continuous.

Frequencies

- A frequency (or absolute frequence) of a value x_i is the number n_i of times the observation has occurred in an experiment or study.
- A relative frequency of a value x_i is the ratio (fraction or proportion) of its absolute frequeny n_i to N the total number of outcomes $f_i = \frac{n_i}{N}$.
- The ascending cumulative frequency associated with the value x_i (denoted n_iac) is the number of data points that are less than or equal to x_i $n_iac = n_{i-1}ac + n_i$.
- The descending cumulative frequency associated with the value x_i (denoted n_iac) is the number of data points that are greater than or equal to x_i $n_idc = n_{i-1}dc - n_{i-1}$.
  - The ascending cumulative frequency tells us how many data points are "below" x_i, while the descending cumulative frequency tells us how many data points are "above" x_i.

Example 1.7

In example 2 1.3:

$$M = \{18, 19, 20, 21, 22, 23\}, N = \sum_{i=1}^{6} n_i = 200.$$

Age x_i (years)	18	19	20	21	22	23
Frequency n_i	40	48	30	22	36	24
Relative Frequency f_i	0.20	0.24	0.15	0.11	0.18	0.12
Ascending C Frequency n_iac	40	88	118	140	176	200
Descending C Frequency n_idc	200	160	112	82	60	24

Example 1.8

In example 3 1.3:

$M = \{[20, 40[, [40, 60[, [60, 80[, [80, 100[, [100, 120[\}$

class	$[20, 40[$	$[40, 60[$	$[60, 80[$	$[80, 100[$	$[100, 120[$
n_i	8	42	25	15	10
$n_i ac$	8	50	75	90	100
$n_i dc$	100	92	50	25	10
f_i	0.08	0.42	0.25	0.15	0.10
$f_i ac$	0.08	0.50	0.75	0.90	1
$f_i dc$	1	0.92	0.50	0.25	0.10

Continuous frequency distribution

The general rule for transferring the data into a continuous quantitative table is that the number of classes is between 5 and 15, but it is up to the statistician to determine k, and L , although there are methods to determine it, in particular:

- Rule of Yule: $k = 2.5\sqrt[4]{N}$
- Rule of Sturge: $k = 1 + 3.322\, log_{10}(N)$

where k is the number of classes, N is the size of the data, the range of the data $R = x_{max} - x_{min}$, L is the class width which can be calculated by: $L = \frac{R}{k}$

Example 1.9

The teachers' travel time from their residence to the university was recorded in minutes as follows:

5	2	28	30	25	20	45	18	5	5
65	10	50	25	29	23	24	46	4	70
60	35	34	23	31	61	33	30	45	31
27	20	6	5	2	44	11	30	2	7
14	3	20	43	41	8	24	16	14	5

$N = 50,\ x_{max} = 70,\ x_{min} = 2,\ R = 70 - 2 = 68$

$k = 1 + 3.322\, log_{10}(50) \simeq 7,\ L = \frac{68}{7} \simeq 10.$

Hence, the frequency table can be formed as follows:

class	[2, 12[	[12, 22[	[22, 32[	[32, 42[	[42, 52[	[52, 62[	[62, 72[
n_i	15	7	14	4	6	2	2
n_iac	15	22	36	40	46	48	50
n_idc	50	35	28	14	10	4	2
f_i	0.30	0.14	0.28	0.08	0.12	0.04	0.04
f_iac	0.30	0.44	0.72	0.80	0.92	0.96	1
f_idc	1	0.70	0.56	0.28	0.20	0.08	0.04

Chapter 2 Graphical Representations

2.1 Introduction

Graphical representations are essential tools in statistics that provide a visual interpretation of data, making complex datasets easier to understand and analyze. Through graphs, charts, and plots, patterns, trends, and outliers in data become more apparent, offering insights that may not be obvious in raw numerical data. Common graphical methods include bar charts, histograms, pie charts, and scatter plots, each serving specific purposes depending on the nature of the data. This chapter explores these tools and their importance in conveying statistical information clearly and effectively.

2.2 Bar Chart - Pie Charts

Simple bar chart

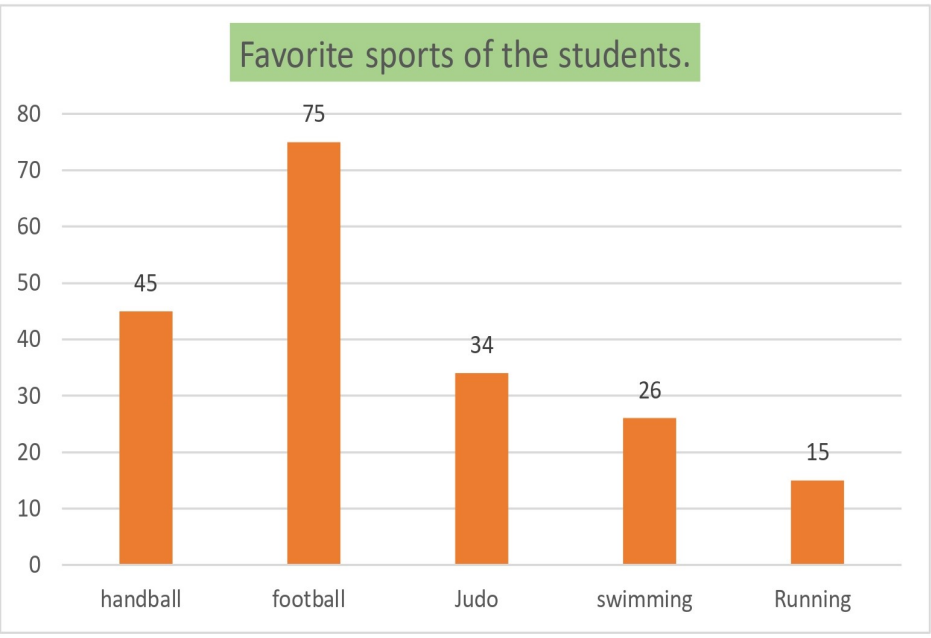
A simple bar chart is used to represent data involving only one variable classified on a spatial, quantitative or temporal basis. In a simple bar chart, we make bars of equal width but variable length. There are two types of bar diagrams namely, Horizontal Bar diagram and Vertical bar diagram. We use horizontal bar diagram for qualitative data or data varying over space, while the vertical bar diagram is associated with quantitative data or time series data.

Example 2.1

A school conducted a survey to know the favorite sports of the students. The table below shows the results of this survey.

Sport	Handball	Football	Judo	Swimming	Running
Nomber of students	42	75	34	26	15

The given data can be represented as



Pie charts

In a pie chart, the various observations or components are represented by the sectors of a circle and the whole circle represents the total value of all the components .

The central angle of a component is $\frac{\text{Value of the component}}{N} \times 360$.

Example 2.2

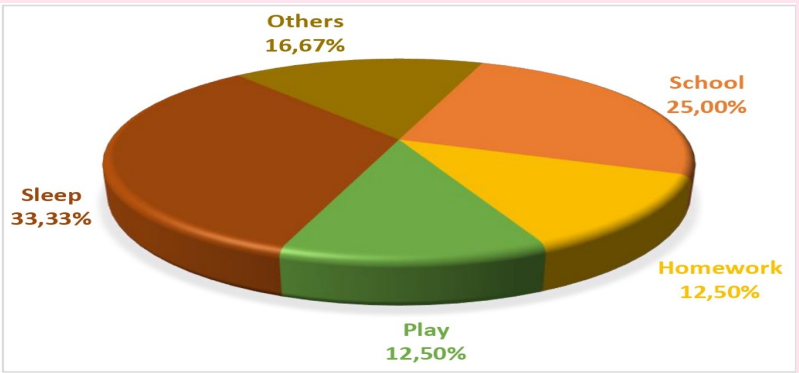
The number of hours spent by a school student on various activities on a working day, is given below.

Activity	School	Homework	Play	Sleep	Others
Nomber of hours	6	3	3	8	4

Construct a pie chart using the angle measurement.

We can calculate the central angles for various the components, and then we get the required pie chart as shown in the figure below.

Activity	**Hours**	**Cenral angle**
School	6	$6/24 \times 360 = 90$
Homework	3	$3/24 \times 360 = 45$
Play	3	$3/24 \times 360 = 45$
Sleep	8	$8/24 \times 360 = 120$
Others	4	$4/24 \times 360 = 60$



2.3 Histogram - Frequency Polygon

Histogram

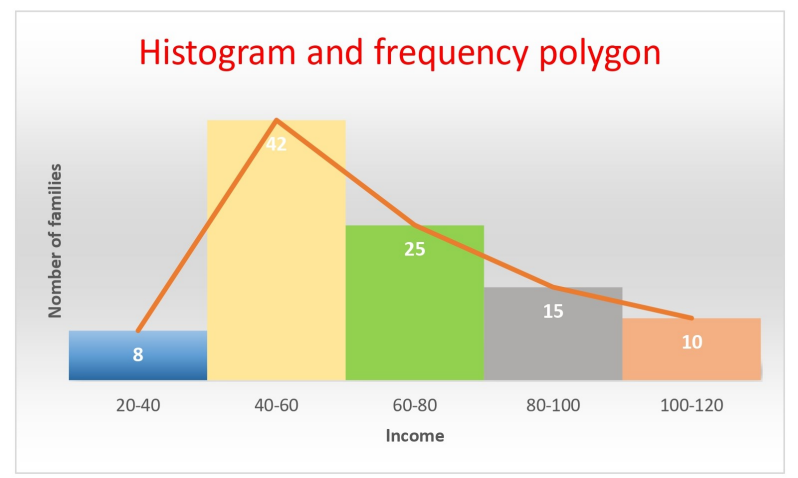
A histogram is a graphic representation of data in a grouped frequency distribution with continuous classes. They resemble bar graphs, but there are no gaps between the consecutive rectangles. Histograms are widely used for ranging data into groups. It is a highly practical graph due to its clarity and simplicity.

Frequency Polygon

A frequency polygon is a graph that displays the distribution of a set of continuous data. It aims to visually represent a group of consistent data issuance, making it easier to interpret and analyse by plotting the frequency of each interval on the vertical axis and the midpoint of each interval on the horizontal axis.

