

# Principles of PROBABILITY AND STATISTICS

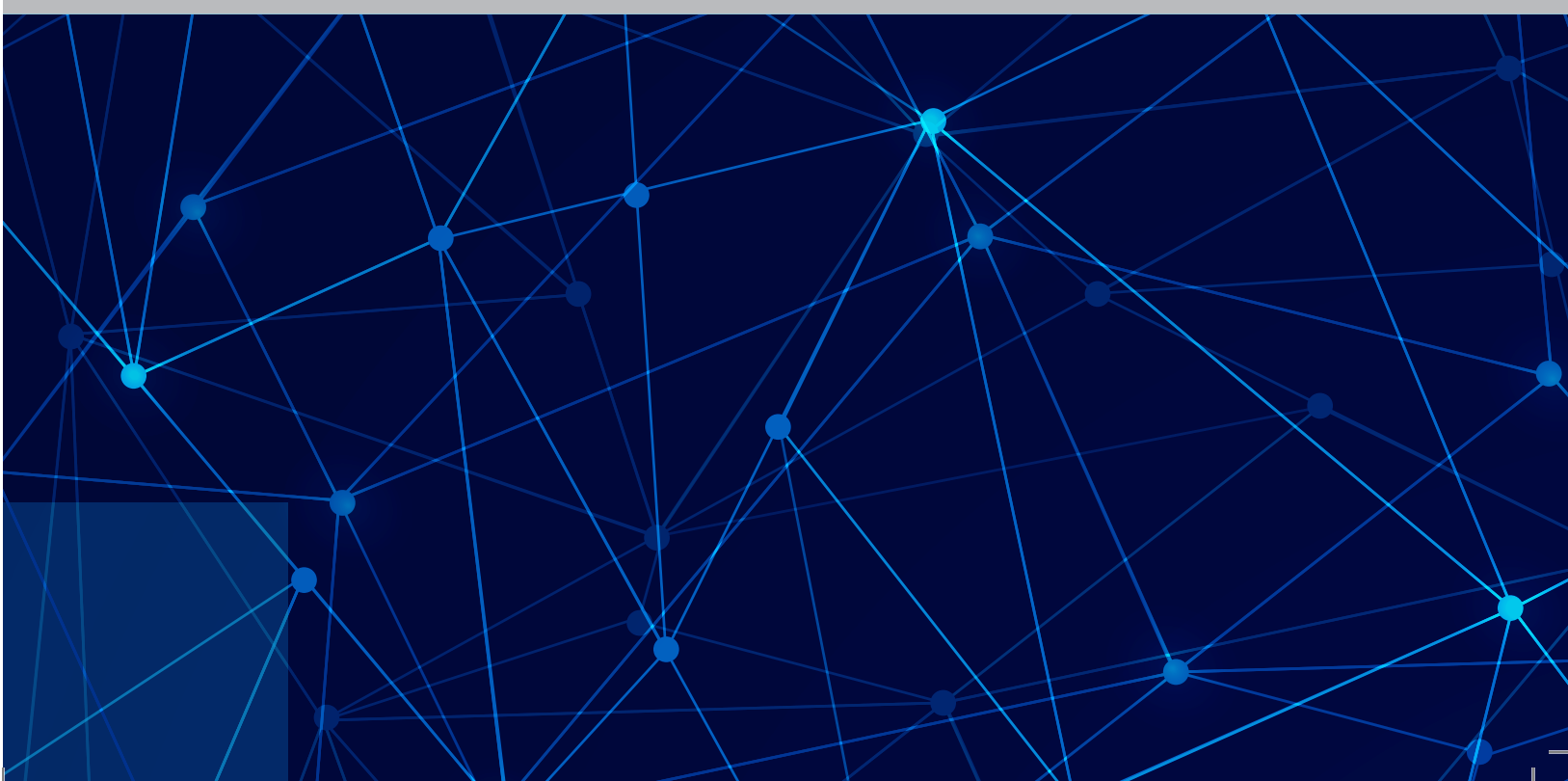
ABDULRAHMAN ABOUAMMOH

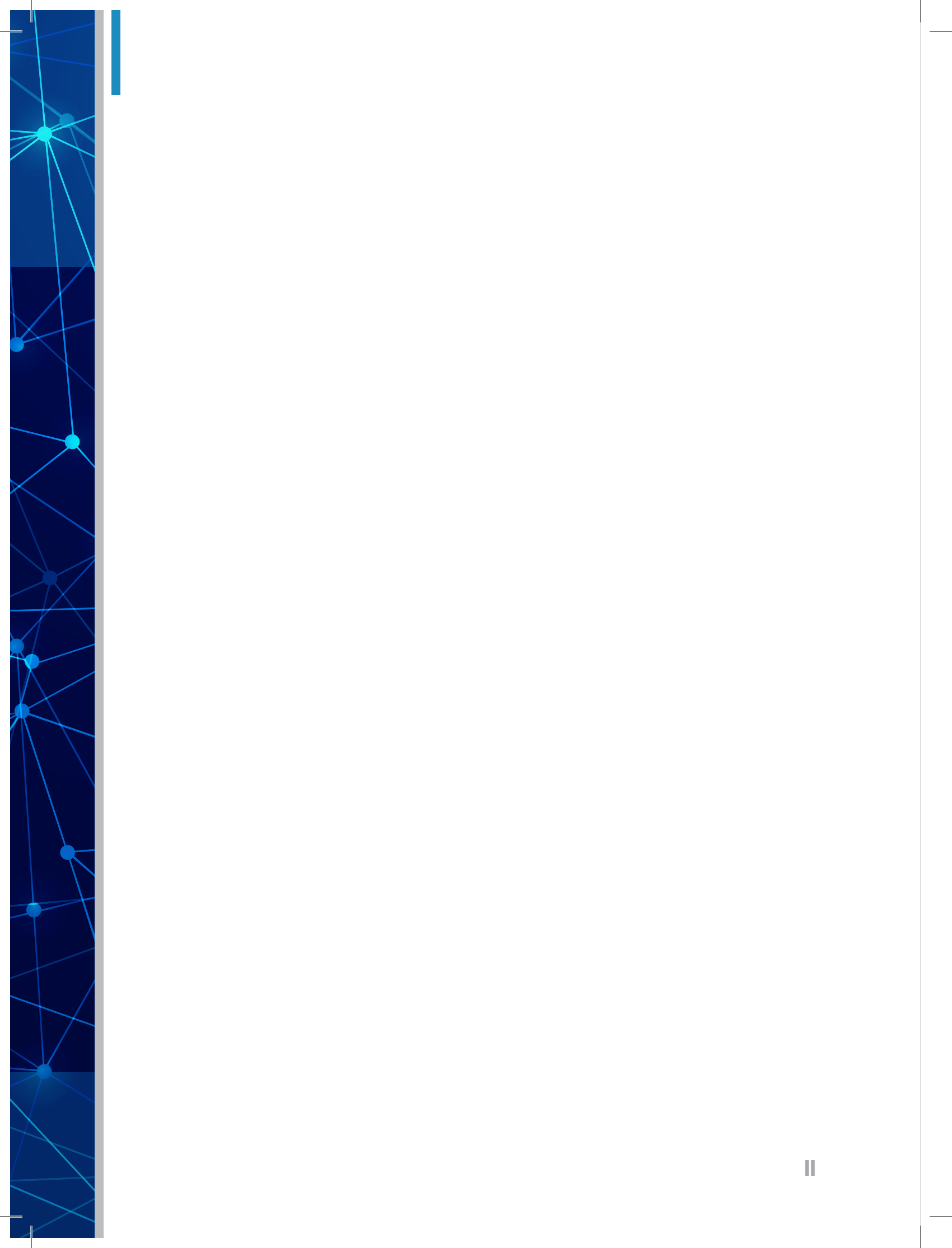
KHALAF SULTAN

MOHAMED KAYID

MANSOUR SHRAHILI

KING SAUD UNIVERSITY





## TABLE OF CONTENTS

<i>Table of Contents</i>	<i>iii</i>
<i>Introduction</i>	<i>vi</i>
<i>Acknowledgements</i>	<i>viii</i>



<b>CHAPTER 1 •• Descriptive Statistics 1</b>	
1.0 Introduction	2
1.2 Basic Concepts and Definitions	4
1.3 Organization and Graphical Representation of the Data	8
1.4 Graphical Representation	16
1.5 Measures of Central Tendency	23
1.6 Measures of Dispersion	40
<b>EXERCISES</b>	<b>46</b>



<b>CHAPTER 2 •• Probability 55</b>	
2.1 Mathematical Concepts	56
2.2 Definitions And Concepts In Probabilty Calculas	60
2.3 Concept Of Probability Function	69
2.4 Conditional Probabilty And Independence Of Event	76
<b>EXERCISES</b>	<b>90</b>

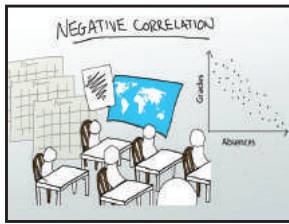


<b>CHAPTER 3 •• Random Variables and Probability Distributions 95</b>	
3.1 Concept Of Random Variables And Their Distributions	96
3.2 Discrete Random Variables and their Distributions	99
3.3 Continuous Random Variables and their Distributions	119
<b>EXERCISES</b>	<b>133</b>



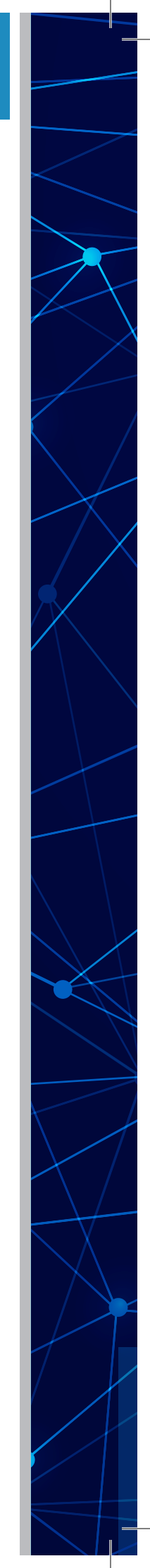
<b>CHAPTER 4 •• Introduction to Statistical Inference 139</b>	
4.1 Definitions and Concepts	140
4.2 Estimation of the Population Mean	147
4.3 Estimation of The Population Proportion	153
4.4 Introduction to Hypotheses Testing	161
4.5 Hypothesis Testing for the Population Mean	165
4.6 Hypothesis Testing for the Population Proportion	173
<b>EXERCISES</b>	<b>177</b>

## TABLE OF CONTENTS



<b>CHAPTER 5 • • Correlation and Regression</b>	<b>183</b>
5.0 Introduction	184
5.1 Linear Correlation Coefficient	186
5.2 Simple Linear Regression	191
<b>EXERCISES</b>	<b>202</b>

<i>References</i>	209
<i>Normal Distribution Table</i>	210



# INTRODUCTION

This one-term foundation course in statistics and probability is written for the first year student at King Saud University. This course is intended for students majoring in science, engineering, architectures, computer science, agriculture, business, actuarial science, finance and operations research.

Our main aim is to present basic statistics and probability concepts in natural and gradual steps. We have done our best present abundant of well-selected interesting examples and exercises. Also, theorems and proofs methods have been avoided.

Thus, these course main objectives are to enhance the understanding of the basic concepts of statistics and probability, to increase the students' computational skills, to strengthen their probabilistic and statistical background and to give them practical flavor of the wide range and interesting application of this branch of science in various fields of life.

The authors have faced two appealing and difficult compromise-able desires. One is to put too much material, give deeper content and utilize stringent mathematical treatment that it is felt necessary for statisticians, probability analysts and other practitioners. On the other hand, the material of this book has to be accessible to audience of freshmen who have just finished their high schools and need academic support, motivation and smooth transition to college life.