Example 2.3

In the first Chapter , we can present the histogram and frequency polygon of the continuous variable mentioned in Example 3 1.3 as follows



☞ Remark

If the classes are of unequal width, we draw the histogram using the adjusted frequency n_i' by the following formula: $n_i' = \dfrac{n_i \times l}{l_i}$

where n_i: concerned class frequency, l_i: concerned class width, l: lowest width

Example 2.4

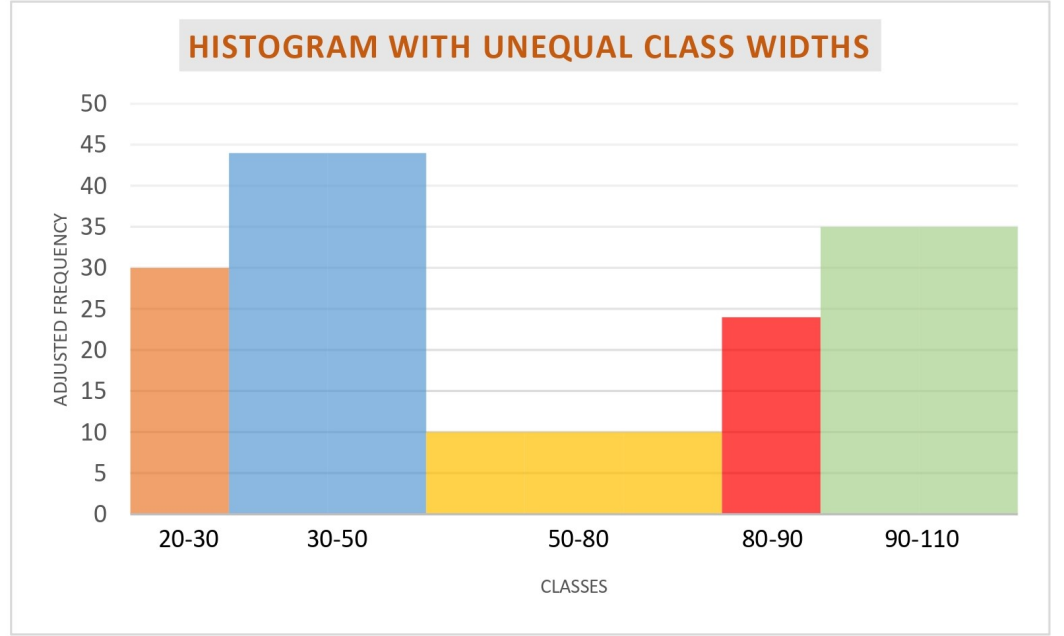
Draw the following frequency distribution with a histogram

Classes	[20, 30[	[30, 50[	[50, 80[	[80, 90[	[90, 110[
n_i	30	88	42	24	36

Solution: First, it is necessary to calculate the value of the adjusted frequency.

Classes	[20, 30[	[30, 50[	[50, 80[	[80, 90[	[90, 110[
n_i	30	88	30	24	70
l_i	10	20	30	10	20
n_i'	30	44	10	24	35

We can therefore plot the histogram using adjusted frequency values.



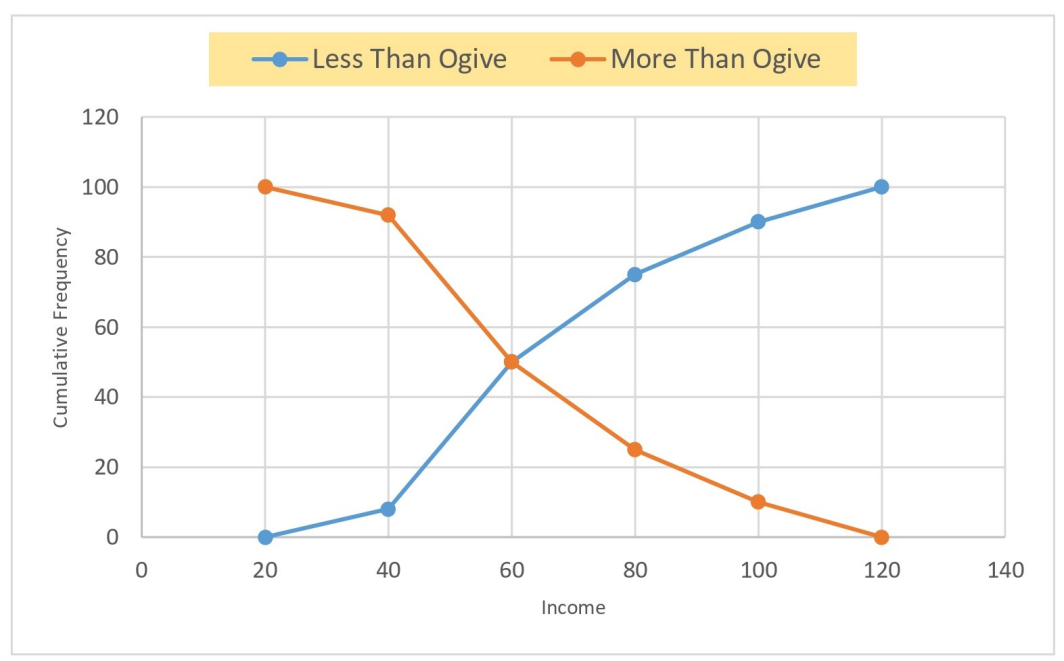
2.4 Ogive (Cumulative Frequency Polygon)

Ogive

An ogive (or Cumulative Frequency polygon) is a polygon of a data set obtained by an individual through the representation of cumulative frequency distribution on a graph. As there are two types of cumulative frequency distribution (Ascending cumulative frequencies and Descending cumulative frequencies), the ogives are also of two types: Less than ogive and More than ogive.

Example 2.5

We come back to example 3 1.3 of chapter 1



Chapter 3 Measures of Location (Central Tendency)

3.1 introduction

The representative value of a data set, generally the central value or the most occurring value that gives a general idea of the whole data set is called the Measure of Central Tendency.
Some of the most commonly used measures of central tendency are: Mode , Mean, Median

3.2 Mode

Definition

The Mode of a statistical series, which we denote by the symbol Mo is the value of that observation which has a maximum frequency corresponding to it.

Mode of Ungrouped Data

Mode of Ungrouped Data can be simply calculated by observing the observation with the highest frequency.

☞ Remark

The greatest frequency can occur at two or more different values. If the data has two modes, the data is bimodal. If the data has more than two modes, the data is multimodal.
If the values are repeated with the same frequency, then there is no mode.

Example 3.1

In the series: $10, 4, 6, 4, 1, 4, 6$. the mode is $MO = 4$.
The series: $5, 6, 7, 5, 7, 8$ has two modes which are 5 and 7.
There is no mode in the series: $2, 5, 3, 3, 5, 2$.

Mode of Grouped Data

The mode of grouped data is the value that appears most often in the data set. It is determined in two ways:

- **Analytical method:** First, we determine the modal class that corresponds to the highest frequency. Inside this class, we can locate the mode value of the data by using the formula: $Mo = a + \frac{\Delta_1}{\Delta_1 + \Delta_2}.l$, where a is the lower limit of modal class, l is the class size, Δ_1 is the difference between the frequency of the modal class and the class preceding it, Δ_2 is the difference between the frequency of the modal class and the class succeding it.
- **Graphical Method** For calculating the mode of the grouped data graphically, the following procedure is adopted.
  Draw a histogram of the data; the modal class is the tallest rectangle. Draw a line from the top right corner of the tallest rectangle to the top right corner of the preceding rectangle. Draw a line from the top left corner of the tallest rectangle to the top left corner of the succeeding rectangle. The abscissa of the point where these two lines intersect represents the Mode

Example 3.2

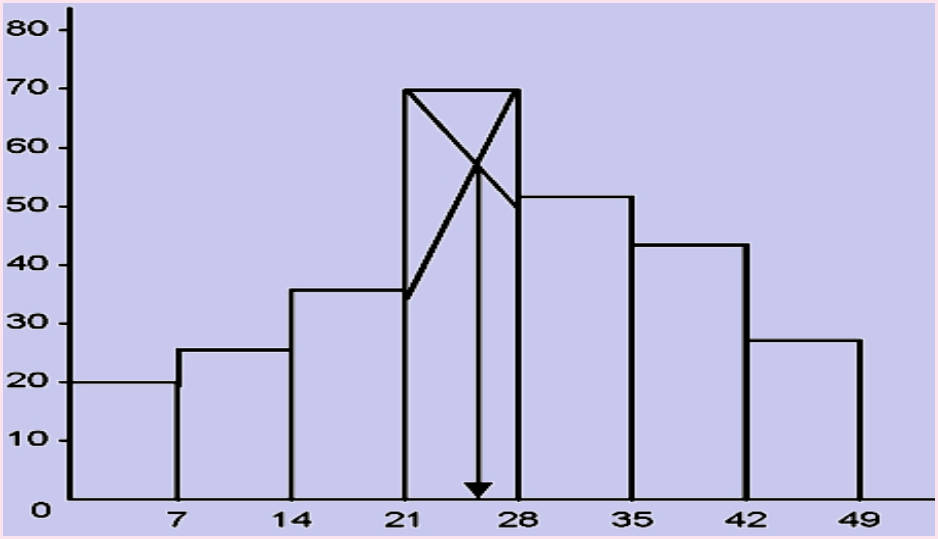
Let us calculate the Mode using the analytical and graphical method for the following distributions:

Classes	[0, 7[	[7, 14[	[14, 21[	[21, 28[	[28, 35[	[35, 42[	[42,49[
Frequency	19	25	36	72	51	43	28

As the class interval with the highest frequency is $[21, 28[$, which has a frequency of 72. Thus, $[21-28[$ is the modal class, $a = 21, l = 7, \Delta_1 = 72-36 = 36, \Delta_2 = 72-51 = 21$.

Therefore, the mode for this set of data is $Mo = 21 + \frac{36}{36+21}.7 = 25,42$.

The following figure shows how to find the mode



☞ Remark

If the classes are unequal width, the adjusted frequency is used instead of the original frequency.

Example 3.3

Calculate the mode in the following distribution:

Classes	[10, 20[	[20, 50[	[50, 70[	[70, 110[	110, 120[
n_i	10	60	30	20	20

Solution: First, it is necessary to calculate the value of the adjusted frequency.

Classes	[10, 20[	[20, 50[	[50, 70[	[70, 110[	[110, 120[
n_i	10	60	30	20	20
l_i	10	30	20	40	10
n_i'	10	20	15	5	20

We notice from the adjusted frequency that there are two modal classes $[20, 50[, [110, 120[$. Thus the two modes are:

$$Mo_1 = 20 + \frac{20-10}{20-10+20-15}.30 = 40 \text{ and } Mo_2 = 110 + \frac{20-5}{20-5+20-0}.10 = 114,29$$

3.3 Arithmetic Mean

Definition

Arithmetic Mean is defined as the ratio of all the values or observations to the total number of values or observations. It is sometimes also called mean, average, or arithmetic average.

Mean Formula

✿ **Direct method:**

For r values in a raw data $x_1, x_2, \ldots, x_r$ the mean is: $\overline{X} = \frac{1}{r}\sum_{i=1}^{r} x_i$

For r values x_i with their frequencies n_i the mean is:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{r} n_i x_i \text{ , where } N = \sum_{i=1}^{r} n_i$$

For r classes $[a_i, b_i[$ with their frequencies n_i the mean is:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{r} n_i c_i \text{ , where } c_i = \frac{a_i + b_i}{2}$$

✿ **Indirect Method:**

We can easily find the arithmetic mean using the indirect method also called the assumed mean method by the following formula: $\overline{X} = A + \frac{\sum n_i(x_i - A)}{N}$.

Where A is the assumed mean (preferably corresponding to the highest frequency)

Example 3.4

1) The Mean of this raw data $10, 11, 15, 16, 18$ is: $\overline{X} = \frac{10+11+15+16+18}{5} = 14$

2) Let's have the following distribution

x_i	20	25	30	40	60
n_i	5	10	12	8	5

The Mean is: $\overline{X} = \frac{20\times 5 + 25\times 10 + 30\times 12 + 40\times 8 + 60\times 5}{40} = 33.25$

Example 3.5

Calculate the arithmetic mean using different methods for the following distribution:

Class	[5, 15[	[15, 25[	[25, 35[	[35, 45[
Frequency	6	12	3	9

Solution: Let the assumed mean be $A = 20$ (the midpoint of modal class)

Class	Midpoint c_i	Frequency n_i	Deviation $d_i = c_i - A$	$n_i c_i$	$n_i d_i$
$[5, 15[$	10	6	-10	60	-60
$[15, 25[$	20	12	0	240	0
$[25, 35[$	30	3	10	90	30
$[35, 45[$	40	9	20	360	180
$\sum$		30		750	150

By direct method: $\overline{X} = \frac{\sum n_i c_i}{N} = \frac{750}{30} = 25$

By indirect method: $\overline{X} = A + \frac{\sum n_i d_i}{N} = 20 + \frac{150}{30} = 20 + 5 = 25$

Other Types of Mean

- **Geometric Mean:** $G = \sqrt[N]{\prod x_i^{n_i}} = e^{\frac{\sum n_i \log x_i}{N}}$
- **Harmonic Mean:** $H = \frac{N}{\sum \frac{n_i}{x_i}}$
- **Quadratic Mean:** $Q = \sqrt{\frac{\sum n_i x_i^2}{N}}$

Example 3.6

Calculate by different types the mean of the following distribution:

x_i	3	5	6	7	9
n_i	2	3	2	3	5

Solution: Let us form this table

x_i	n_i	x_i^2	$\log x_i$	$n_i x_i$	$\frac{n_i}{x_i}$	$n_i x_i^2$	$n_i \log x_i$
3	2	9	1.10	6	0.67	18	2.20
5	3	25	1.61	15	0.60	75	4.83
6	2	36	1.79	12	0.33	72	3.58
7	3	49	1.95	21	0.43	147	5.85
9	5	81	2.20	45	0.56	405	11
Total	15			99	2.59	717	27.46

Using the above formulas, we obtain

$$\overline{X} = \frac{99}{15} = 6.60,\ G = e^{\frac{27.46}{15}} = 6.23,\ H = \frac{15}{2.59} = 5.79,\ Q = \sqrt{\frac{717}{15}} = 6.91$$

3.4 Median

Definition

The median of a statistical series, which we denote by the symbol Me, is that value that divides the distribution into two equal parts such that the number of observations above it is equal to the number of observations below it.

Median of Ungrouped Data

To calculate the Median, the observations must be arranged in ascending or descending order. If the total number of observations is N then there are two cases:

✿ **Case 1:** N is odd $Me = x_{\frac{N+1}{2}}$

✿ **Case 2:** N is Even $Me = \frac{\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right)}{2}$

Example 3.7

If the observations are : $25, 36, 31, 23, 22, 26, 38, 28, 20$, then the Median is given by arranging the data in ascending order $20, 22, 23, 25, \underline{26}, 28, 31, 36, 38$.
$N = 9$ which is odd then $Me = x_{\frac{9+1}{2}} = x_5 = 26$

Example 3.8

If the observations are : $25, 36, 31, 23, 22, 26, 38, 28, 20, 32$, then the Median is given by arranging the data in ascending order $20, 22, 23, 25, \underline{26}, \underline{28}, 31, 32, 36, 38$.
$N = 10$ which is even then $Me = \frac{x_5 + x_6}{2} = \frac{26 + 28}{2} = 27$.

☞ **Remark**

The greatest frequency can occur at two or more different values. If the data has two modes, the data is bimodal. If the data has more than two modes, the data is multimodal.
If the values are repeated with the same frequency, then there is no mode.

Median of Discrete serie

In case the data is discrete with repetition, we determine the median rank $\frac{N}{2}$ among the values of the ascending cumulative frequency , the corresponding value $x_{\frac{N}{2}}$ is the median.

Example 3.9

This table shows the results of 28 students in physics. Calculate the median for this data.

Result	4	7	9	10.5	11.5	12	15	17
Frequency	3	5	4	8	2	3	2	1

Solution: Create the following table for the given data.

Result	4	7	9	10.5	11.5	12	15	17
Frequency	3	5	4	8	2	3	2	1
Cumulative Frequency	3	8	12	20	22	25	27	28

As $\frac{N}{2} = 14$ thus, the median is: $Me = x_{14} = 10.5$.

☞ **Remark**

If the value of the median rank $\frac{N}{2}$ is present among the values of the ascending cumulative frequency, the median is the average of the two values before and after the median rank

$$Me = \frac{(x_{\frac{N}{2}} + x_{\frac{N}{2}+1})}{2}$$

Example 3.10

Calculate the median for the following data.

Result	4	7	9	10.5	11.5	12	15	17
Frequency	3	7	4	6	2	3	2	1

Solution: Create the following table for the given data.

Result	4	7	9	10.5	11.5	12	15	17
Frequency	3	7	4	6	2	3	2	1
Cumulative Frequency	3	10	14	20	22	25	27	28

In this example, $\frac{N}{2} = 14$ is one of the ascending cumulative frequency, thus, the median is:

$$Me = \frac{(x_{14} + x_{15})}{2} = \frac{9 + 10.5}{2} = 9.75.$$

Median of Grouped Data

The median of grouped data is the value that appears most often in the data set. It is determined in two ways:

✿ **Analytical method:** First we identify the Median class whose cumulative frequency is greater

than and near to $\frac{N}{2}$. The following formula is then applied to determine the actual median value.

$$Me = a + \frac{(\frac{N}{2} - n_{i-1}ac)}{n_i}.l \ ,$$

where a is the lower limit of median class, l is the class size, N is total number of observations, $n_{i-1}ac$ is the cumulative frequency of the class preceding the median class , n_i the frequency of median class.

✿ **Graphical Method** The median of a given series can also be calculated through its graphical representation in the form of less than or more than ogive. Through ogives, the median can be determined in two ways:

1) Presenting the data of given series on a graph, simultaneously in the form of both less than ogive and more than ogive. The point where these two ogives meet is the median value.

2) Presenting the data in the form of the ascending (or desending) cumulative polygone separately, where at the level of the middle of the frequencies, we draw a horizontal line until it touches the cumulatie polygon and then we drop a column on the horizontal axis to obtain the median .

Example 3.11

Calculate the median for the following data, and determine it graphically.

Class	[20, 40[	[40, 60[	[60, 80[	[80, 100[	[100, 120[
Frequency	5	10	12	8	5

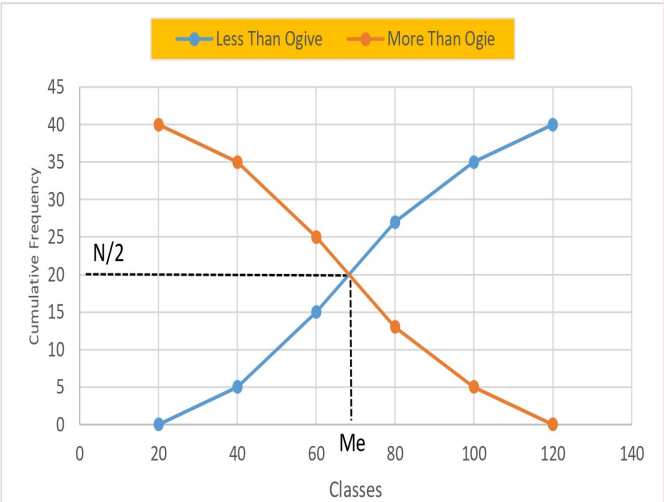
Solution: Create the following table for the given data.

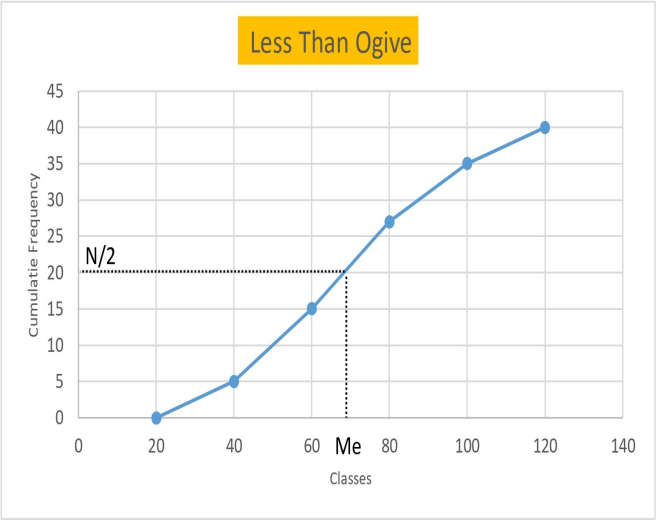
Class	[20, 40[	[40, 60[	[60, 80[	[80, 100[	[100, 120[
Frequency	5	10	12	8	5
Cumulative Frequency	5	15	27	35	40

■ As $\frac{N}{2} = 20$ thus, $[60 - 80[$ is the median class, $a = 60, l = 20, n_i = 12, n_{i-1}ac = 15$. Therefore, the median for this set of data is:

$$Me = 60 + \frac{20 - 15}{12}.20 = 68,33.$$

■ The following figures show how to find the median





3.5 Quantiles

Definition

A quantile is a value that divides a probability distribution into parts with equal probability. Quantiles are often used to summarize data and to compare different distributions. To find them, we follow the same procedure for finding the median.

Most common quantiles

The most common quantiles are:

- **Quartiles**

  The quartiles divide a ranked data set into four equal parts. These three measures are denoted first quartile (denoted by $Q1$), second quartile (denoted by $Q2$) and third quartile (denoted by $Q3$). The second quartile is the same as the median Me of a data set. The first quartile is the value of the middle term among the observations that are less than the median and the third quartile is the value of the middle term among the observations that are greater than the median.

- **Deciles**

  The deciles are the values that separate a distribution into ten equal parts, where each part

contains the same number of observations.
For a given distribution, there are nine deciles, which we denote by: $D_1, D_2, \ldots, D_9$.
The decile D_i indicates the value where $10i\%$ of the observations occur below this value and $(100-10i)\%$ of the observations occur above this value. For example, the eighth decile D_8 is the value where 80% of the observations fall below this and 20% occur above it.

✿ **Percentiles**

The percentiles divide the data into 100 equal parts.
For a given distribution, there are 99 percentiles, which we denote by: $P_1, P_2, \ldots, P_{99}$.

Example 3.12

Calculate $Q_1, Q_3, D_3, D_8, P_{35}, P_{95}$, for this data.

X_i	10	11	12	13	14	15	16	17
n_i	30	20	40	35	35	25	5	10

Solution:

Let's create the following table for that data.

X_i	10	11	12	13	14	15	16	17
n_i	30	20	40	35	35	25	5	10
n_iac	30	50	90	125	160	185	190	200

- $\frac{N}{4} = 50$ is one of the ascending cumulative frequency, thus the first quartile is:

$$Q_1 = \frac{(x_{50} + x_{51})}{2} = \frac{11+12}{2} = 11.5.$$

- $\frac{3N}{4} = 150$, thus the third quartile is: $Q_3 = x_{150} = 14$
- $\frac{3N}{10} = 60$, thus the third decile is: $D_3 = x_{60} = 12$
- $\frac{8N}{10} = 160$ s one of the ascending cumulative frequency, thus the eighth decile is:

$$D_8 = \frac{(x_{160} + x_{161})}{2} = \frac{14+15}{2} = 14.5.$$

- $\frac{35N}{100} = 70$, thus the thirty-fifth percentile is: $P_{35} = x_{70} = 12$
- $\frac{95N}{100} = 190$ is one of the ascending cumulative frequency, thus the ninety-fifth percentile is:

$$P_{95} = \frac{(x_{190} + x_{191})}{2} = \frac{16+17}{2} = 16.5.$$

Example 3.13

in this table, we give the results of students on a mathematics exam

Result	[0, 4[	[4, 8[	[8, 12[	[12, 16[	[16, 20[
Frequency	5	12	25	20	10

Calculate $Q_1, Q_3, D_2, D_7, P_{52}, ,$ for the above data.