We have used our experience in teaching similar statistics and probability courses on various levels of university stages to preserve mathematical correctness and rigor and evade tedious, cumbersome or advanced mathematical manipulation.

The book comprises of five chapters and covers most necessary knowledge and skills for freshmen students. Learning objectives and expected specific comprehension and abilities, by students, are stated at the beginning of every chapter. Instructors are requested to check how much the present material of the book has participated in achieving these objectives and the authors would be grateful for any feedback in this context. Readers are advised to study the five chapters in their sequel of presentation.

We have intended to give more exercises that are similar to the ideas presented in the examples of every chapter for affirming general skills. Also, additional exercises are included to give students deeper understanding of various concepts, improve their calculation ability, enhance their thinking skills and show a wider range of applications of the subject in every chapter.

Chapter one gives basic descriptive statistics that include graphical and tabular presentation of sample data, summarizing data, measures of central tendency and measures of dispersions, Chapter two presents the concepts of probability, random experiments, axioms of probability, possible outcomes of experiments, counting techniques, conditional probability and independence. Chapter three covers the ideas of discrete and continuous random variable, some commonly used discrete and continuous distributions, using the normal table and the distribution of

sample mean or sample proportion. Chapter four discusses the point and interval estimation of the population mean and proportion and includes the testing hypothesis for the mean and proportions. Finally chapter five abridges the earlier ideas for two related populations by introducing the simple linear correlation. Looks into scatter plots and derives the simple linear regression model.

**The Authors**  
**August 2017**

## ACKNOWLEDGMENTS

We would like to express our sincere thanks to Dr. Nami Aljehani, Dean of The Common First Year at King Saud University, and Dr. Abdulmajeed Aljerawi, Vice Dean for Academic Affairs, for their support and continuous encouragement throughout this work.

We would like to thank Prof. Dr. Hamid Al-Oklah for his reviewing, revising and realization for this book.

We thank Mr. Fadi Hasan for his support in editing, reviewing, designing and producing figures.

Finally, we are grateful for useful conversation with all faculty members in Statistics and Operations Research Department and in Basic Sciences Department at King Saud University.



# CHAPTER 1

## DESCRIPTIVE STATISTICS



### LEARNING OBJECTIVES

After completing this chapter, you should be able to:

1. Distinguish between quantitative and qualitative data.
2. Organize types of raw data into tables.
3. Represent data into different types of graphs.
4. Calculate some measures of location and interpret the meaning.
5. Calculate some measures of dispersion.

- SECTION 1.0 INTRODUCTION
- SECTION 1.1 BASIC CONCEPTS AND DEFINITIONS
- SECTION 1.2 ORGANIZATION AND GRAPHICAL REPRESENTATION OF THE DATA
- SECTION 1.3 GRAPHICAL REPRESENTATION
- SECTION 1.4 MEASURES OF CENTRAL TENDENCY
- SECTION 1.5 MEASURES OF DISPERSION

## *Section 1.0*

# INTRODUCTION

We live in an age which is referred to as the information age. We come across a wide variety of information in our day to day life in the form of graphs, facts, news, tables. The major sources of information include magazines, newspaper, televisions and various means of communication. This information may be related to profit history of a firm, growth of a nation, weather conditions, personal information, physical characteristics of human and many more.

A collection of information collected by means of experiments or real life events and stored in a proper format is called as **data**. The word data is derived from a Latin word 'datum'. There are many ways of collecting data some of which are experiments, interviews and questionnaires.

For instance, students' performance data (include number of students, marks obtained, age), country's growth data (include literacy rate, crime rates, health conditions), company's performance data (include sales, revenue generated, profit, share price).

Our lives, nowadays, are becoming more and more data oriented. We believe that data leads to power and success. Earlier when the technology was not well equipped only humans could collect data and the volume of data was very small but with the help of innovative technologies like computers, internet, digital storage, we are now able to collect and store huge volumes of data efficiently.

Every part of our lives utilizes data in one form or the other and helps us in the decision making process. For example, for buying a new car we look at the performance data of different cars which is obtained from people who have been using cars from a significant time. Similarly, for investing in a particular stock at Saudi Tadwul we look at the historic performance of that stock.

With the advancement in research and technology, we are now able to use historic data to predict the upcoming events. We can now forecast the weather conditions up to three days or one week from the current day using the weather condition data about current and previous days. There are numerous fields where data has proven its importance including biology, social sciences, and business.

It is therefore essential for us to know the ways of extracting meaningful information from such data. These ways of extracting information are studied in a branch of mathematical sciences called statistics which is derived from a Latin word 'status' meaning a state.

Statistics is a branch of science deals with collection, organization, presentation, analysis and interpretation of data.

## Section 1.1

# BASIC CONCEPTS AND DEFINITIONS

Earlier statistics was only confined to collection of data which was useful to only state but with the advancement in time, statistics' scope broadened and it was not only restricted to the collection and presentation of data but also concerned with interpretation of data and making inferences from the data. Statistical methods are subdivided in two main categories:

### DEFINITION 1.1.1 (Descriptive Statistics)

Those statistical methods or techniques which are used for presenting and summarizing data in either tables or graphs form. It includes construction of charts, graphs and tables and calculation of averages, percentiles, dispersions and other descriptive measures.

► **EXAMPLE 1.1.1** The following are types of descriptive data:

- a. Data involving human physical characteristics include average height, average weight.
- b. Human population data includes the age-wise, gender-wise proportion of the population etc.

### DEFINITION 1.1.2 (Inferential Statistics)

Those statistical methods or techniques which are used for making conclusions or inferences about the entire population using the observation from the samples by using the. It includes point estimation, interval estimation, hypothesis testing, statistical modeling, clustering and many more methods based on probability theory.

► **EXAMPLE 1.1.2** If we want to measure the income of all the citizens of The Kingdom of Saudi Arabia, (KSA) it is not feasible to measure income of all citizens individually. We therefore draw a sample from the KSA population and make conclusions about the income of every citizen by using the sample.

Both the descriptive and the inferential statistics are closely related to each other. It is a common practice to look at the organized and summarized data obtained using descriptive statistics to select the appropriate inferential method to be used.

Statistics, nowadays, is applicable in every field of science and can be used to solve problems in the field of medical, agriculture, politics, economics and technology. The following topic will introduce various concepts used in statistics.

**DEFINITION 1.13 (Population)**

It is a set of all individuals, persons, objects or historical events which are of some interest to the statistician to make inferences for a specific problem or experiment.

► **EXAMPLE 1.1.3**

The following sets represent populations:

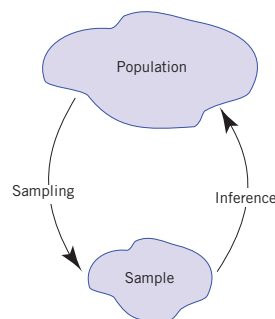
- i. A set of all students in a country,
- ii. A set of income of all citizens of a country.

Population as a whole is never used to collect information and make inferences due to its size. Always a subset of the population which reflects all the major characteristics of the population is used for collecting the information.

**DEFINITION 1.1.4 (Sample)**

A sample is a subset of population which is used to collect information and to make inferences about the entire population.

Figure 1.1.1 below illustrates the relationship between population and sample. We use different sampling techniques like simple random sampling, systematic sampling etc to generate a sample from the population. After sampling, statistical methods are used to make inferences and generalizations about the population considered for the experiment.



**Figure 1.1.1** (Relationship between population and sample)

**DEFINITION 1.1.5 (Parameter)**

It is a numerical characteristics of a population that summarize the data for the entire population.

## SECTION 1.1 BASIC CONCEPTS AND DEFINITIONS

### DEFINITION 1.1.6 (Statistic)

Statistic is a function of a sample (it is a numerical characteristic for this sample).

Parameters are usually unknown and sample statistics are used to estimate them. In other words, one use statistic to make inference about a parameter.

► **EXAMPLE 1.1.4** Consider a problem of finding out the proportion  $p$  of people aged 20-30 years old having height less than 5 feet in a population. ◀

In this problem the actual proportion  $p$  of the people aged 20-30 years old having height less than 5 feet is the parameter and the proportion  $\hat{p}$  calculated from a sample of the people aged 20-30 years old having height less than 5 feet is the statistic using which can make inference about  $p$ .

### DEFINITION 1.1.7 (Variables)

A variable is a characteristic, feature or factor that varies from one individual to another in a population.

Some of the characteristics of the individual may be similar to another individual but not all the individuals will have exactly same characteristics.

► **Example 1.1.5** Variables for countries are population, gross domestic product, sex ratio, birth and death rates, area, literacy rates. Different countries will have different values for each of these variables. Variables for humans are height, weight, sex, marital status, and eye color. In this example it can be seen that there might be some humans who will have same values for different variables say sex, marital status and eye color but not all humans have same values for these variables. ◀

### TYPES OF VARIABLES

Data are usually disaggregated according to their generation of variables in two main types: qualitative data and quantitative data.

### DEFINITION 1.1.8 (Qualitative (Categorical) Variables)

These variables can only take values which are non-numerical.

For example, marital status, eye-color, gender, hair-color, country, city of a human. Such variables cannot be ordered and no mathematical operations (addition, subtraction, multiplication or division) can be performed on them. These are measured on a nominal scale.

**DEFINITION 1.1.9 (Quantitative Variables)**

These are the variables which take numerical values.

For example, population, height, weight, temperature, revenue etc. Such variables can be ordered in increasing or decreasing and can undergo mathematical operations.

It can be seen in various cases that some quantitative variables only take values that are a whole number for example the numbers of accidents in a particular city, the numbers of laptops sold in a day, the numbers of goals scored by a football player, the numbers of children in a society, such variables are called as discrete variables. Discrete variables can assume only a finite number of variables. On the other hand, variables such as weight of person, distance between the cities, temperature, income, sex ratio, can be measured accurately and are called as continuous variables. There are no finite numbers of values which a continuous variable can take.

## Section 1.2

# ORGANIZATION AND GRAPHICAL REPRESENTATION OF THE DATA

In many problems, statisticians are provided with information in a format which is not well organized. Such set of information is known as raw data.

► **EXAMPLE 1.2.1 (For Qualitative Data)** Consider the month of the birth of 25 members of a community:

June	July	January	December	March
March	April	September	August	May
May	February	July	February	June
June	April	February	November	August
January	July	April	June	December

► **EXAMPLE 1.2.2 (For Discrete Quantitative Data)** Consider the number of children in 40 families of a society

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

► **EXAMPLE 1.2.3 (For Continuous Quantitative Data)** Consider the height of 25 adult men (in centimeters)

170	180	175	176	172
173	183	171	169	174
180	190	186	189	192
167	175	170	178	191
165	177	183	181	179



The data presented in Examples 1.2.1, 1.2.2, and 1.2.3 are called ungrouped data. An ungrouped data contains information on each member of a sample. It is always useful to organize raw data in proper format as it helps to approximately analyze the data in hand. There are various useful ways to illustrate the data in graphs and tables. To begin understanding different ways of representing data one should understand the concepts of frequency, relative frequency, cumulative frequency and percent frequency.

The number of observations of a particular class or category in the data is called as the frequency of that class. In addition to the frequency, we have relative frequency and percent frequency of a class which are defined as

$$\text{Relative Frequency of a class} = \frac{\text{Frequency of that class}}{\text{Total number of observations}}$$

$$\text{Percent Frequency of a class} = (\text{Relative Frequency}) \times 100\%$$

### FREQUENCY TABLE (FOR QUALITATIVE DATA)

For qualitative data, in the frequency table, all the classes of the variables are mentioned along with their frequency, relative frequency and percent frequency. It is a good practice to mention the total number of observation in the last row of the table.