Solution: Let's create the following table for that data.

Class	[0, 4[	[4, 8[	[8, 12[	[12, 16[	[16, 20[
n_i	5	12	25	20	10
$n_i ac$	5	17	42	62	72

- As $\frac{N}{4} = 18$ thus, $Q_1 \in [8, 12[$, $a = 8, l = 4, n_i = 25, n_{i-1}ac = 17$.
  Therefore, the first quantile for this set of data is:

$$Q_1 = a + \frac{(\frac{N}{4} - n_{i-1}ac)}{n_i}.l = 8 + \frac{18-17}{25}.4 = 8.16.$$

- As $\frac{3N}{4} = 54$ thus, $Q_3 \in [12, 16[$, $a = 12, l = 4, n_i = 20, n_{i-1}ac = 42$.
  Therefore, the third quantile for this set of data is:

$$Q_3 = a + \frac{(\frac{3N}{4} - n_{i-1}ac)}{n_i}.l = 12 + \frac{54-42}{20}.4 = 14.4.$$

- As $\frac{2N}{10} = 14.4$ thus, $D_2 \in [4, 8[$, $a = 4, l = 4, n_i = 12, n_{i-1}ac = 5$.
  Therefore, the second decile for this set of data is:

$$D_2 = a + \frac{(\frac{2D}{10} - n_{i-1}ac)}{n_i}.l = 4 + \frac{14.4-5}{12}.4 = 7.17.$$

- As $\frac{7N}{10} = 50.4$ thus, $D_7 \in [12, 16[$, $a = 12, l = 4, n_i = 20, n_{i-1}ac = 42$.
  Therefore, the seventh decile for this set of data is:

$$D_7 = a + \frac{(\frac{7D}{10} - n_{i-1}ac)}{n_i}.l = 12 + \frac{50.4-42}{20}.4 = 13.68.$$

- As $\frac{55N}{100} = 39.6$ thus, $P_{55} \in [8, 12[$, $a = 8, l = 4, n_i = 25, n_{i-1}ac = 17$.
  Therefore, the fifty fifth percentile for this set of data is:

$$P_{55} = a + \frac{(\frac{55N}{100} - n_{i-1}ac)}{n_i}.l = 8 + \frac{39.6-17}{25}.4 = 8.16.$$

Chapter 4 Measures of Dispersion (Variation)

4.1 introduction

The measures of dispersion are non-negative real numbers, that help to gauge the spread of data about a central value, and to determine how stretched or squeezed the given data is.
Along with measures of central tendency, measures of variability give you descriptive statistics that summarize your data.
To measure the degree of dispersion of these values by the central value we use some dispersion characteristics such as: the range, the variance, the standard deviation, mean deviation, quartile deviation and coefficient of variation.

4.2 Range

Definition

The range denoted R or the interval of variation of a statistical series is the difference between the highest value and the smallest value of the statistical variable $R = H - S$

Example 4.1

Find the range and coefficient of range of the data set $8, 12, 5, 6, 8, 2, 15$
Solution: we have: $H = 15, S = 2$, thus the range is: $R = 15 - 2 = 13$

4.3 Mean Deviation

Definition

The mean deviation gives the the arithmetic mean of the data's absolute deviation about the central points. These central points could be the mean, median, or mode.
It expresses the intensity of clustering of data around each other and the extent of their homogeneity around a central value.

✿ **In the case of ungrouped data**

the mean deviation of n values $x_1, x_2, \ldots, x_r$, from their arihmetic mean $\overline{X}$, is the number noted by $e_{\overline{X}}$ defined as follows: $e_{\overline{X}} = \frac{1}{r}\sum_{i=1}^{r}|x_i - \overline{X}|$

✿ **In the case of grouped discrete data**

the mean deviation $e_{\overline{X}}$ of r values $x_1, x_2, \ldots, x_r$, associated with the frequencies $n_1, n_2, \ldots, n_r$ respectively is the number defined as follows:

$$e_{\overline{X}} = \frac{1}{N}\sum_{i=1}^{r} n_i|x_i - \overline{X}| \text{ , where } N = \sum_{i=1}^{r} n_i$$

✿ **In the case of grouped continuous data**

For r classes $[a_i, b_i[$ with their frequencies n_i the mean deviation is:

$$e_{\overline{X}} = \frac{1}{N}\sum_{i=1}^{r} c_i|x_i - \overline{X}| \quad \text{, where } c_i = \frac{a_i + b_i}{2}$$

Example 4.2

- We consider the series:4, 7, 9, 10, 13, 17 . The mean is: $\overline{X} = 10$, the absolute deviations from the mean are: 6, 3, 1, 0, 3, 7, therefore $e_{\overline{X}} = \frac{1}{6}(6+3+1+0+3+7) = 3.33$
- We consider the following grouped discrete data:

x_i	7	9	11	13	15
n_i	3	7	5	4	1

To calculate the mean deviation, we create this table:

x_i	n_i	$n_i x_i$	$\lvert x_i - \overline{X}\rvert$	$n_i\lvert x_i - \overline{X}\rvert$
7	3	21	3.3	9.9
9	7	63	1.3	9.1
11	5	55	0.7	3.5
13	4	52	2.7	10.8
15	1	15	4.7	4.7
Total	20	206		38

The arithmetic mean is: $\overline{X} = \frac{206}{20} = 10.3$

Based on the above mentioned formula, the Mean Deviation will be: $e_{\overline{X}} = \frac{38}{20} = 1.9$

- For the grouped continuous distribution, let's go back to the data in Example 3.5

Class	**Midpoint** c_i	**Frequency** n_i	$\lvert d_i\rvert = \lvert c_i - \overline{X}\rvert$	$n_i c_i$	$n_i\lvert d_i\rvert$
$[5, 15[$	10	6	15	60	90
$[15, 25[$	20	12	5	240	60
$[25, 35[$	30	3	5	90	15
$[35, 45[$	40	9	15	360	135
$\sum$		30		750	300

The arithmetic mean is: $\overline{X} = \frac{750}{30} = 25$

Based on the above mentioned formula, the Mean Deviation will be: $e_{\overline{X}} = \frac{300}{30} = 10$

4.4 Variance

Definition

The variance measures the spread or dispersion of a set of data points around their mean. It is calculated as the average of the squared differences between each data point and the mean.

✿ **In the case of ungrouped data**

the variance of r values $x_1, x_2, \ldots, x_r$, from their arihmetic mean $\overline{X}$, is the number noted by σ^2 defined as follows:

$$\sigma^2 = \frac{1}{r}\sum_{i=1}^{r}|x_i - \overline{X}|^2$$

✿ **In the case of grouped discrete data**

the variance σ^2 of r values $x_1, x_2, \ldots, x_r$, associated with the frequencies $n_1, n_2, \ldots, n_r$ respectively is defined as follows:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{r} n_i|x_i - \overline{X}|^2 \text{ , where } N = \sum_{i=1}^{r} n_i$$

✿ **In the case of grouped continuous data**

For r classes $[a_i, b_i[$ with their frequencies n_i the variance is:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{r} n_i|c_i - \overline{X}|^2 \text{ , where } c_i = \frac{a_i + b_i}{2}$$

4.5 Standard Deviation

Definition

The standard deviation is the square root of the variance. $\sigma = \sqrt{V}$

Example 4.3

✿ We consider the series: $10, 11, 12, 15, 17$.
The mean is: $\overline{X} = 13$, the absolute deviations from the mean are: $3, 2, 1, 2, 4$.
Therefore $\sigma^2 = \frac{1}{5}(3^2 + 2^2 + 1^2 + 2^2 + 4^2) = 6.8$ and $\sigma = \sqrt{6.8} = 2.6$.

✿ We consider the following grouped discrete data:

x_i	5	10	15	20	25
n_i	2	8	3	4	3

To calculate the mean deviation, we create this table:

x_i	n_i	$n_i x_i$	$\lvert x_i - \overline{X}\rvert$	$n_i \lvert x_i - \overline{X}\rvert^2$
5	2	10	9.5	180.5
10	8	80	4.5	162
15	3	45	0.5	0.75
20	4	80	5.5	121
25	3	75	10.5	330.75
Total	20	290		795

The arithmetic mean is: $\overline{X} = \frac{290}{20} = 14.5$

Based on the above mentioned formula, $\sigma^2 = \frac{795}{20} = 39.75$, and $\sigma = \sqrt{39.75} = 6.30$

✿ For the grouped continuous distribution, let's go back to the data in Example 3.5

Class	**Midpoint** c_i	**Frequency** n_i	$\lvert d_i\rvert = \lvert c_i - \overline{X}\rvert$	$n_i c_i$	$n_i \lvert d_i\rvert^2$
$[5, 15[$	10	6	15	60	1350
$[15, 25[$	20	12	5	240	300
$[25, 35[$	30	3	5	90	75
$[35, 45[$	40	9	15	360	2025
$\sum$		30		750	3750

The arithmetic mean is: $\overline{X} = \frac{750}{30} = 25$

Based on the above mentioned formula, $\sigma^2 = \frac{3750}{30} = 125$, and $\sigma = \sqrt{125} = 11.18$

4.6 Quartile Deviation

Definition

The quartile deviation is the half of the difference between the third quartile and the first quartile in a given data set.

$$W = \frac{Q_3 - Q_1}{2}$$

4.7 Coefficient of Variation

Definition

The coefficient of variation is the ratio of the standard deviation to the mean of the data set. It is a dimensionless quantity and is usually given as a percentage. It helps to compare two data sets

on the basis of the degree of variation.

$$CD = \frac{\sigma}{\overline{X}} \times 100$$

Example 4.4

The number of goals scored by two teams in a 30 football matches is given below. With the help of Coefficient of Variation, determine which team is more consistent.

Number of goals	0	1	2	3	4	5
Number of matches (Team A)	8	5	4	8	2	3
Number of matches (Team B)	2	11	9	3	3	2

Solution:

Let's create the following table for that data.

x_i	0	1	2	3	4	5	$\sum$
n_i	8	5	4	8	2	3	30
f_i	2	11	9	3	3	2	30
$n_i x_i$	0	5	8	24	8	15	60
$f_i x_i$	0	11	18	9	12	10	60
$\lvert x_i - \overline{X}_A \rvert$	2	1	0	1	2	3	
$\lvert x_i - \overline{X}_B \rvert$	2	1	0	1	2	3	
$n_i \lvert x_i - \overline{X}_A \rvert^2$	32	5	0	8	8	27	80
$f_i \lvert x_i - \overline{X}_B \rvert^2$	8	11	0	3	12	18	52

The teams A and B have the same mean goals scored

$\overline{X}_A = \frac{1}{N}\sum n_i x_i = \frac{60}{30} = 2$ and $\overline{X}_B = \frac{1}{N}\sum f_i x_i = \frac{60}{30} = 2$

The two standard deviations are: $\sigma_A = \sqrt{\frac{80}{30}} = 1.63$ and $\sigma_B = \sqrt{\frac{52}{30}} = 1.32$

The coefficients of variation are: $CD_A = \frac{1.63}{2} = 0.815$ and $CD_B = \frac{1.32}{2} = 0.66$.

We have $CD_B < CD_A$, so Team B is more consistent

4.8 Coefficient of Quartile Deviation

Definition

The coefficient of quartile deviation is the ratio of the difference between the third quartile and the first quartile to the sum of the third and first quartiles.

$$CDQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 4.5

Let's return to Example 3.13

$$CDQ = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{14.4 - 8.16}{14.4 + 8.16} = 0.28$$

Chapter 5 Measures of Shape

5.1 Introduction

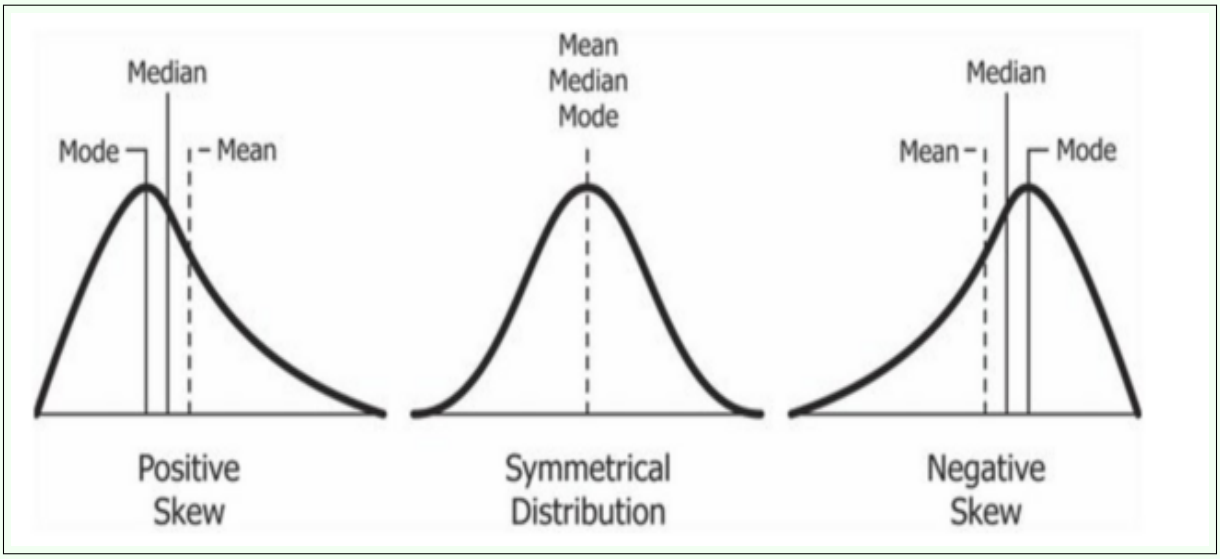
In the previous two chapters, we discussed measures of location and dispersion measures. these measures can describe the distribution, but they are not sufficient to describe the nature of the distribution. For this purpose, we use other two statistical measures that compare the shape to the normal curve called Skewness and Kurtosis, which quantify the symmetry and the "tailedness" of the distribution, respectively.

5.2 Skewness

Definition

Skewness is a statistical number that tells us if a distribution is symmetric or not. Skewness indicates how much the shape of a distribution deviates from a normal distribution, which is perfectly symmetric.

If the skewness is positive, the right tail is longer, and the bulk of the data is concentrated on the left (right-skewed). If the skewness is negative, the left tail is longer, and the data is concentrated on the right (left-skewed). A skewness of zero suggests a symmetric distribution, such as the normal distribution.



Formulas of Measures

There are several measures of skewness:

1) **Karl Pearson's coefficient of skewness**

$$SK_P = \frac{Mean - Mode}{\sigma}$$

where σ= standard deviation

- In case the mode is indeterminate, then we take

$$SK_P = \frac{3(Mean - Median)}{\sigma}$$

2) **Bowley's coefficient of skewness**

This measure is based on quartiles. For a symmetrical distribution, it is seen that $Q1$ and Q_3 are equidistant from the median. Thus $(Q_3 - Med) - (Med - Q_1)$ can be taken as an absolute measure of skewness.

A relative measure of skewness, known as Bowley's coefficient (SK_B), is given by

$$SK_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

The above formula can be converted to

$$SK_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

3) **Kelly's coefficient of skewness**

Kelly developed another measure of skewness, which is based on percentiles or deciles.

- the formula of Kelly's coefficient of skewness based on percentiles is as follows

$$SK_K = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$$

- the formula of Kelly's coefficient of skewness based on deciles is as follows

$$SK_K = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

4) **Fisher's coefficient of Skewness**

Fisher's coefficient Skewness, also known as the third moment of the distribution, is a measure of the asymmetry of a data distribution relative to its mean. It is calculated as follows:

- For n values $x_1, x_2, \ldots, x_n$, the coefficient is

$$SK_F = \frac{1}{n.\sigma^3} \sum_{i=1}^{n} (x_i - \overline{X})^3$$

- For r values $x_1, x_2, \ldots, x_r$, associated with the frequencies $n_1, n_2, \ldots, n_r$ respectively, the coefficient is

$$SK_F = \frac{1}{N.\sigma^3} \sum_{i=1}^{r} n_i (x_i - \overline{X})^3$$

where σ = standard deviation , $N = \sum_{i=1}^{r} n_i$

Example 5.1

The marks of 10 students in the Mathematics exam are 8, 8, 9, 10, 12, 13, 13, 13, 14, 15
Determine the shape of this distribution in terms of skewness and its sign
1) Using the three measures of central tendency, then comparing them.
2) Using the Pearson coefficient.
3) Using the Fisher skewness coefficient.

Solution

1) Determination of the shape of the distribution in terms of skewness and its sign using the three measures of central tendency:

- The Mean is

$$\overline{X} = \frac{8+8+9+10+12+13+13+13+14+15}{10} = \mathbf{11.5}$$

- The Median is

$$Me = \frac{x_{(\frac{10}{2})} + x_{(\frac{10}{2}+1)}}{2} = \frac{x_5 + x_6}{2} = \frac{12+13}{2} = \mathbf{12.5}$$

- The mode is the most frequent value, which is $Mo = \mathbf{13}$

We notice that $\overline{X} < Me < Mo$. Hence, the distribution is skewed to the left (left-skewed), and negatively skewed.

2) Determination of the shape using the Pearson coefficient.

We need to determine the standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{10}(x_i - \overline{X})^2}{10}}$$

After substitution, we find that $\sigma = \mathbf{2.41}$

Pearson's coefficient of skewness is

$$SK_P = \frac{Me - Mo}{\sigma} = \frac{11.5 - 13}{2.41} = \mathbf{-0.62}$$

Since $SK_P < 0$, the distribution is left-skewed and negatively skewed.

3) Determination of the shape using the Fisher skewness coefficient.

The Fisher coefficient is

$$SK_F = \frac{1}{10}\sum_{i=1}^{10}\left(\frac{x_i - \overline{X}}{\sigma}\right)^3$$

After substitution, we find that Fisher's coefficient of skewness $SK_F = \mathbf{-0.025}$.
Since $SK_F < 0$, the distribution is left-skewed and negatively skewed.

Example 5.2

The data in this table relates to the time (in minutes) it takes 80 students from home to university.

Time	[0, 8[	[8, 16[	[16, 24[	[24, 32[	[32, 40[	[40, 48[
Number of students	12	15	20	14	10	9

1) Calculate the mean, the median, the mode and the standard deviation of times.
2) Determine the shape of this distribution in terms of skewness and its sign by comparing the three measures of central tendency.
3) Determine the shape of this distribution in terms of skewness and its sign using the Fisher skewness coefficient.

Solution

1) Let's create the following table for that data.

Class	[0, 8[	[8, 16[	[16, 24[	[24, 32[	[32, 40[	[40, 48[	$\sum$
n_i	12	15	20	14	10	9	80
c_i	4	12	20	28	36	44	/
$n_i c_i$	48	180	400	392	360	396	1776
$n_i ac$	12	27	47	61	71	80	/
$(x_i - \overline{X})^2$	331.24	104.04	4.84	33.64	190.44	475.24	/
$(x_i - \overline{X})^3$	- 6028.57	- 1061.21	- 10.65	195.11	2628.07	10360.23	/
$n_i(x_i - \overline{X})^2$	3974.88	1560.6	96.8	470.96	1904.4	4277.16	12284.8
$n_i(x_i - \overline{X})^3$	- 72342.82	- 15918.12	- 212. 96	2731.57	26280.72	93242.09	33780.48

■ The Mean is

$$\overline{X} = \frac{\sum_{i=1}^{6} n_i c_i}{\sum_{i=1}^{6} n_i} = \frac{1776}{80} = \mathbf{22.2}$$

■ The median class is $[16, 24[$ and the median is

$$Me = a + \frac{(\frac{N}{2} - n_{i-1} ac)}{n_i} \times l = 16 + \frac{40 - 27}{20} \times 8 = \mathbf{21.2}$$

■ The modal class is $[16, 24[$ and the mode is

$$Mo = a + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times l = 16 + \frac{5}{5 + 6} \times 8 = \mathbf{19.63}$$

■ The Standard Deviation is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{6} n_i (x_i - \overline{X})^2}{\sum_{i=1}^{6} n_i}} = \sqrt{\frac{12284.8}{80}} = \mathbf{12.39}$$

2) We notice that $\overline{X} > Me > Mo$. Hence, the distribution is skewed to the right (right-skewed), and positively skewed.

3) Determination of the shape using the Fisher skewness coefficient.

The Fisher coefficient is

$$SK_F = \frac{1}{\sigma^3.N}\sum_{i=1}^{6} n_i(x_i - \overline{X})^3 = \frac{33780.48}{(12.39)^3 \times 80} = \mathbf{0.22}$$

Since $SK_F > 0$, the distribution is right-skewed and positively skewed.