► **EXAMPLE 1.2.4** Construct the frequency table for the data in Example 1.2.1.

**The Answer:** The frequency table for the given data in Example 1.2.1 is as follows:

**Table 1.2.1** (Frequency table for birth months of 25 persons)

Month	Frequency	Relative Frequency	Percent Frequency
January	2	$2/25 = 0.08$	$0.08 \times 100 = 8\%$
February	3	$3/25 = 0.12$	$0.12 \times 100 = 12\%$
March	2	$2/25 = 0.08$	$0.08 \times 100 = 8\%$
April	3	$3/25 = 0.12$	$0.12 \times 100 = 12\%$
May	2	$2/25 = 0.08$	$0.08 \times 100 = 8\%$
June	4	$4/25 = 0.18$	$0.18 \times 100 = 18\%$
July	3	$3/25 = 0.12$	$0.12 \times 100 = 12\%$

SECTION 1.2 ORGANIZATION AND GRAPHICAL REPRESENTATION OF THE DATA

Month	Frequency	Relative Frequency	Percent Frequency
August	2	$2/25 = 0.08$	$0.08 \times 100 = 8\%$
September	1	$1/25 = 0.04$	$0.04 \times 100 = 4\%$
October	0	$0/25 = 0$	$0 \times 100 = 0\%$
November	1	$1/25 = 0.04$	$0.04 \times 100 = 4\%$
December	2	$2/25 = 0.08$	$0.08 \times 100 = 8\%$
<b>Total</b>	$n = \sum_i f_i = 25$	<b>1</b>	<b>100%</b>

Using the  $\sum_i f_i$  notation, we can denote the sum of frequencies of all classes by  $\sum_i f_i$ . Hence

$$\begin{aligned} \sum_i f_i &= f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + f_7 + f_8 + f_9 + f_{10} + f_{11} + f_{12} \\ &= 2 + 3 + 2 + 3 + 2 + 4 + 3 + 2 + 1 + 0 + 1 + 2 = 25 \end{aligned}$$

► **EXAMPLE 1.2.5** Constructing the frequency table for Example 1.2.2.

**The Answer:** The frequency table for the given data in Example 1.2.2 is as follows:

**Table 1.2.2** (Frequency table for number of children in 40 families)

No. of Children	Frequency	Relative Frequency	Percent Frequency
0	12	0.300	30%
1	9	0.225	22.5%
2	9	0.225	22.5%
3	4	0.100	10%
4	5	0.125	12.5%
5	1	0.025	2.5%
<b>Total</b>	<b>40</b>	<b>1</b>	<b>100%</b>

► **EXAMPLE 1.2.6** Consider the blood groups of the 40 persons below.

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

Construct the frequency table for the above data.

**The Answer:** The frequency table for the above example is given as follows:

**Table 1.2.3** (Frequency Table for blood group of 40 persons)

Blood group	Frequency	Relative Frequency	Percent Frequency
O	16	0.40	40 %
A	18	0.45	45 %
B	4	0.10	10 %
AB	2	0.05	5 %
<b>Total</b>	<b>40</b>	<b>1</b>	<b>100%</b>

**FREQUENCY TABLE (FOR DISCRETE QUANTITATIVE DATA)**

For discrete quantitative data, in the frequency table, all the values of the variables are mentioned along with their frequency, relative frequency and percent frequency. This one is quite similar to the frequency table for qualitative data.

► **EXAMPLE 1.2.7** Consider a sample of 40 students. We want to see in how many subjects each student failed in the 5th standard.

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

Then we find the frequency table for this data as follow.

**Table 1.2.4** (Frequency table for 40 students who failed)

No. of subjects in which student failed	Frequency	Relative Frequency	Percent Frequency
0	8	0.20	20 %
1	18	0.45	45 %
2	12	0.30	30 %
3	2	0.05	5 %
<b>Total</b>	<b>40</b>	<b>1</b>	<b>100%</b>

**FREQUENCY DISTRIBUTION TABLE (FOR CONTINUOUS QUANTITATIVE DATA)**

A continuous quantitative variable can take indefinite number of values and hence it is impossible to create a frequency distribution table which gives frequency of each possible value of the variable. We therefore define sets of classes of that variable to calculate the frequency.

## SECTION 1.2 ORGANIZATION AND GRAPHICAL REPRESENTATION OF THE DATA

For continuous quantitative data, frequency distribution table with  $k$  classes is constructed in the following manner:

- a. Calculate the range (we denote it by  $R$ ) of the data which is given by the difference of the greatest  $x_\ell$  and the smallest  $x_s$  value of the data. This means:

$$R = x_\ell - x_s$$

- b. The number of classes (or categories) to be formed. Generally, 5-20 categories are good for the analysis. The length (or height) of each class (we denote it by  $C$ ) is determined by the following relation:

$$\text{Length} = \frac{\text{Range}}{\text{No. of classes}} \quad \Leftrightarrow \quad C = \frac{R}{k}$$

- c. If the length of a class as calculated, using the above formula comes out to be a fraction value  $t$ . In this case we can take a value  $u$  greater than  $t$ , or take (if the length of the class is greater than 1) the smallest integer greater than the fraction as the class length.
- d. The lower limit of the first class limit is the minimum value in the data. The upper limit of the first class limit calculated by adding the number  $(C - 1)$  to the lower limit of the current class. The lower limit of the next class limit calculated by adding 1 to the upper limit of the previous class limit, and we become the upper limit of this class by adding the number  $(C - 1)$  to the lower limit of this class. As such, the rest of the class limits are built.
- e. To make the class boundaries subtract 0.5 unit from the lower limit of each class limit and add 0.5 unit to the upper limit of each class limit. This means that the length class  $C$ , which is previously calculated is for the class boundaries.
- f. It is common practice to represent the classes in frequency table of a continuous quantitative variable by the class midpoint. Class midpoint is the center value of any class of the variable and it is defined in the following manner:

$$\text{Class Midpoint} = \frac{\text{Upper limit} + \text{Lower limit}}{2}$$

- g. Determine the value of the ascending cumulative frequency (ACF) of a class. This value gives us the value of how many items are less than the upper limit of the class boundary corresponding to that class.

**REMARKS 1.2.1**

1. When raw data are dumped in a frequency distribution table, they are emptied in the class boundaries. We note here that it is unwanted to dump the data in frequency distribution table if the number data is less than 32.
2. If a value of data equals to the upper limit of the class boundaries, they are placed in the following class boundaries. Therefore, we will use the symbol ( $\rightarrow$ ) rather than ( $-$ ) in the class boundaries.
3. If we do not know the number of classes to use, then we can calculate the number of classes ( $k$ ) by the following relation:

$$k = \left\lfloor 3.322 \log n \right\rfloor$$

Where  $\lfloor x \rfloor$  is the greatest integer number equal or less than  $x$ . For example  $\lfloor 5.76 \rfloor = 5$ .

► **EXAMPLE 1.2.8** In the shopping center recorded sales of traditional accessories for girls, whose prices are between 1 and 25 SR, we had the following data estimated at SR.

4	1	7	9	12	16	17	7	12	19
22	24	3	2	8	6	13	24	14	11
18	16	23	20	1	2	6	25	15	7
11	12	16	17	21	22	15	17	14	5
7	8	12	13	20	23	13	19	18	12

We will construct the frequency distribution table for this data by using 5 classes. Where we have:

First, we calculate the range of the given data. Note that we have in that example, the greatest value is 25 and the smallest value is 1. Therefore, we have:

$$R = 25 - 1 = 24$$

The length of class is given by

$$C = 24/5 = 4.8$$

We will take  $C = 5$ , Therefore, the length of class limit equals to

$$C - 1 = 5 - 1 = 4.$$

Table 1.2.5

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Less than	Ascending Cumulative Frequency (ACF)
1 – 5	0.5 → 5.5	3	7	0.14	5.5	7
6 – 10	5.5 → 10.5	8	9	0.18	10.5	7+9 = 16
11 – 15	10.5 → 15.5	13	14	0.28	15.5	7+9+14 = 30
16 – 20	15.5 → 20.5	18	12	0.24	20.5	7+9+14+12 = 42
21 – 25	20.5 → 25.5	21	8	0.16	25.5	7+9+14+12+8 = 50
Total	-----	-----	50	1	-----	-----

► **EXAMPLE 1.2.9** Consider the mileage of 40 cars per liter of fuel in a particular city given below:

12	16	15	12	19	17	18	16	14	13
12	20	12	15	16	20	16	15	12	18
16	17	19	15	16	17	15	16	15	14
12	13	14	15	16	17	18	19	20	20

We construct the frequency table for this data in the following manner:

We have the range for the given data equal to  $R = 20 - 12 = 8$ .

Now we determine the number of classes using the following relation:

$$k = \lceil 3.322 \log n \rceil = \lceil 3.322 \log 40 \rceil = \lceil 5.322 \rceil = 5$$

So the class length equal to  $C = 8/5 = 1.6$ .

We will take  $C = 2$ , Therefore, the length of class limit equal to  $C - 1 = 2 - 1 = 1$ .

Table 1.2.6

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency
12 – 13	11.5→13.5	12.5	8	0.20	8
14 – 15	13.5→15.5	14.5	10	0.25	18
16 – 17	15.5→17.5	16.5	12	0.30	30
18 – 19	17.5→19.5	18.5	6	0.15	36
20 – 21	19.5→21.5	20.5	4	0.10	40
Total	-----	-----	40	1	-----

**CUMULATIVE RELATIVE AND CUMULATIVE PERCENTAGES FREQUENCIES**

The cumulative relative frequencies are obtained by dividing the cumulative frequencies by the total number of observations in the data. The cumulative percentages frequencies are obtained by multiplying the cumulative relative frequencies by 100.

► **EXAMPLE 1.2.10** We will construct the cumulative relative and the cumulative percentages frequencies by using the data presented in the previous Example 1.2.9.

We have:

**Table 1.2.7**

Class Boundaries	Frequency	Ascending Cumulative Frequency	Ascending Cumulative Relative Frequencies	Ascending Cumulative Percentages Frequencies
13.5→15.5	8	8	$8/40 = 0.20$	$0.20 \times 100 = 20$
15.5→17.5	10	18	$18/40 = 0.45$	$0.45 \times 100 = 45$
17.5→19.5	12	30	$30/40 = 0.75$	$0.75 \times 100 = 75$
19.5→21.5	6	36	$36/40 = 0.90$	$0.90 \times 100 = 90$
13.5→15.5	4	40	$40/40 = 1$	$1.00 \times 100 = 100$
Total	40	-----	-----	-----

## Section 1.3

# GRAPHICAL REPRESENTATION

### DEFINITION 1.3.1 (Pie Charts)

A pie chart is a simple way of representing the proportion of each class or category of data on a circular disk, so that each category is allocated a circular sector representing it.

Pie chart is a disk which is divided into the same number of pies in which there are classes. Each pie represents a class and its width is in accordance with its relative frequency. Pie charts are useful for nominal or ordinal categories.

For graph a pie chart for data we calculate the measure of the central angle for the class ( $i$ ) by the following relation:

$$(\text{Relative Frequency of class } (i)) (360) = \dots \text{ degree}$$

► **EXAMPLE 1.3.1** The pie chart for Example 1.2.6 is given as follows:

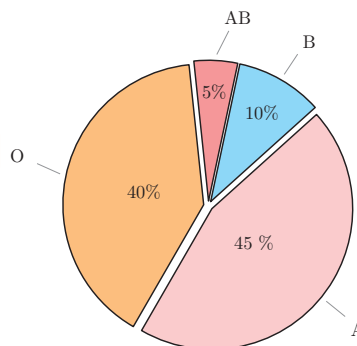


Figure 1.3.1 (Pie Chart for Example 1.2.6)

### DEFINITION 1.3.2 (Bar Charts)

In a bar chart, the frequency of each class is represented by a bar. The height of the bar corresponds to the frequency of the class. The width of the bar doesn't matter.