5.3 Kurtosis

Definition

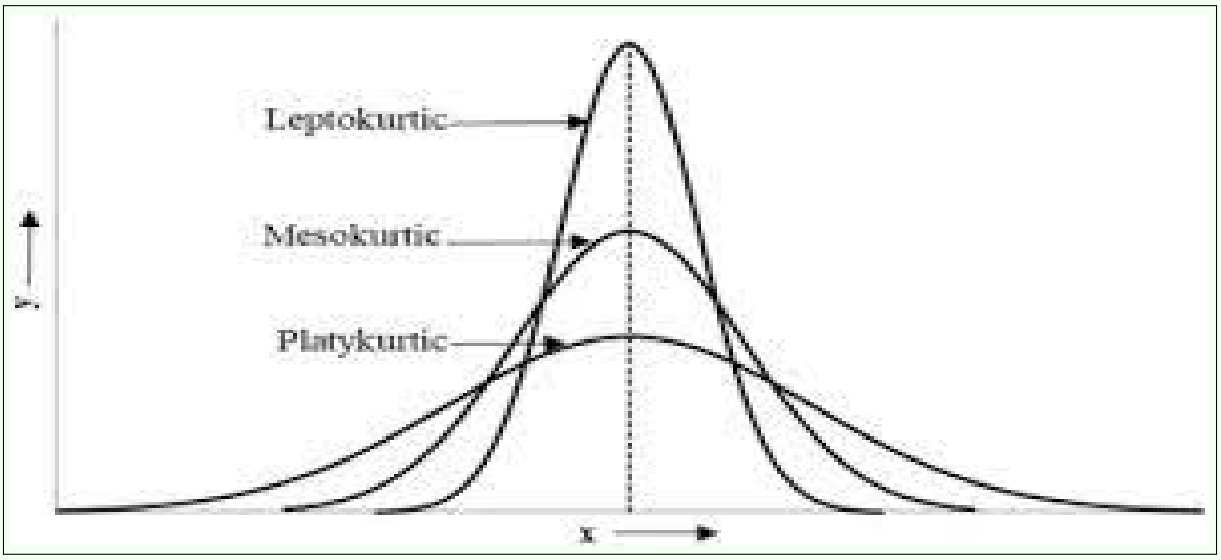
Kurtosis is a Greek word means bulginess. In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of Kurtosis of a distribution is measured relative to the peakedness of normal curve.

There are three types of distributions:

Leptokurtic: Sharply peaked with fat tails, and less variable.

Mesokurtic: Medium peaked

Platykurtic: Flattest peak and highly dispersed.



Formulas of Measures

The most important measures of Kurtosis are:

1) **Karl Pearson's Measures of Kurtosis**

The Karl Pearson's Measures of Kurtosis is calculated as follows:

- For n values $x_1, x_2, \ldots, x_n$, the coefficient is

$$K = \frac{1}{n.\sigma^4}\sum_{i=1}^{n}(x_i - \overline{X})^4$$

- For r values $x_1, x_2, \ldots, x_r$, associated with the frequencies $n_1, n_2, \ldots, n_r$ respectively, the coefficient is

$$K = \frac{1}{N.\sigma^4}\sum_{i=1}^{r} n_i(x_i - \overline{X})^4$$

where σ = standard deviation , $N = \sum_{i=1}^{r} n_i$

When $K = 3$, the curve is normal .

When $K > 3$, the curve is more peaked than the normal curve and is called as leptokurtic.

When $K < 3$, the curve is less peaked than the normal curve and is called as platykurtic curve.

2) **Kelly's Measure of Kurtosis**

Kelly has given a measure of kurtosis based on percentiles. The formula is given by

$$\boxed{K = \frac{P_{75} - P_{25}}{P_{90} - P_{10}}}$$

When $K = 0.263$, the curve is normal .

When $K > 0.263$, the curve is platykurtic .

When $K < 0.263$, the curve is leptokurtic.

Example 5.3

Let's give this data : 2, 3, 4, 5, 6

Find the Pearson's coefficient of Kurtosis. What do you conclude?

Solution

The Mean is

$$\overline{X} = \frac{2+3+4+5+6}{5} = \mathbf{4}$$

The Standard Deviation is

$$\sigma = \sqrt{\frac{(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2}{5}} = \mathbf{1.41}$$

the Pearson's coefficient of Kurtosis is

$$K = \frac{1}{5.(1.41)^4}\left[(2-4)^4 + (3-4)^4 + (4-4)^4 + (5-4)^4 + (6-4)^4\right] = \mathbf{1.7}$$

Since $K < 3$, the curve is platykurtic curve.

Chapter 6 Bivariate Descriptive Statistics

6.1 Introduction

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. For example, we can study a set of employees according to age and gender, or a set of individuals according to their weight and heigh

6.2 Generalities

let X and Y be two variables respectively possessing the values $x_1, x_2, \ldots, x_p$ and $y_1, y_2, \ldots. y_q$

Bivariate Series

the bivariate statistical series defined by X and Y is the set of pairs (x_i, y_j) associated with the N individuals of the population

Contingency Table

A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present the possible relationships between two sets of categorical data. Entries in the body of the table are called joint frequencies. The cells which contain the sum of the initial counts by row and by column are called marginal frequencies.

The bivariate statistical series can be presented as follows:

X \ Y	y_1	y_2	...	y_j	...	y_q	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2\bullet}$
⋮							⋮
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i\bullet}$
⋮							⋮
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet q}$	N

The frequency of the couple (x_i, y_j) is n_{ij}, and its relative frequency is $f_{ij} = \dfrac{n_{ij}}{N}$.
The marginal frequencies of X, denoted $n_{i\bullet}$, (respectively of Y, denoted $n_{\bullet j}$) are the sums of frequencies per column (respectively per rows).
The marginal relative frequencies of X, denoted $f_{i\bullet}$, (respectively of Y, denoted $f_{\bullet j}$) are the sums of relatives frequencies per column (respectively per rows).

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}, \quad n_{\bullet j} = \sum_{i=1}^{p} n_{ij}, \quad f_{i\bullet} = \sum_{j=1}^{q} f_{ij}, \quad f_{\bullet j} = \sum_{i=1}^{p} f_{ij}$$

☞ Remark

✿ The "sum of the row totals" equals the "sum of the column totals" . This value is also the sum of all counts from the interior cells, $\sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{j=}^{q}\sum_{i=1}^{p} n_{ij} = N$

✿ The sum of all relative frequencies is 1, $\sum_{i=1}^{p} f_{i\bullet} = \sum_{j=1}^{q} f_{\bullet j} = \sum_{j=}^{q}\sum_{i=1}^{p} f_{ij} = 1$

Marginal Distributions

The p couples $(x_i, n_{i\bullet})$ form the marginal distribution of variable X.
The q couples $(y_j, n_{\bullet j})$ form the marginal distribution of the variable$Y$

x_i	x_1	x_2	...	x_p
Marginal Frequency	$n_{1\bullet}$	$n_{2\bullet}$	...	$n_{p\bullet}$

y_j	y_1	y_2	...	y_q
Marginal Frequency	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet q}$

Conditional Distributions

A conditional distribution is a distribution of values for one variable that exists when you specify the values of other variables. The series obtained for column j, noted $X_{Y=y_j}$, is called the conditional series of X given $Y = y_j$. Similarly, the series obtained for column i, noted $Y_{X=x_i}$, is called the conditional series of Y given $X = x_i$.

x_i	x_1	x_2	...	x_p	**Total**
Frequency	n_{1j}	n_{2j}	...	n_{pj}	$n_{\bullet j}$

y_j	y_1	y_2	...	y_q	**Total**
Frequency	n_{i1}	n_{i2}	...	n_{iq}	$n_{i\bullet}$

Statistical Independence

We say that the variables X and Y are independent if, and only if for all the pair (i, j):

$$f_{ij} = f_{i\bullet} \times f_{\bullet j} \Leftrightarrow n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

Example 6.1

An experiment was carried out on 200 people to study the relationship between age X and sleep time Y, the following table was obtained

Y / X	$[6,8[$	$[8,10[$	$[10,12[$	$[12,14[$
$[1,3[$	0	1	4	31
$[3,11[$	0	3	5	20
$[11,19[$	1	8	21	14
$[19,31[$	0	25	a	2
$[31,59[$	22	13	7	0

1) Determine the value of a.
2) Determine the marginal distriutions of X and Y
3) Determine the conditional distributions of X when $8 \leqslant Y < 12$
4) Determine The conditional distributions of Y when $1 \leqslant X < 11$
5) Are the variables X and Y independent?

Solution

1) We know that the sum of all joint frequencies equals $N = 200$.
Then: $1+4+31+3+5+20+1+8+21+14+25+a+2+22+13+7 = 200$
$a + 177 = 200 \Rightarrow a = 23$

2) Let us form this table to facilitate the calculations

x_i / y_j	$[6,8[$	$[8,10[$	$[10,12[$	$[12,14[$	$n_{i\bullet}$
$[1,3[$	0	1	4	31	36
$[3,11[$	0	3	5	20	28
$[11,19[$	1	8	21	14	44
$[19,31[$	0	25	23	2	50
$[31,59[$	22	13	7	0	42
$n_{\bullet j}$	23	50	60	67	200

The marginal distributions of X:

x_i	$[1,3[$	$[3,11[$	$[11,19[$	$[19,31[$	$[31,59[$	**Total**
$n_{i\bullet}$	36	28	44	50	42	200
$f_{i\bullet}$	0.18	0.14	0.22	0.25	0.21	1

The marginal distributions of Y:

y_j	$[6,8[$	$[8,10[$	$[10,12[$	$[12,14[$	**Total**
$n_{\bullet j}$	23	50	60	67	200
$f_{\bullet j}$	0.115	0.25	0.30	0.335	1

3) The conditional distribution of X when $8 \leqslant Y < 12$:

x_i	$[1,3[$	$[3,11[$	$[11,19[$	$[19,31[$	$[31,59[$	$\sum$
Frequency=h_i	1+4=5	3+5=8	8+21=29	25+23=48	13+7=20	110

4) The conditional distribution of Y given $1 \leqslant X < 11$:

y_j	$[6,8[$	$[8,10[$	$[10,12[$	$[12,14[$	$\sum$
Frequency=k_j	0+0=0	1+3=4	4+5=9	31+20=51	64

5) We have $n_{11} = 0$ and $\dfrac{n_{1\bullet} \times n_{\bullet 1}}{N} = \dfrac{36 \times 23}{200} \neq 0$. Then X and Y are not independent

6.3 Numerical Description

In a bivariate statistical series (X, Y), the means and variances are given respectively by

Means and Variances

The marginal means are: $\overline{X} = \dfrac{1}{N}\sum_{i=1}^{p} n_{i\bullet}x_i$ and $\overline{Y} = \dfrac{1}{N}\sum_{i=1}^{q} n_{\bullet j}y_j$

The conditional means are: $\overline{X}_{Y=y_j} = \dfrac{1}{n_{\bullet j}}\sum_{i=1}^{p} n_{ij}x_i$ and $\overline{Y}_{X=x_i} = \dfrac{1}{n_{\bullet j}}\sum_{i=1}^{q} n_{ij}y_j$

The marginal variances are: $V_X = \dfrac{1}{N}\sum_{i=1}^{p} n_{i\bullet}x_i^2 - \overline{X}^2$ and $V_Y = \dfrac{1}{N}\sum_{i=1}^{q} n_{\bullet j}y_j^2 - \overline{Y}^2$

The standard deviations of X and Y are given, respectively, by: $\sigma_X = \sqrt{V_X}$ and $\sigma_Y = \sqrt{V_Y}$

Example 6.2

Let us return to the example 4.1 considered earlier, calculate respectively:

1) The marginal means, marginal variances of X and Y
2) The conditional mean of X when $8 \leqslant Y < 12$
3) The conditional mean of Y when $1 \leqslant X < 11$

Solution

1) Let us form the marginal distributions to facilitate the calculations

x_i \ y_j	7	9	11	13	$n_{i\bullet}$	$n_{i\bullet}x_i$	$n_{i\bullet}x_i^2$
2	0	1	4	31	36	72	144
7	0	3	5	20	28	196	1372
15	1	8	21	14	44	660	9900
25	0	25	23	2	50	1250	31250
45	22	13	7	0	42	1890	85050
$n_{\bullet j}$	23	50	60	67	200	4068	127716
$n_{\bullet j}y_j$	161	450	660	871	2142		
$n_{\bullet j}y_j^2$	1127	4050	7260	11323	23760		

The marginal means: $\overline{X} = \frac{1}{N}\sum_{i=1}^{5} n_{i\bullet}x_i = \frac{4068}{200} = 20.34$

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{4} n_{\bullet j}y_j = \frac{2142}{200} = 10.71$$

The marginal varicenes: $V_X = \frac{1}{N}\sum_{i=1}^{5} n_{i\bullet}x_i^2 - \overline{X}^2 = \frac{127716}{200} - 20.34^2 = 224.86$

$$V_Y = \frac{1}{N}\sum_{i=1}^{4} n_{\bullet j}y_j^2 - \overline{Y}^2 = \frac{23760}{200} - 10.71^2 = 4.10$$

2) To determine the conditional mean of X when $8 \leqslant Y < 12$, it is enough to form this table

x_i	2	7	15	25	45	$\sum$
Frequency=h_i	1+4=5	3+5=8	8+21=29	25+23=48	13+7=20	110
h_ix_i	10	56	435	1200	900	2601

$$\overline{X}_{8\leqslant Y<12} = \frac{\sum h_ix_i}{\sum h_i} = \frac{2601}{110} = 23.65$$

3) Similarly, we form this table to determine the conditional mean of Y given $1 \leqslant X < 11$

y_j	7	9	11	13	$\sum$
Frequency=k_j	0+0=0	1+3=4	4+5=9	31+20=51	64
k_jy_j	0	36	99	663	798

$$\overline{Y}_{1\leqslant X<11} = \frac{\sum k_jy_j}{\sum k_j} = \frac{798}{64} = 12.47$$

Covariance

The covariance between X and Y, denoted by $Cov(X, Y)$ is defined as

$$Cov(X,Y)=\overline{(X-\overline{X})(Y-\overline{Y})}=\frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}(x_i-\overline{X})(y_j-\overline{Y})$$

Practical formula for calculation

$$Cov(X,Y)=(\overline{XY})-(\overline{X}).(\overline{Y})=\frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}x_iy_j-(\overline{X}).(\overline{Y})$$

☞ Remark

The covariance satisfies the following properties

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

$Cov(X,X)=V_X$

$Cov(X+Y,Z)=Cov(X,Z)+Cov(Y,Z)$

$Cov(aX+b,cY+d)=acCov(X,Y)$

Example 6.3

Find the covariance between x and Y for this distribution

x_i \ y_j	- 2	0	2	3
2	3	4	0	6
3	4	3	3	2
4	2	3	3	2

Solution

Let us form the marginal distributions

y_j	$n_{\bullet j}$	$n_{\bullet j}y_j$
- 2	9	- 18
0	10	0
2	6	12
3	10	30
$\sum$	35	24

x_i	$n_{i\bullet}$	$n_{i\bullet}x_i$
2	13	26
3	12	36
4	10	40
$\sum$	35	102

$\overline{X}=\frac{1}{N}\sum_{i=1}^{3}n_{i\bullet}x_i=\frac{102}{35}=2.91$ and $\overline{Y}=\frac{1}{N}\sum_{i=1}^{4}n_{\bullet j}y_j=\frac{35}{35}=0.68$

Then, we form this table to facilitate the calculations

x_i \ y_j	-2	0	2	3	$\sum_{j=1}^{4} n_{ij}y_j$	$x_i \sum_{j=1}^{4} n_{ij}y_j$
2	3	4	0	6	12	24
3	4	3	3	2	4	12
4	2	3	3	2	8	32
						68

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - (\overline{X}).(\overline{Y}) = \frac{68}{35} - (2.91)(0.68) = -0.036$$

Correlation Coefficient

The correlation coefficient between x and Y , denoted by $\rho(X,Y)$ is defined as

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

☞ Remark

The correlation coefficient satisfies the following proprieties

- ✿ The correlation coefficient $\rho(X,Y)$ is a number in the interval $[-1,1]$
- ✿ $\rho(X,Y)$ close to $+1 \Leftrightarrow$ a strong positive linear correlation (large x values go with large y).
- ✿ $\rho(X,Y)$ close to $-1 \Leftrightarrow$ a strong negative linear correlation (large x values go with small y).
- ✿ $\rho(X,Y)$ close to $0 \Leftrightarrow$ no or weak linear correlation. Other nonlinear trends may be possible.

Example 6.4

Return to example 4.3 and determine $\rho(X,Y)$

Solution

Firstly we must calculate marginal variances V_X and V_Y, for this we form this table

x_i \ y_j	-2	0	2	3	$n_{i\bullet}$	$n_{i\bullet}x_i^2$
2	3	4	0	6	13	52
3	4	3	3	2	12	108
4	2	3	3	2	10	160
$n_{\bullet j}$	9	10	6	10	35	320
$n_{\bullet j}y_j^2$	36	0	24	90	150	

The marginal varicenes: $$V_X = \frac{1}{N}\sum_{i=1}^{3} n_{i\bullet}x_i^2 - \overline{X}^2 = \frac{320}{35} - (2.91)^2 = 0.67$$

$$V_Y = \frac{1}{N}\sum_{i=1}^{4} n_{\bullet j} y_j^2 - \overline{Y}^2 = \frac{150}{35} - (0.68)^2 = 3.82$$

The correlation coefficient: $\rho(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y} = \dfrac{-0.036}{\sqrt{0.67}.\sqrt{3.82}} = -0.023$

6.4 Scatter plots and Linear Adjustments

Scotter plots

Scatter plots are a graphical way of displaying the relationship between two variables. Each data point is plotted as a dot, with the x-coordinate representing the value of one variable and the y-coordinate representing the value of the other variable.

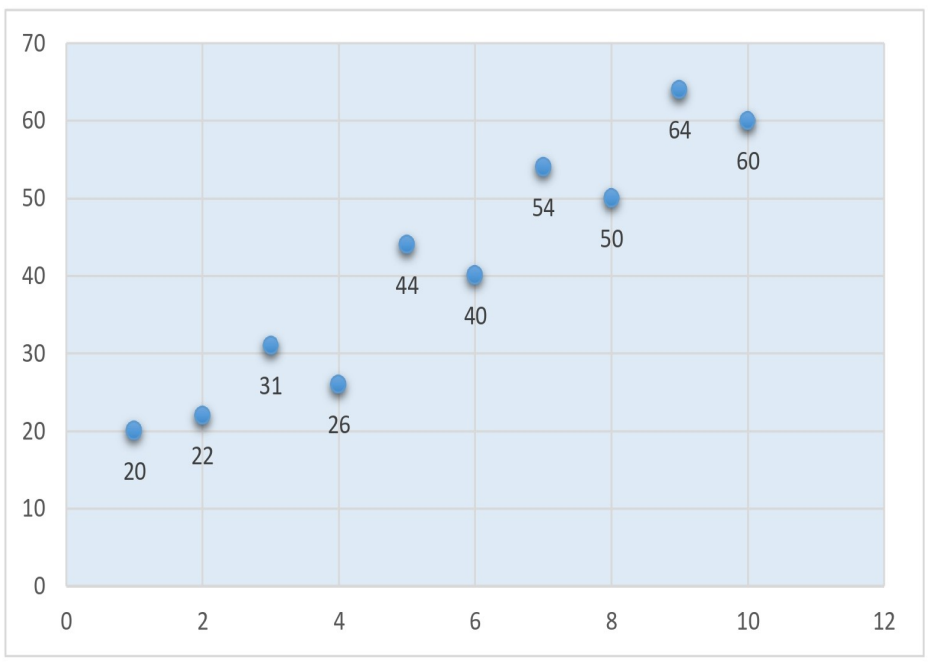
Scatter plots can be used to identify patterns and trends in the data. For example, if there is a positive correlation between the two variables, the points will tend to cluster in the upper right and lower left quadrants of the plot. If there is a negative correlation, the points will tend to cluster in the upper left and lower right quadrants of the plot.

Example 6.5

The following table gives the number of agricultural operations in a region according to their surface area in hectares

Surface X	1	2	3	4	5	6	7	8	9	10
Number of Operations Y	20	22	31	26	44	40	54	50	64	60

we can plot the previous data through this scatter plot



The mean point is denoted by $G(\overline{X}, \overline{Y})$, where $\overline{X} = \dfrac{\sum x_i}{10} = 5.5$ and $\overline{Y} = \dfrac{\sum y_j}{10} = 41.1$

Linear Adjustements

Adjusting (fitting) a set of points consists of selecting a simple curve that is as close as possible to this scatter plot. When the points on a scatterplot are close together, we can look for a straight line that reflects that alignment. This line is called the adjustment line, regression line or a trend line (also known as a line of best fit)

A line of best fit is a straight line that minimizes the distance between itself and a set of data points. It can be used to describe the relationship between two variables, to make predictions about one variable given the value of the other variable, and to compare different datasets. It can be calculated using a variety of methods, including Mayar method and the least squares method. The least squares method finds the line that minimizes the sum of the squared distances between the data points and the line.

Mayer Line Method

The Mayer line method is a procedure for generating a regression line for a given scatter plot by calculating means (averages).

The following steps are used to find the rule of the Mayer line and to make predictions from a bivariate data set.

- Order the coordinates according to the independent variable.
- Separate the distribution into 2 equal groups, if possible.
- Calculate the mean points of each group (G_1 and G_2).
- The line (G_1G_2) is the adjustment line using the Mayer method.

Example 6.6

Return to example 4.5 and determine the equation of the adjustment line using the Mayer method. What do you predict about the number of agricultural operations if the surface area is 20 hectares? plot the Mayer's line

Solution

- The distribution contains 10 pairs of data. The five pairs whose surface area ranges from 1 to 5 hectares make up the first group. The other five pairs form the second group.
- Find the mean of the x and y values of each group to make 2 points G_1 and G_2.

$$\overline{X_1} = \frac{1+2+3+4+5}{5} = 3 \text{ and } \overline{Y_1} = \frac{20+22+31+26+44}{5} = 28.6$$

$$\overline{X_2} = \frac{6+7+8+9+10}{5} = 8 \text{ and } \overline{Y_2} = \frac{40+54+50+64+60}{5} = 53.6$$

Then $G_1(3, 28.6)$ and $G_2(8, 53.8)$

- Find the rule of the regression line that passes through the points G_1 and G_2.