These are a simple representation of classes and their frequencies and provide a simple way to compare the frequencies of different classes. The bars can be represented vertical and horizontal manner.

The below is the bar chart representation for the data in Example 1.2.6:

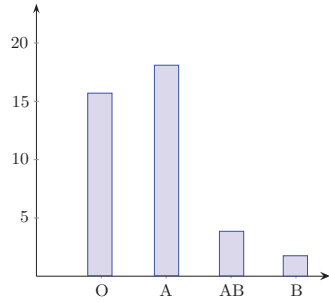


Figure 1.3.2-a (Vertical Bar Chart)

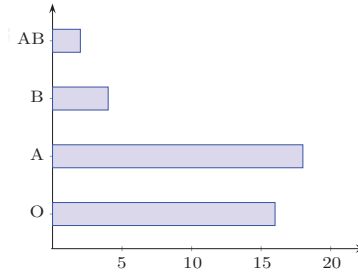


Figure 1.3.2-b (Horizontal Bar Chart)

**DEFINITION 1.3.3 (Two Directional Bar Charts)**

Such bar charts can represent both positive and negative values of different classes.

► **EXAMPLE 1.3.2** Consider the following changes in income of a company from January to June.

Table 1.3.1

Month	Change in Income
January	-4 %
February	14 %
March	6 %
April	-10 %
May	-4 %
June	5 %

The following is the two-way bar chart for the above data:

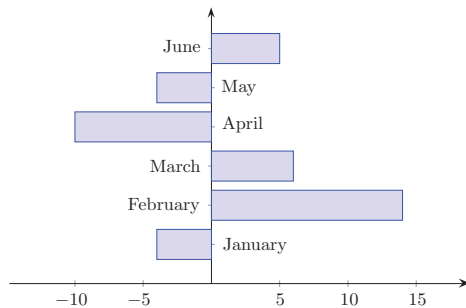


Figure 1.3.3 (Two Directional Bar Chart)

## SECTION 1.3 GRAPHICAL REPRESENTATION

### DEFINITION 1.3.4 (Multiple Bar Charts)

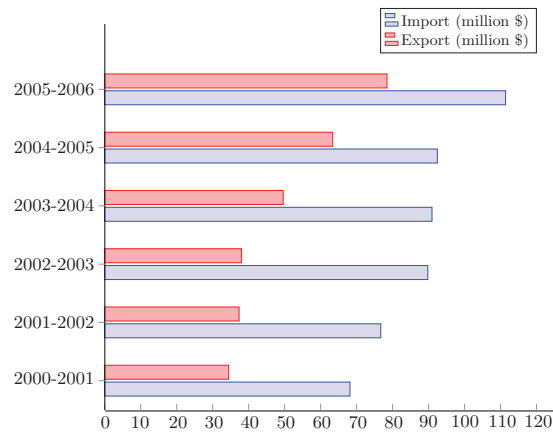
Such bar charts are used to represent multiple inter related variables by clustering bars side by side.

► **EXAMPLE 1.3.3:** To demonstrate a multiple bar chart, consider the following import export data.

**Table 1.3.2**

Year	Imports \$ (in billions)	Exports \$ (in billions)
2000→2001	68.15	34.44
2001→2002	76.71	37.33
2002→2003	89.78	37.98
2003→2004	90.95	49.59
2004→2005	92.43	63.35
2005→2006	111.39	78.44

The following is the multiple bar chart for the above data:



**Figure 1.3.4** (Horizontal Bar Chart)

In multiple bar charts the interrelated variables are represented by bars of different colors to make the comparisons easier.

### DEFINITION 1.3.5 (Component (Stacked) Bar Chart)

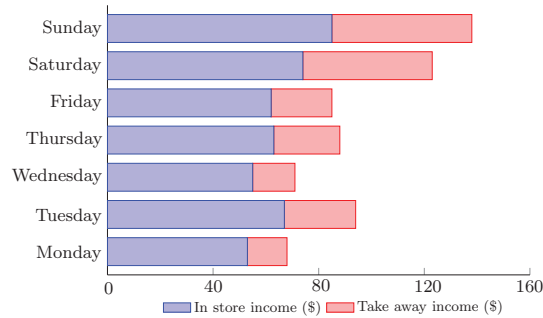
Such bar charts when each class or category has components, which make up the class. Each of the components is represented by a section in the bar whose size is proportional to its contribution in the class.

► **EXAMPLE 1.3.4** Let us consider the income at a café in a particular week.

**Table 1.3.3**

Day	In store Income (in \$)	Take away Income (in \$)	Total income (in \$)
Monday	53	15	68
Tuesday	67	27	94
Wednesday	55	16	71
Thursday	63	25	88
Friday	62	23	85
Saturday	74	49	123
Sunday	85	53	138

The following is the component (Stacked) bar chart for the above data:



**Figure 1.3.5** (Horizontal Bar Chart)

**DEFINITION 1.3.6 (Histogram)**

Histogram is similar to bar chart but they both have a basic difference that is in histograms, classes of the variable are adjacent to each other and the rectangular bars must touch each other. Histograms are generally used to represent quantitative data. The class intervals in a histogram are called as bins. To study the properties of the data, statisticians usually vary the bin size to make inferences about the distribution of variable.

► **EXAMPLE 1.3.5** Consider the data from Example 1.2.9, where we have:

Class Boundaries	11.5 – 13.5	13.5 – 15.5	15.5 – 17.5	17.5 – 19.5	19.5 – 21.5	Total
Frequency	8	10	12	6	4	40

Then the histogram for the given data is as follow:

## SECTION 1.3 GRAPHICAL REPRESENTATION

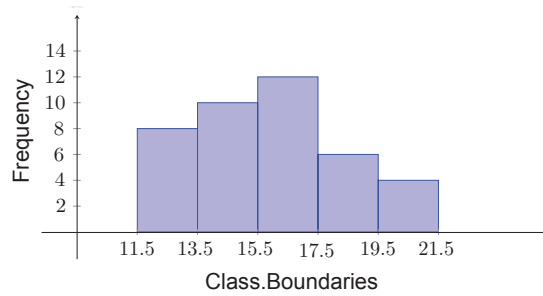


Figure 1.3.6

### REMARK 1.3.1

Histogram, on the basis of its shape, is classified into the following types:

#### I- Symmetric Histogram:

In such types of histogram, the bins in the middle have higher frequency and the frequency values keeps decreasing as one move towards the boundary from both left and right hand sides. Some symmetric histograms have an appearance that of a bell (See the following figures). The tails on both left hand and right hand side are equivalent.

On the basis of number of peaks, symmetric histogram can be classified into:

**Unimodal:** Histogram with one peak.

**Bimodal:** Histogram with two peaks.

**Multimodal:** Histogram with more than two peaks.

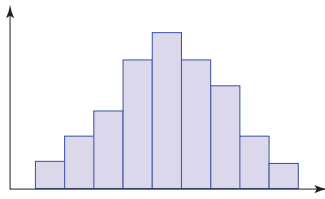


Figure 1.3.7-a (Unimodal Histogram)

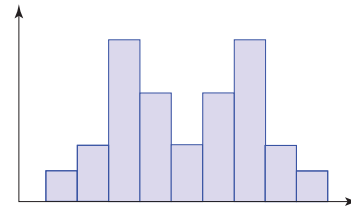


Figure 1.3.7-b (Bimodal Histogram)

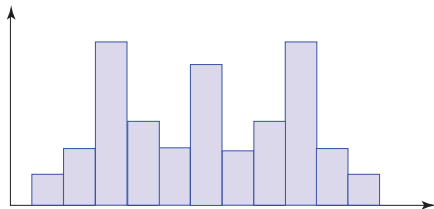


Figure 1.3.7-c (Multimodal Histogram)

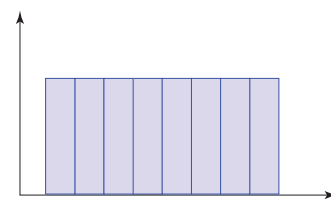


Figure 1.3.7-d (Uniform Histogram)

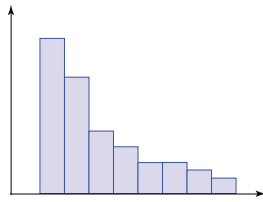
#### II- Symmetric Skewed Histogram

Before talking about twisted distributions, give the following definition.

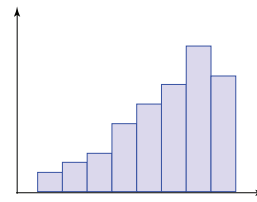
**DEFINITION 1.3.7 (Skewedness)**

Histograms are called as skewed if they are non-symmetric. In such histograms, bins on one side have very high frequency which decreases as we move to the other side. The side with lower frequency is said to have a longer tail.

A right skewed histogram is a histogram, which has longer tail on the right side and left skewed histogram is one, which has longer tail on the left side.



**Figure 1.3.8-a:** (Right Skewed Histogram)



**Figure 1.3.8-b:** (Left Skewed Histogram)

**DEFINITION 1.3.8 (Polygon)**

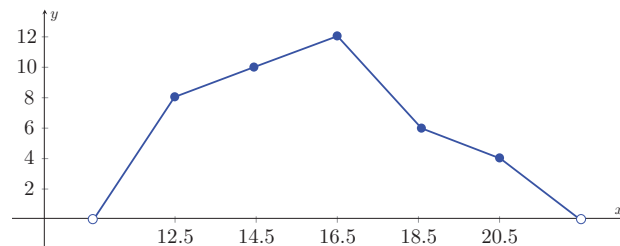
The frequency polygon is a polygon which connects with a straight line the points  $(x_i, f_i)$ , whereas  $x_i$  and  $f_i$  is the midpoint and the frequency of class boundary  $i$  respectively, and closes from the left to the center of the previous class, and from right to center of the subsequent class after last class.

Polygons are useful to compare two or more distributions of the variables in the same graph.

► **EXAMPLE 1.3.6** Consider the frequency table of car mileage data from Example 1.2.9.

Class Boundaries	11.5 – 13.5	13.5 – 15.5	15.5 – 17.5	17.5 – 19.5	19.5 – 21.5	Total
Frequency	8	10	12	6	4	40

As we can see the class midpoints are 12.5, 14.5, 16.5, 18.5 and 20.5 respectively. We now construct the frequency polygon for this data



**Figure 1.3.9:** (Polygon frequency chart for mileage data)

## SECTION 1.3 GRAPHICAL REPRESENTATION

### DEFINITION 1.3.9 (Ogive)

The cumulative frequency polygon is a polygon which connects with a straight line the points  $(b_i, F_i)$ , whereas  $b_i$  and  $F_i$  is the upper bound and the cumulative frequency of class boundary  $i$  respectively, and closes from the left to the beginning of first class.

For a less than cumulative frequency, the ogive is known as a less than ogive and for a greater than cumulative frequency the ogive is known as a greater than ogive.

The less than ogive curve for the data from Example 1.2.9 is presented in the following graph.

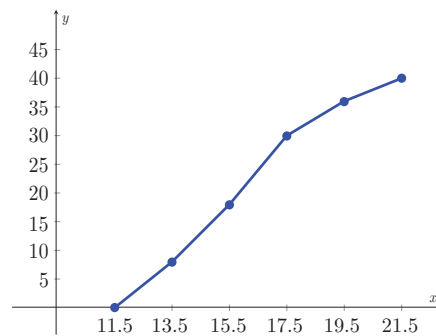


Figure 1.3.10: (Less than ogive curve for mileage data)

## Section 1.4

# MEASURES OF CENTRAL TENDENCY

In the above methods, we became familiar with the concepts that help us to organize raw data in tables and graphs. We can understand important features of the complete data by looking at some characteristics of the data only instead of considering the complete data. Most set of data seem to possess central values which characterize the data in hand. Such phenomenon is called existence of central tendency. We will now understand the concepts of mean, median, mode and quartiles of the data. Some of these are only defined for quantitative data but some are defined for both.