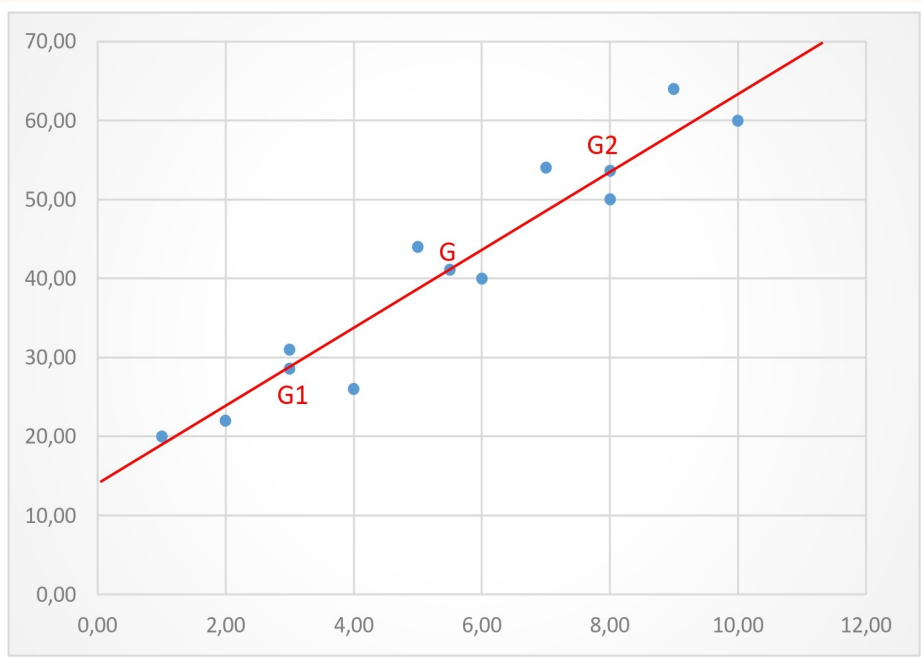
Since this is a straight line, the rule has the form $y = ax + b$. First we calculate the slope (a)

$$a = \frac{\overline{Y_2} - \overline{Y_1}}{\overline{X_2} - \overline{X_1}} = \frac{53.6 - 28.6}{8 - 3} = 5$$

Next, we replace a by 5 and the x and y variables by the coordinates of one of the 2 points. Then, we isolate b. $y = 5x + b$ and $b = 28.6 - 5(3) = 13.6$

✿ The rule of the Mayer line is: $y = 5x + 13.6$

If the surface area is 20 hectares, the number of operations is $y = 5(20) + 13.6 \simeq 114$
The line of Mayer is as follows



6.5 Nonlinear Adjustments

In some cases, linear adjustment is not justified (low or no linear correlation). This does not exclude a (non-affine) dependence between the two variables.

Exponential Adjustment

We suppose that the variables X and Y are related by an exponential relationship of the form:

$$y = be^{ax}$$

In this case, this equation can be transformed by passing to logarithms:

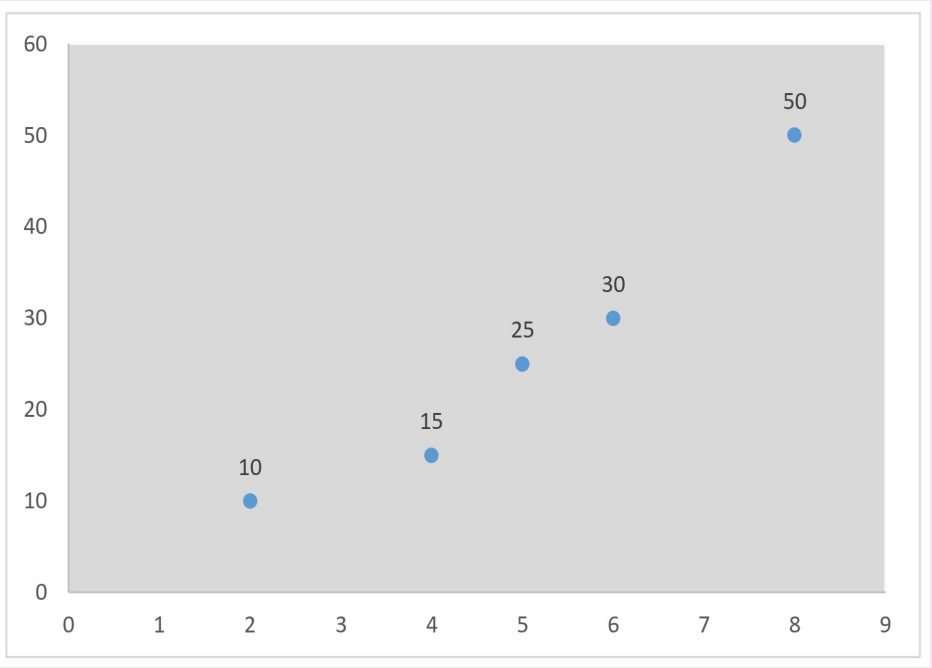
$$\ln y = z = \ln b + ax$$

Then, we return to a linear adjustment between X and $Z = \ln(Y)$

$$a = \frac{Cov(X,Z)}{V(X)} \text{ and } \ln b = \overline{Z} - a\overline{X}$$

Example 6.7

Adjust the scatter plot below by an exponential function



Solution

Firstly we must form this table

i	x_i	y_i	$z_i = \ln y_i$	x_i^2	$x_i z_i$
1	2	10	2.30	4	4.60
2	4	15	2.71	16	10.84
3	5	25	3.22	25	16.10
4	6	30	3.40	36	20.40
5	8	50	3.91	64	31.28
$\sum$	25	130	15.54	145	83.22

Let's determine the parameters of this model

$$\overline{X} = \frac{1}{N}\sum x_i = \frac{25}{5} = 5, \quad V(X) = \frac{1}{N}\sum x_i^2 - (\overline{X})^2 = \frac{145}{5} - (5)^2 = 4$$

$$\overline{Z} = \frac{1}{N}\sum z_i = \frac{15.54}{5} = 3.108$$

$$Cov(X,Z) = \frac{1}{N}\sum x_i z_i - (\overline{X})(\overline{Z}) = \frac{83.22}{5} - (5)(3.108) = 1.104$$

$$a = \frac{Cov(X,Z)}{V(X)} = \frac{1.104}{4} = 0.276$$

$$\ln b = \overline{Z} - a\overline{X} = 3.108 - (0.276)(5) = 1.728 \Rightarrow b = 5.629$$

Therefore, the adjustment function is :

$$y = 0.28e^{5.63x}$$

Adjustment by a Power Function

We suppose that the variables X and Y are related by an exponential relationship of the form:

$$y = bx^a$$

In this case, this equation can be transformed by passing to logarithms:

$$\ln y = \ln b + a \ln x$$

Then, we return to a linear adjustment between $\ln X$ and $\ln(Y)$, and we can use the formulas presented previously:

$$a = \frac{Cov(\ln X, \ln Y)}{V(\ln X)} \text{ and } \ln b = \overline{\ln Y} - a\overline{\ln X}$$

and therefore we can deduce the parameters of the power model

Example 6.8

A group of 10 people is studied according to two statistical characteristics: their masses (in kilograms) and their heights (in centimeters). The results obtained are recorded in the table below:

Mass(kg)	60	65	65	70	72	75	78	80	80	88
Height(cm)	153	164	166	172	174	178	179	181	182	186

Adjust the above distributions by a power function.

Solution

Firstly we must form this table

i	x_i	y_i	$\ln x_i$	$\ln y_i$	$(\ln x_i)^2$	$(\ln x_i)(\ln y_i)$
1	60	153	4.094	5.030	16.761	20.593
2	65	164	4.174	5.100	17.426	21.287
3	65	166	4.174	5.112	17.422	21.337
4	70	172	4.249	5.148	18.054	21.874
5	72	174	4.277	5.195	18.293	22.219
6	75	178	4.317	5.182	18.636	22.371
7	78	179	4.357	5.187	18.983	22.600
8	80	181	4.382	5.199	19.202	22.782
9	80	182	4.382	5.204	19.202	22.804
10	88	186	4.477	5.226	20.044	23.397
$\sum$	×	×	42.883	51.583	184.023	221.264

Let's determine the parameters of this model

$$\overline{\ln X} = \frac{1}{N}\sum \ln x_i = \frac{42.883}{10} = 4.288, \quad V(\ln X) = \frac{184.023}{10} - (4.288)^2 = 0.015$$

$$\overline{\ln Y} = \frac{1}{N}\sum \ln y_i = \frac{51.583}{10} = 5.158$$
$$Cov(\ln X, \ln Y) = \frac{1}{N}\sum \ln x_i \ln y_i - (\overline{\ln X})(\overline{\ln Y}) = \frac{221.264}{10} - (4.288)(5.158) = 0.0089$$
$$a = \frac{Cov(\ln X, \ln Y)}{V(\ln X)} = \frac{0.0089}{0.015} = 0.593$$
$$\ln b = \overline{\ln Y} - a\overline{\ln X} = 5.158 - (0.593)(4.288) = 2.615 \Rightarrow b = 13.667$$

Therefore, the adjustment function is :

$$y = 13.667x^{0.593}$$

Parabolic Adjustment

We suppose that the points $M_i(x_i, y_i)$ are located approximately at a parabola of equation:

$$y = ax^2 + bx + c$$

We determine the values of a, b and c by minimizing the function:

$$D(a,b,c) = \sum_{i=1}^{n}(ax_i^2 + bx_i + c - y_i)^2$$

It implies that: $\frac{\partial D}{\partial a}(a,b,c) = 0$, $\frac{\partial D}{\partial b}(a,b,c) = 0$ and $\frac{\partial D}{\partial c}(a,b,c) = 0$.

$$\frac{\partial D}{\partial a} = 2\sum_{i=1}^{n} x_i^2(ax_i^2 + bx_i + c - y_i), \quad \frac{\partial D}{\partial b} = 2\sum_{i=1}^{n} x_i(ax_i^2 + bx_i + c - y_i)$$

and $$\frac{\partial D}{\partial c} = 2\sum_{i=1}^{n}(ax_i^2 + bx_i + c - y_i)$$

The cancellation of these three partial derivatives leads us to solve the following system:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cN = \sum_{i=1}^{n} y_i \end{cases}$$

Example 6.9

We consider the following statistical series with two variables X and Y:

X	2	4	5	6	8	10
Y	10	25	30	32	25	12

Adjust the above distributions by a function of the form: $y = ax^2 + bx + c$

Solution

Let's form this table

i	x_i	y_i	x_i^2	x_i^3	x_i^4	x_iy_i	$x_i^2y_i$
1	2	10	4	8	16	20	40
2	4	25	16	64	256	100	400
3	5	30	25	125	625	150	750
4	6	32	36	216	1296	192	1152
5	8	25	64	512	4096	200	1600
6	10	12	100	1000	10000	120	1200
$\sum$	35	134	245	1925	16289	782	5142

Using the least squares method, we seek to solve the following system of equations:

$$\begin{cases} 16289a + 1925b + 245c &= 5142 \\ 1925a + 245b + 35c &= 782 \\ 245a + 35b + 6c &= 134 \end{cases}$$

Whose solution is:

$$a = -1.264, \quad b = 15.358, \quad c = -15.636$$

Therefore, the adjustment function is :

$$y = -1.264x^2 + 15.358x - 15.636.$$

Bibliography

[1] Ross, Sheldon M. Introductory statistics. Academic Press, 2017.

[2] Holický, Milan. Introduction to probability and statistics for engineers. Springer Science and Business Media, 2013.

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

[4] Moore, David S, George P, McCabe. Introduction to the practice of statistics. WH Freeman/Times Books/Henry Holt and Co, 1989.

[5] Bluman, Allan. Elementary Statistics: A step by step approach 9e. McGraw Hill, 2014.

[6] Touati M S, Educational publication on descriptive statistics (in Arabic) for LMD students of University of El Oued, 2016.