### DEFINITION 1.4.1 (Mean)

Let  $x_1, x_2, \dots, x_n$  be numerical data, then one defines the mean (we denote it by  $\bar{x}$ ) is following relation:

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

So we note that the mean of a data set is defined as the arithmetic average of variable values i.e. sum of all the values divided by the number of values. Therefore, mean is only defined for quantitative data.

► **EXAMPLE 1.4.1 (For Raw Data):** Calculate the mean of the following:

20, 18, 15, 15, 14, 12, 11, 9, 7, 6, 4, 1

According to the definition of the mean we have:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \frac{20 + 18 + 15 + 15 + 12 + 14 + 11 + 9 + 7 + 6 + 4 + 1}{12} = \frac{132}{12} = 11 \end{aligned}$$

Hence the mean of the above data is 11.

We consider data in the following Frequency distribution table:

Table 1.4.1

N.o.C	Class Boundaries	Class Midpoint	Frequency	Ascending Cumulative Frequency (ACF)
1	$b_0 \rightarrow b_1$	$x_1$	$f_1$	$F_1 = f_1$
2	$b_1 \rightarrow b_2$	$x_2$	$f_2$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots \vdots \vdots$	$\vdots$	$\vdots$	$\vdots \vdots \vdots \vdots$
$k-1$	$b_{k-2} \rightarrow b_{k-1}$	$x_{k-1}$	$f_{k-1}$	$F_{k-1} = f_1 + f_2 + \dots + f_{k-1}$
$k$	$b_{k-1} \rightarrow b_k$	$x_k$	$f_k$	$F_k = f_1 + f_2 + \dots + f_k$
<b>Total</b>	-----	-----	$\sum f_i$	-----

Then we can calculate the mean for these data by the following relation:

$$\bar{x} = \frac{1}{\sum f_i} \sum_{i=1}^n f_i x_i$$

Where, for discrete quantitative data,  $f_i$  represents the frequency of the variable  $x_i$  and for continuous quantitative data,  $f_i$  represents the frequency of the class of the variable and  $x_i$  is the midpoint of that class. The examples below will help the reader in better understanding.

► **EXAMPLE 1.4.2 (Discrete quantitative data)** Let us consider the following data:

No. of subjects in which student failed	Frequency
0	8
1	18
2	12
3	2
<b>Total</b>	<b>40</b>

The mean of the number of subjects in which students failed is calculated as:

$$\bar{x} = \frac{0 \times 8 + 1 \times 18 + 2 \times 12 + 3 \times 2}{8 + 18 + 12 + 2} = \frac{0 + 18 + 24 + 6}{40} = \frac{48}{40} = 1.2$$

As it can be seen that the mean is 1.2 which is close to 1 so one can make a conclusion that most of the students from the sample failed in 1 which is true as 1 has the highest frequency among all of them.

► **EXAMPLE 1.4.3 (Continuous quantitative data)** Let us consider the following data:



Table 1.4.2

Class Boundaries	Class Midpoint	Frequency
11.5→13.5	12.5	8
13.5→15.5	14.5	10
15.5→17.5	16.5	12
17.5→19.5	18.5	6
19.5→21.5	20.5	4
<b>Total</b>	-----	<b>40</b>

The mean of the mileage of the cars is calculated as:

$$\bar{x} = \frac{12.5 \times 8 + 14.5 \times 10 + 16.5 \times 12 + 18.5 \times 6 + 20.5 \times 4}{40} = \frac{636}{40} = 15.9$$

### Advantages of the Mean

- It is quick and easy to compute.
- All the values are considered for calculating the mean of a variable.
- It is one and only one value for a set of data.

### Disadvantages of the Mean

- Mean is not defined for qualitative data.
- Since it considers all the observed values, it is highly affected by the extreme values.
- It becomes not applicable if a data is lost.

#### DEFINITION 1.4.2 (Weighted Mean)

In a weighted mean (we denote it by  $\bar{x}_w$ ), some of the observations are given more weight than the others. This is used when some values are more significant than others.

Let the observed values be  $x_1, x_2, \dots, x_n$  and the weights corresponding to each value are  $w_1, w_2, \dots, w_n$  then the weighted mean is calculated as:

$$\bar{x}_w = \frac{1}{\sum w_i} \sum_{i=1}^n w_i x_i$$

When all the weights are equal then the weighted mean is equal to the ordinary mean.

► **EXAMPLE 1.4.4** A person wants to decide which car is better for him to purchase using the following rating system:

look 20%, mileage 30%, Engine 50%

Car A gets 8 (out of 10) in engine, 6 for mileage and 7 for looks.

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

Car B gets 9 for engine, 4 for mileage and 6 for looks.

Car ratings are calculated by calculating the weighted mean.

$$\bar{x}_A = \frac{0.5 \times 8 + 0.3 \times 6 + 0.2 \times 7}{(0.5 + 0.3 + 0.2)} = 7.2$$

$$\bar{x}_B = \frac{0.5 \times 9 + 0.3 \times 4 + 0.2 \times 6}{(0.5 + 0.3 + 0.2)} = 6.9$$

Hence Car A has higher rating and hence is better for him than Car B. ◀

One can use weighted mean to compare between the objects in hand or to make decisions using this measure.

### DEFINITION 1.4.3 (Median)

Median (we denote it by  $\tilde{x}$ ) is that value which divides the data in half after ordering them, in ascending or descending order such that one-half of the data is less than or equal to the median and the other half is greater than or equal to the median. The following graph explains the concept of median.

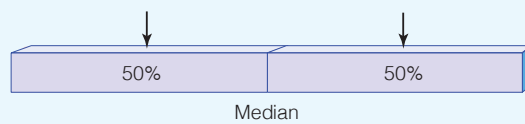


Figure 1.4.1: The concept of median

### MEDIAN FOR RAW DATA

To calculate the median we arrange the data in the increasing order and suppose it  $x_1, x_2, \dots, x_n$ .

Then the median given by the following relation:

$$\tilde{x} := \begin{cases} \frac{x_{\frac{n+1}{2}}}{2} & \text{for } n \text{ odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{for } n \text{ even} \end{cases}$$

► **EXAMPLE 1.4.5** Calculate the median of the finishing times of 7 bike racers who had finishing times as

$$28, 22, 26, 29, 21, 23, 24.$$

We first arrange them in increasing order

$$\begin{array}{ccccccc} 21 & 22 & 23 & 24 & 26 & 28 & 29 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{array}$$

Since there are 7 (odd number) observations, the median is given by:

$$\tilde{x} = x_{\frac{7+1}{2}} = x_4 = 24$$

Hence the median is 24. We can see that 3 observations are less than 24 and 3 observations are greater than 24.

► **EXAMPLE 1.4.6** Calculate the median of the finishing times of 8 bike racers who had finishing times as:

$$28, 22, 26, 29, 21, 23, 24, 35$$

In the similar manner as the above example, we first arrange the data as:

$$21, 22, 23, 24, 26, 28, 29, 35$$

$$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad x_8$$

Since there are 8 (even number) observations, the median is given by:

$$\tilde{x} = \frac{x_{\frac{8}{2}} + x_{\frac{8}{2}+1}}{2} = \frac{x_4 + x_5}{2} = \frac{24 + 26}{2} = 25$$

The median is 25. Again, we can observe that there are 4 observations are less than 25 and 4 observations are greater than 25.

### MEDIAN FOR FREQUENCY TABLE

For discrete quantitative data with a frequency table, the median is calculated by the following steps:

- First, evaluate the cumulative frequency of the data.
- If the total number of observations is  $n$  then we consider the smallest cumulative frequency greater than  $\frac{n}{2}$ .
- The value of the data corresponding to that cumulative frequency is the required median

► **EXAMPLE 1.4.7** Consider the following data:

**Table 1.4.3**

No. of subjects in which student failed	Frequency	ACF
0	8	8
1	18	8+18 = 26
2	12	26+12 = 38
3	2	38 +2 = 40
Total	40	-----

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

For this example,  $n = 40$ . Therefore,  $\frac{n}{2} = 20$ . The smallest cumulative frequency greater than 20 is 26. The value of the data corresponding to that cumulative frequency is 1. Hence, the median is 1.

### MEDIAN FOR FREQUENCY DISTRIBUTION TABLE

We consider data in a frequency distribution table as in Table 1.4.1. Then for continuous quantitative data, the median is calculated by the following steps:

- First find the class whose cumulative frequency is the smallest cumulative frequency among those which are greater than  $\frac{1}{2} \sum f_i$ . Similar to the steps to find median of discrete quantitative data. Such a class is called as median class.
- We use the following formula to calculate the median for continuous quantitative data

$$\tilde{x} := \tilde{L} + \frac{\frac{1}{2} \sum f_i - (\tilde{F} - \tilde{f})}{\tilde{f}} \times C$$

Where,  $\tilde{L}$  is the lower limit of the median class,

$\tilde{F}$  is the cumulative frequency of the median class,

$\tilde{f}$  is the frequency of the median class,

$C$  is the class length of the median class.

Note that the median class is the first class whose cumulative frequency is greater than or equal to half the sum of the frequencies.

The examples below will illustrate the above steps for better understanding of the reader.

► **EXAMPLE 1.4.8** We will calculate the median for the data in Example 1.4.3.

**Table 1.4.4**

Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency
11.5→13.5	12.5	8	0.20	8
13.5→15.5	14.5	10	0.25	18
15.5→17.5	16.5	12	0.30	30
17.5→19.5	18.5	6	0.15	36
19.5→21.5	20.5	4	0.10	40
<b>Total</b>	-----	<b>40</b>	<b>1</b>	-----

First, we find the median class. In this example  $n = 40$ , therefore,  $\frac{1}{2} \sum f_i = 20$ . Therefore the median class is  $15.5 - 17.5$ . So using the formula:

$$\tilde{x} := \tilde{L} + \frac{\frac{1}{2} \sum f_i - (\tilde{F} - \tilde{f})}{\tilde{f}} \times C$$

We have  $\tilde{L} = 15.5$ ,  $\tilde{F} = 30$ ,  $\tilde{f} = 12$  and  $C = 2$  then.

$$\tilde{x} = 15.5 + \frac{20 - (30 - 12)}{12} \times 2 = 15.5 + 0.33 = 15.83$$

#### DEFINITION 1.4.4 (The Mode)

The mode (we denote it by  $\hat{x}$ ) of a set of raw data is the value, which has occurred maximum number of times i.e. has the highest frequency.

#### REMARKS 1.4.1

- One can use the mode for qualitative and quantitative data.
- In case if the highest frequency is constant for all data, then the data is said to have no mode.
- There is a possibility that multiple values have highest frequency in such situations the variable has more than one mode and is said to be multimodal.

#### MODE FOR RAW DATA

To find the mode of raw data we will see the following examples.

#### ► EXAMPLE 1.4.9:

The finishing times of 7 bike racers who had finishing times as:

28, 22, 26, 29, 21, 23, 24.

Table 1.4.5

Finishing Time	21	22	23	24	26	28	29
Frequency	1	1	1	1	1	1	1

In this example since the highest frequency is 1. There is no mode.

Whereas if the below data is considered for 10 bike racers

28, 22, 26, 29, 21, 23, 28, 28, 25, 29.

Table 1.4.6

Finishing Time	21	22	23	25	26	28	29
Frequency	1	1	1	1	1	3	2

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

For this data, the mode is equal to 28.

► **EXAMPLE 1.4.10** We consider the data in Example 1.2.6 (the data about the blood group of 40 persons), where we had:

<b>Blood group</b>	O	A	B	AB
<b>Frequency</b>	16	18	4	2

So we find that the mode of the blood group is A.

► **EXAMPLE 1.4.11** Consider the following data representing the age (in years) of 15 students:

<b>12</b>	<b>11</b>	<b>13</b>	<b>14</b>	<b>13</b>
<b>12</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>12</b>
<b>12</b>	<b>13</b>	<b>14</b>	<b>11</b>	<b>13</b>

The ages 12 and 13 in this example have the highest frequency. Hence the variable is said to be bimodal i.e. having two modes 12 and 13.

► **EXAMPLE 1.4.12:**

The data below represents the outcome when a fair dice is rolled 25 times.

<b>1</b>	<b>5</b>	<b>2</b>	<b>4</b>	<b>6</b>
<b>3</b>	<b>1</b>	<b>5</b>	<b>2</b>	<b>4</b>
<b>1</b>	<b>6</b>	<b>2</b>	<b>4</b>	<b>2</b>
<b>2</b>	<b>1</b>	<b>5</b>	<b>2</b>	<b>6</b>
<b>2</b>	<b>4</b>	<b>2</b>	<b>3</b>	<b>4</b>

The frequency table is given by

**Table 1.4.7**

<b>Outcome</b>	1	2	3	4	5	6
<b>Frequency</b>	4	8	2	5	3	3

The mode of the outcomes of dice for the above data is equal to 2

### MODE FOR FREQUENCY TABLE

To find the mode of frequency table we will see the following example.

► **EXAMPLE 1.4.13** Find the mode of the following marks (out of 10) obtained by 20 students:

4, 6, 5, 9, 3, 2, 7, 7, 6, 5, 4, 9, 10, 10, 3, 4, 7, 6, 9, 9

The frequency table is given by:

**Table 1.4.8**

Marks (out of 10)	Frequency
2	1
3	2
4	3
5	2
6	3
7	3
9	4
10	2

The mode of the marks obtained by the student is 9.

### MODE FOR FREQUENCY DISTRIBUTION TABLE

We consider data in a frequency distribution table as in **Table 1.4.1**. Then the mode for those data is calculated by the following relation:

$$\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} C$$

Where,  $\hat{L}$  is the lower limit of the mode class,

$d_1$  is the difference between the frequency of the class and the frequency of the previous class directly,

$d_2$  is the difference between the frequency of the class and the frequency of the next class directly,

$C$  is the class length of the mode class.

Note that the class of mode is the class whose frequency is greater than the frequency of the previous and subsequent classes directly, and that this class is not extremity. In other words, the first and last class in the distribution is not seen as classes modal.

#### ► EXAMPLE 1.4.14:

Consider the time taken by 21 persons in the run race.

**Table 1.4.9**

Seconds	Frequency
50.5 → 55.5	2
55.5 → 60.5	7
60.5 → 70.5	8
65.5 → 75.5	4

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

The above data has the modal class 60.5-65.5 as this class has the highest frequency. Therefore, we have:

$$\begin{aligned}\hat{x} &= \hat{L} + \frac{d_1}{d_1 + d_2} C \\ &= 60.5 + \frac{8 - 7}{(8 - 7) + (8 - 4)} \times 5 = 61.5\end{aligned}$$

### ► EXAMPLE 1.4.15:

The minutes spent per week by the teenagers in watching movies is given by.

Table 1.4.10

Number of Minutes per week	Number of Teenagers
0 → 99.5	26
99.5 → 199.5	32
199.5 → 299.5	65
299.5 → 399.5	75
399.5 → 499.5	60
499.5 → 599.5	42
<b>Total</b>	<b>300</b>

Then we find the class modal for the above data is 299.5-399.5 because it has the highest frequency 75. Therefore, we have:

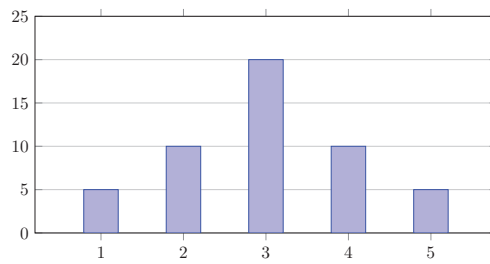
$$\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} C = 299.5 + \frac{10}{10 + 15} \times 100 = 339.5$$

### THE RELATIONSHIPS AMONG THE MEAN, MEDIAN AND MODE

Mean, median and mode are all the measures of central tendency but sometimes it becomes difficult to choose which measure is an appropriate representation of the data in hand. We will consider the cases when the data is symmetric and asymmetric (or skewed) and learn the relationship between these measures.

For a symmetric data, all the three measures i.e. mean, median and mode are equal and lie in the center of the data as can be seen from the below plot.



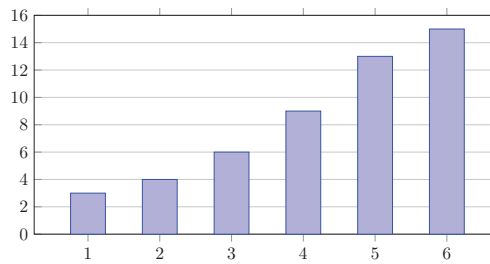


**Figure 1.4.2** (A symmetric distribution with mode = mean = median = 3)

For this data,

$$\text{Mode} = \text{Mean} = \text{Median} = 3.$$

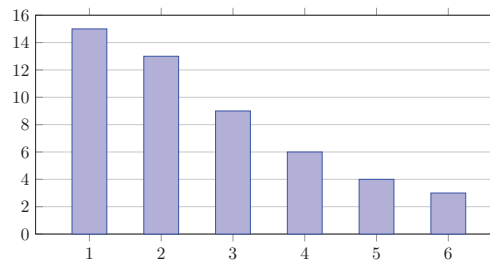
For a left skewed data, we have mode to be greatest followed by median which is followed by the mean. The presence of extreme values on the left side drag the mean towards left and hence mean is the smallest of all for left skewed data. The plot below shows the relationship between these measures.



**Figure 1.4.3** (A left skewed distribution with mode greater than median greater than mean)

For the above data, the mode is 6, the median is 5 and the mean is 4.4.

For a right skewed data, we have mean to be largest followed by median which is followed by the mode. Here also the extreme values on the right side drag the mean towards right and hence mean is the largest of all for right skewed data. The plot below shows the relationship between these measures.

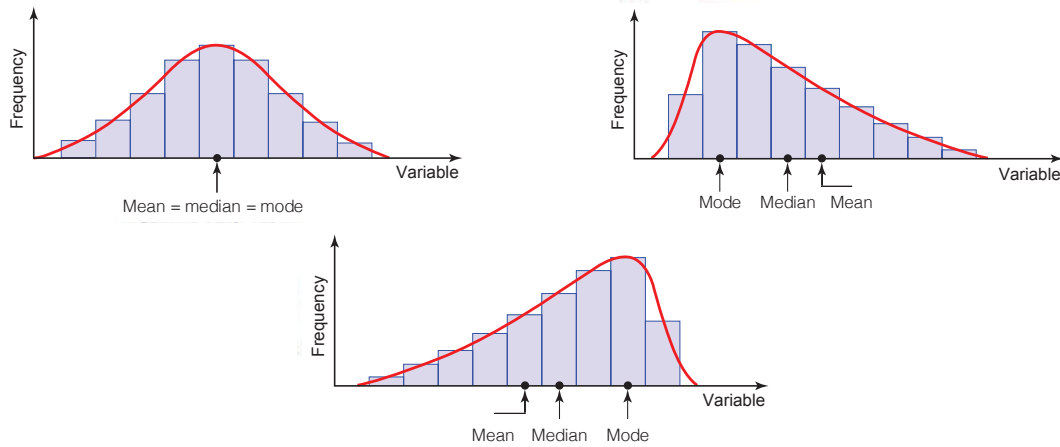


**Figure 1.4.4** (A right skewed distribution with mean greater than median greater than mode)

For the above data, the mean is 2.6, the median is 2 and the mode is 1.

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

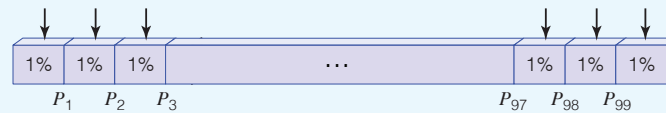
In general, we can explain the relationship among the three measures of tendency using the following graph.



**Figure 1.4.5** (The relationship among the three measures of tendency)

### DEFINITION 1.4.5 (Variance Percentiles)

The percentiles (we denote them by  $P_1, P_2, \dots$  and  $P_{99}$ ) of a variable divide the observed values into 100 equal parts. Median is 50<sup>th</sup> percentile and it divides data into two equal halves. The percentile  $P_1$  divides the observed data into 1% from bottom and 99% from top. Similarly any  $j^{\text{th}}$  percentile,  $P_j$  divides the observed value into two parts such that  $j$  % observed values are below this value and  $(100 - j)$  % observed values are above this value. The following graph explains the concept of percentiles.



**Figure 1.4.6** (The concept of percentiles)

### How can we calculate the percentages?

Let  $x_1, x_2, \dots, x_n$  be arranged data. Then:

- We calculate the **rank** of  $r^{\text{th}}$  percentile, whose denote by  $p_r$ , and is calculated by the following relation:

$$p_r = \frac{r(n+1)}{100} \quad ; r = 1, 2, \dots, 99$$

- The  $r^{\text{th}}$  **percentile**  $P_r$  can be calculated by the following relation:

$$P_r = x_k + s(x_{k+1} - x_k) \quad ; r = 1, 2, \dots, 99$$

Where  $k$  is the integer part of  $p_r$ , and the number  $s$  is the rest of  $p_r$ .

► **Example 1.4.16** Calculate the 35<sup>th</sup> percentile of the data given below.

40, 51, 92, 10, 36, 60, 70, 36, 36, 40, 80, 39, 53, 56, 60, 60, 70, 72, 88, 92, 50, 92, 20, 70, 38, 95, 56, 60, 88, 70.

**Solution:** We first arrange the data in the increasing (or ascending) order as follows:

10, 20, 36, 36, 36, 38, 39, 40, 40, 50, 51, 53, 56, 56, 60, 60, 60, 60, 70, 70, 70, 70, 72, 80, 88, 88, 92, 92, 92, 95.

Here we have  $n = 30$ . So

$$p_{35} = \frac{r(n+1)}{100} = \frac{35(30+1)}{100} = 10.85$$

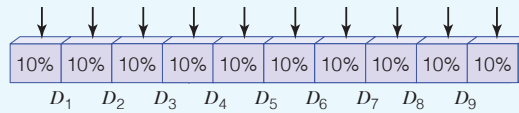
Also we have  $k = 10$  and  $s = 0.85$ . Therefore, we become that:

$$\begin{aligned} P_{35} &= x_k + s(x_{k+1} - x_k) = x_{10} + 0.85(x_{11} - x_{10}) \\ &= 50 + 0.85(51 - 50) = 50.85 \end{aligned}$$

Some commonly used percentiles are deciles and quartiles.

**DEFINITION 1.4.6 (Deciles)**

The percentiles (we denote them by  $D_1, D_2, \dots$  and  $D_9$ ) in multiple of 10 or equivalently deciles divide the data into 10 equal parts.  $D_1$  is the 10<sup>th</sup> percentile;  $D_2$  is 20<sup>th</sup> percentile and so on till  $D_9$  which is the 90<sup>th</sup> percentile. The following graph explains the concept of deciles.



**Figure 1.4.7** (The concept of deciles)

**How can we calculate the deciles?**

Let  $x_1, x_2, \dots, x_n$  be arranged data. Then:

- We calculate the **rank** of  $r$ th decile, whose denote by  $d_r$ , and is calculated by the following relation:

$$d_r = \frac{r(n+1)}{10} \quad ; r = 1, 2, \dots, 9$$

- The  $r^{\text{th}}$  **decile**  $D_r$  can be calculated by the following relation:

$$D_r = x_k + s(x_{k+1} - x_k) \quad ; r = 1, 2, \dots, 9$$

Where  $k$  is the integer part of  $d_r$ , and the number  $s$  is the rest of  $d_r$ .

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

For example, refer to the example above we calculate the sixth decile.

We have  $n = 30$ . So

$$d_6 = \frac{r(n+1)}{10} = \frac{6(30+1)}{10} = 18.6$$

Also we have  $k = 18$  and  $s = 0.6$ . Therefore, we become that:

$$\begin{aligned} P_{35} &= x_k + s(x_{k+1} - x_k) = x_{18} + 0.6(x_{19} - x_{18}) \\ &= 60 + 0.6(70 - 60) = 66 \end{aligned}$$

### DEFINITION 1.4.7 (Quartiles)

The Quartiles (we denote them by  $Q_1$ ,  $Q_2$  and  $Q_3$ ) quartiles divide the data into 4 equal parts. The first quartile  $Q_1$  is 25<sup>th</sup> percentile, the second quartile  $Q_2$  is 50<sup>th</sup> percentile which is also the median of the data and the third quartile  $Q_3$  is 75<sup>th</sup> percentile. The following graph explains the concept of quartiles.

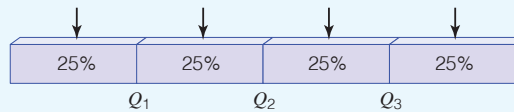


Figure 1.4.8 (The concept of quartiles)

### How can we calculate the quartiles?

Let  $x_1, x_2, \dots, x_n$  be arranged data. Then:

- We calculate the **rank** of  $r^{\text{th}}$  quartile, whose denote by  $q_r$ , and is calculated by the following relation:

$$q_r = \frac{r(n+1)}{4} \quad ; r = 1, 2, 3$$

- The  $r^{\text{th}}$  **decile**  $Q_r$  can be calculated by the following relation:

$$Q_r = x_k + s(x_{k+1} - x_k) \quad ; r = 1, 2, 3$$

Where  $k$  is the integer part of  $q_r$ , and the number  $s$  is the rest of  $q_r$ .

For example, refer to the example above we calculate the first quartile.

We have  $n = 30$ . So

$$q_1 = \frac{r(n+1)}{4} = \frac{(30+1)}{4} = 7.75$$

Also we have  $k = 7$  and  $s = 0.75$ . Therefore, we become that:

$$\begin{aligned} Q_1 &= x_k + s(x_{k+1} - x_k) = x_7 + 0.75(x_8 - x_7) \\ &= 39 + 0.75(40 - 39) = 39.75 \end{aligned}$$

► **EXAMPLE 1.4.17** Find the quartiles for the data given below

$$28, 22, 26, 29, 21, 23, 24.$$

**Solution:** We first arrange the data in the increasing order as follows:

$$21, 22, 23, 24, 26, 28, 29$$

Then:

For  $Q_1$  we have the rank

$$q_1 = \frac{r(n+1)}{4} = \frac{(7+1)}{4} = 2$$

Also we have  $k = 2$  and  $s = 0$ . Therefore, we become that:

$$Q_1 = x_k + s(x_{k+1} - x_k) = x_2 = 22$$

For  $Q_2$  (the median) we have the rank  $q_2 = \frac{r(n+1)}{4} = \frac{2(7+1)}{4} = 4$

Also we have  $k = 4$  and  $s = 0$ . Therefore, we become that:

$$Q_2 = x_k + s(x_{k+1} - x_k) = x_4 = 24$$

For  $Q_3$  we have the rank

$$q_3 = \frac{r(n+1)}{4} = \frac{3(7+1)}{4} = 6$$

Also we have  $k = 6$  and  $s = 0$ . Therefore, we become that:

$$Q_3 = x_k + s(x_{k+1} - x_k) = x_6 = 28$$

► **EXAMPLE 1.4.18** Find the quartiles of the data given below

$$7, 8, 15, 36, 39, 40, 41, 56$$

**Solution:** We first arrange the data in the increasing order as follows:

$$7, 8, 15, 36, 39, 40, 41, 56$$

In a similar way to the above we find:

$$q_1 = 2.25 \Rightarrow Q_1 = 8 + 0.25(15 - 8) = 9.75$$

$$q_2 = 4.5 \Rightarrow Q_2 = 36 + 0.5(39 - 36) = 37.5$$

$$q_3 = 6.75 \Rightarrow Q_3 = 40 + 0.75(41 - 40) = 40.75$$

An important use of quartiles is to determine whether a value  $x$  of given data is an extreme value.

## SECTION 1.4 MEASURES OF CENTRAL TENDENCY

### DEFINITION 1.4.8 (Extreme Values)

We say that a value  $x$  of given data is said to be extreme if one of the following relations is realizing:

$$x < Q_1 - 1.5(Q_3 - Q_1) \quad \text{or} \quad x > Q_3 + 1.5(Q_3 - Q_1)$$

► **EXAMPLE 1.4.19** Refer to the Example 1.4.17, then we find:

$$Q_1 - 1.5(Q_3 - Q_1) = 9.75 - 1.5(40.75 - 9.75) = -36.75$$

$$Q_3 + 1.5(Q_3 - Q_1) = 40.75 + 1.5(40.75 - 9.75) = 87.25$$

So the given data haven't an extreme value.

### DEFINITION 1.4.9 (Five Numbers)

These is a summary of the variable data which includes the below mentioned five characteristics

Smallest value,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , Largest value

The five numbers summary for Example 1.4.18 is given by **7**, **9.75**, **37.5**, **40.75** and **56**.

### DEFINITION 1.4.10 (Box Plot)

The box plot of given data is the graphical representation of its five numbers summary.

The steps to construct a box plot are as follows:

- First evaluate the five numbers summary for the given data.
- Draw an axis either horizontal or vertical on which the summary obtained can be located.
- Consider that horizontal axis is drawn; above the axis mark the quartiles, the minimum and the maximum and join them with a horizontal line.
- Draw vertical lines on all the three quartiles and join them making the box.

### REMARK 1.4.2

Extreme values are represented on the box plot by stars (\*) or dotes (•) above the corresponding values.

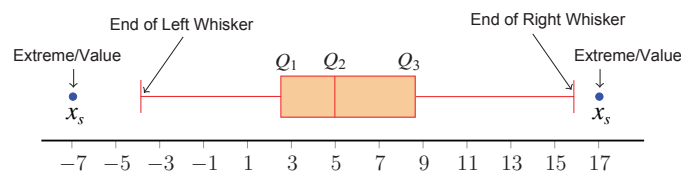


Figure 1.4.9

Note that the end of right whisker can be calculated as follow:

- If the data have not great extreme value, the end of right whisker equal to greatest value in the data.
- If the data have great extreme value, the end of right whisker equal to:

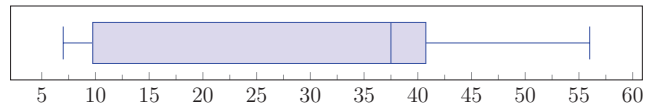
$$HF := Q_3 + 1.5(Q_3 - Q_1)$$

- If the data have not small extreme value, the end of left whisker equal to smallest value in the data.

- If the data have small extreme value, the end of left whisker equal to:

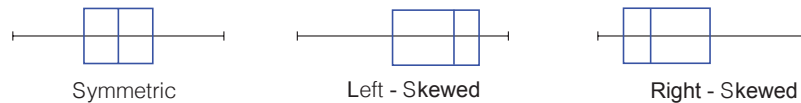
$$LF := Q_1 - 1.5(Q_3 - Q_1)$$

The box plot for the data given in Example 1.4.18 with five numbers summary is 7, 9.75, 37.5, 40.75 and 56 is given below.



**Figure 1.4.10** (Box plot for Example 1.4.18)

One can use box plot to determine whether data is symmetric or skewed. The following graph represents three different sets of data displaying the symmetric and the skewed data.



**Figure 1.4.11** (Types of data using box plot)

## Section 1.5

# MEASURES OF DISPERSION

In statistics, dispersion gives us the information about how the data is spread out. With the dispersion measures we can tell whether the data is stretched or squeezed. In other words, measures of variation provide us how the data is distributed. Measures of central tendency and dispersion together provide a good summary of the data in hand. We will now look at some measures of dispersion.

### DEFINITION 1.5.1 (Variance for raw data)

Let  $x_1, x_2, \dots, x_n$  be raw data with mean  $\bar{x}$ . Then the variance of this data (we denote it by  $S^2$ ) given by the following relation:

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

### REMARKS 1.5.1

- The variance unit is in square unit.
- We can also calculate the variance of raw data by using the following formula:

$$S^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right]$$

The above formula is nothing but a manipulation of the previous formula and the proof of this is left as an exercise for the reader.

### DEFINITION 1.5.2 (Standard Deviation)

The standard deviation (one denote it by  $S$ ) is the square root of the variance denoted by  $S^2$  and calculated by:

$$S = +\sqrt{S^2}$$

### REMARKS 1.5.2

- Standard deviation is the best measure of dispersion.



- b. This is used whenever the mean is used as the measure of central tendency. A small value of standard deviation indicates that the values of the variable tend to be close to the mean whereas a large value indicates that they tend to be far from the mean. (See the following graph).

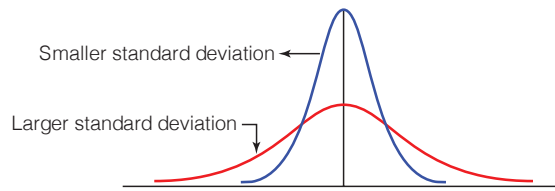


Figure 1.5.1

► **Example 1.5.1** Let 2, 3, 6, 8, 10, 13 and 14 be given data. Then we find the mean of this data is:

$$\bar{x} = \frac{2 + 3 + 6 + 8 + 10 + 13 + 14}{7} = \frac{56}{7} = 8$$

Now to calculate the variance and standard deviation for the given data we will build the following table:

Table 1.5.1

Variable value	Squared Variable values	Deviation from mean $(x_i - \bar{x})$	Squared Deviation $(x_i - \bar{x})^2$
2	4	$2 - 8 = -6$	36
3	9	-5	25
6	36	-2	4
8	64	0	0
10	100	2	4
13	169	5	25
14	196	6	36
$\sum_{i=1}^7 x_i = 56$	$\sum_{i=1}^7 x_i^2 = 578$	0	$\sum_{i=1}^7 (x_i - \bar{x})^2 = 130$

Therefore, we have

$$S = +\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = +\sqrt{\frac{130}{6}} = +\sqrt{21.667} = 4.65$$

We can solve the above example using this formula and calculate the standard deviation and the variance.

## SECTION 1.5 MEASURES OF DISPERSION

$$\begin{aligned}
 S^2 &= \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] \\
 &= \frac{1}{6} \left( 578 - \frac{56^2}{7} \right) = \frac{1}{6} (578 - 448) = \frac{1}{6} (130) = 21.667
 \end{aligned}$$

And then we have  $S = +\sqrt{21.667} = 4.65$ .

The variance comes out to be exactly same as in the previous case and so does the standard deviation.

### REMARKS 1.5.3

- We know that variance is the sum of squares of deviations and since squares are always positive or zero, therefore, we have  $S^2 \geq 0$ .
- Since the standard deviation is the positive square root of the variance, then is  $S \geq 0$ .
- The equality holds when all the deviations are zero, in that case all the values will be equal to a constant. Therefore, we can say  $S = 0$  if and only if all observed values of the variable are equal.

### DEFINITION 1.5.3 (Range for Raw Data)

We have introduced the definition of the range earlier when building a frequency distribution table, then we have  $R = x_\ell - x_s$ .

### DEFINITION 1.5.4 (Range for A Frequency Distribution Table)

We consider data in a frequency distribution table as in Table 1.4.1. Then the range is defined as follow:

$$R = x_k - x_1$$

Where  $x_k$  is the middle point of last class, and  $x_1$  is the middle point of the first class.

### ▶ EXAMPLES 1.5.2

- We consider the following sets of data:

$$X : 4, 8, 7, 3, 5, 10, 24, 5$$

$$Y : 10, 7, 9, 11, 11, 8, 9, 7$$

Then we find the range:

$$\text{For data } (X) \text{ equal to } R_X = x_\ell - x_s = 24 - 3 = 21.$$

$$\text{For data } (Y) \text{ equal to } R_Y = x_\ell - x_s = 11 - 7 = 4.$$

2. We consider the data in table 1.2.5 (frequency distribution table). Then we find the range of data equal to  $R = x_k - x_1 = 21 - 3 = 18$ .

#### DEFINITION 1.5.5 (Interquartile Range)

The Interquartile Range (one denote it by  $IQR$ ) of given data is defined as the difference between the first quartile and the third quartile.

$$IQR = Q_3 - Q_1$$

$IQR$  approximately gives us the range of the middle 50% of the observed value and hence it is also sometimes called as mid-spread.

► **Example 1.5.3:** Find the Interquartile range of the data given in Example 1.4.18.

**Solution:** In the Example 1.4.18 we calculated the quartiles of the given data which were as  $Q_1 = 9.75$ ,  $Q_2 = 37.5$  and  $Q_3 = 40.75$ . Therefore, we get that:

$$IQR = Q_3 - Q_1 = 40.75 - 9.75 = 31$$

#### DEFINITION 1.5.6 (Coefficient of Variation)

Let  $x_1, x_2, \dots, x_n$  be raw data with mean  $\bar{x} \neq 0$  and standard deviation  $S$ . Then the coefficient of variation (we denote it by  $CV$ ) is calculated as:

$$CV = \frac{S}{\bar{x}} \times 100 \%$$

One fact is worth noticing that with the five-number summary we can find the Range and the interquartile range. We can also find the median using the five-number summary.

#### DEFINITION 1.5.7 (z-scores)

Let  $x_1, x_2, \dots, x_n$  be raw data with mean  $\bar{x}$  and standard deviation  $S > 0$ . Then the standard score ( $z$ -scores and one denote it by  $z$ ) of data converts the data in such manner that the resultant data have mean 0 and a standard deviation 1. The following formula is used to calculate the standard score of a data:

$$z = \frac{x - \bar{x}}{S}$$

► **Example 1.5.4:** Let 2, 5, 3, 3, 7 be given data. Then to calculate the  $z$ -scores for this data we must calculate the mean and standard deviation of data. We find  $\bar{x} = 4$  and  $S = 2$ . So we get:

## SECTION 1.5 MEASURES OF DISPERSION

$$z_1 = \frac{x_1 - \bar{x}}{S} = \frac{2 - 4}{2} = \frac{-2}{2} = -1, \quad z_2 = \frac{x_2 - \bar{x}}{S} = \frac{5 - 4}{2} = \frac{1}{2}, \quad z_3 = \frac{x_3 - \bar{x}}{S} = \frac{3 - 4}{2} = \frac{-1}{2},$$

$$z_4 = \frac{x_4 - \bar{x}}{S} = \frac{3 - 4}{2} = \frac{-1}{2}, \quad z_5 = \frac{x_5 - \bar{x}}{S} = \frac{7 - 4}{2} = \frac{3}{2}$$

### THE CHEBYSHEV'S RULE

The standard deviation of data tells us about the spread of data. Using this theorem one can intuitively understand the significance of standard deviation. The theorem is as follows:

At least  $1 - \frac{1}{K^2}$  of the data lies within  $k$  standard deviation of the mean i.e. in the interval  $\bar{x} \pm kS$ , where  $k$  is any positive whole number greater than 1.

Using the above theorem, we can say that at least 75% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints  $\bar{x} \pm 2S$ , and at least 88.89% of the data lie within three standard deviations of the mean, that is, in the interval with endpoints  $\bar{x} \pm 3S$ .

### THE EMPIRICAL RULE

If a data set has an approximately bell-shaped relative frequency histogram,

1. Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints  $\bar{x} \pm S$ .
2. Approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints  $\bar{x} \pm 2S$ .
3. Approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints  $\bar{x} \pm 3S$ .

The following graph illustrates the concept of the empirical rule.

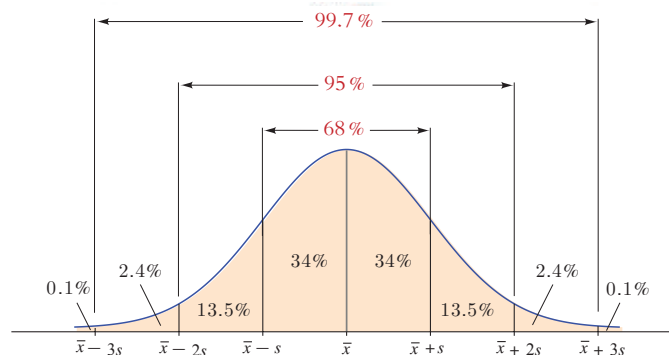


Figure 1.5.2 (Illustration of the concept of the empirical rule)

► **Example 1.5.5** Scores of some tests have a bell-shaped distribution with mean  $\mu = 100$  and standard deviation  $S = 10$ . Discuss what the Empirical Rule implies concerning individuals with scores of 110, 120, and 130.

**Solution:** The Empirical Rule states that:

1. Approximately 68% of the IQ scores in the population lie between 90 and 110,
2. Approximately 95% of the IQ scores in the population lie between 80 and 120, and,
3. Approximately 99.7% of the IQ scores in the population lie between 70 and 130.

We can use the empirical rule to determine what percentage of the values lies between given two points.

► **EXAMPLE 1.5.6** The mean price of apartments in a certain Saudi city is Saudi Riyals (SR) 500000 with standard deviation 100000. We will determine the price range for which at least 75% of the houses will sell.

Chebyshev's rule states that three-fourths, or 75% of the data values will fall within 2 standard deviations of the mean. Thus,

$$\begin{aligned}(\bar{x} - 2S, \bar{x} + 2S) &= (50000 - 2 \times 10000, 50000 + 2 \times 10000) = (30000, 70000) \\(\bar{x} - 2S, \bar{x} + 2S) &= (500000 - 2 \times 100000, 500000 + 2 \times 100000) \\ &= (300000, 700000)\end{aligned}$$

Hence, at least 75% of all apartments sold will have a price range from SR 300000 to SR 700000.



1. The value of  $\pi$  till 50 decimal places is given below:

3.14159265358979323846264338327950288419716939937510

- Make a frequency table of the digits from 0 to 9 after the decimal point.
- What are the most and the least frequently occurring digits?

2. Using the data shown in the following table:

No. of subjects in which student failed	Frequency
0	8
1	18
2	12
3	2
<b>Total</b>	<b>40</b>

- Represent them graphically using pie chart and Bar chart.
  - Compute the range of this data.
  - Compute the mean of this data.
3. The following data give the results of a sample survey. The letters A, B and C represent the three categories:

A	C	B	A	C	B	C	C	C	B
C	B	C	B	C	C	B	C	C	C
A	B	C	C	B	C	B	A	C	C

- Prepare a frequency table of this data.
  - Calculate the relative frequencies and percentages for all symbols.
  - What percentage of the elements belongs to category B?
  - Draw a bar chart and pie chart for the frequency table.
4. The following data give the results of a sample survey. The letters Y, N and D represent the three categories:

N	N	N	Y	Y	Y	N	Y	D	N
Y	Y	Y	Y	N	Y	Y	N	N	D
D	Y	Y	D	D	N	N	N	Y	N
Y	Y	N	N	Y	Y	N	N	D	Y

- Prepare a frequency distribution table.

- b. Calculate the relative frequencies and percentages for all categories.
  - c. What percentage of the elements belongs to category Y?
  - d. Draw a bar chart and pie chart for the given data.
5. A company manufactures car batteries of a particular type. The lives (in years) of 40 such batteries were recorded as follows:

2.6	3.0	3.7	3.2	2.2	4.1	3.5	4.5	4.6	3.8
3.5	2.3	3.2	3.4	3.8	3.2	4.6	3.7	2.9	3.6
2.5	4.4	3.4	3.3	2.9	3.0	4.3	2.8	3.5	4.2
3.5	3.2	3.9	3.2	3.2	3.1	3.7	3.4	3.2	2.6

Construct a frequency distribution table for this data, using class intervals of size 0.5 starting from the interval 2 – 2.5 .

6. The distance (in km) of 40 engineers from their residence to their place of work were found as follows:

5	3	10	20	25	11	13	7	12	31
19	10	12	17	18	11	32	17	16	2
7	9	7	8	3	5	12	15	18	3
12	14	0.5	9	6	15	15	7	6	12

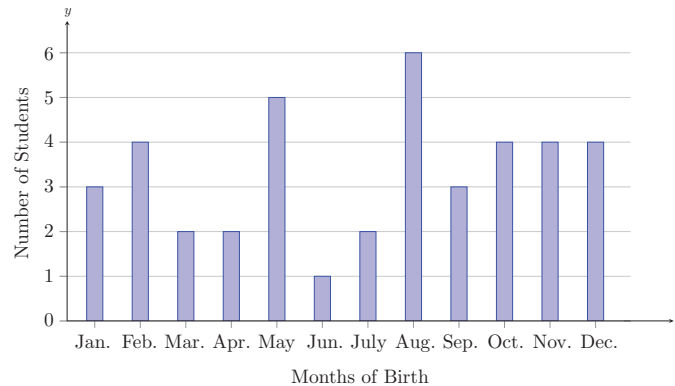
- a. Construct a frequency distribution table with class size 5 for the data given above.
  - b. Draw the histogram for the data of frequency distribution table.
  - c. Draw the polygon for the data of frequency distribution table.
  - d. Draw the ogive for the data of frequency distribution table.
  - e. How many engineers have residence at distance less than 20 km from their workplace?
  - f. How many engineers have residence at distance more than 15 km from their workplace?
7. Thirty children were asked about the number of hours they watched TV programs in the previous week. The results were found as follows:

8	10	12	14	12	10	8	6	4	2
10	3	4	12	2	8	15	1	17	6
1	6	2	3	5	12	5	8	4	8
3	2	8	5	9	6	8	7	14	12

- a. Construct a frequency distribution table for this data.

## EXERCISES

- b. Draw the histogram, polygon and ogive for the frequency distribution table.
  - c. Draw the polygon for the data of frequency distribution table.
  - d. Draw the ogive for the data of frequency distribution table.
8. In a particular section of Class X, 40 students were asked about the months of their birth and the following graph was prepared for the data so obtained:



Observe the bar graph given above and answer the following questions:

- a. How many students were born in November?
  - b. In which month were the maximum numbers of students born?
9. A sample of 100 children was asked how many times they play computer games for a period of one week. The following table gives the frequency distribution of their responses

Number of times of playing	Number of children
0 – 3	23
4 – 7	40
8 – 11	28
12 – 15	6
16 – 19	3

- a. Find the class midpoints?
  - b. Do all classes have the same width? If so, what is this width?
  - c. Prepare the relative frequency and percentage distribution columns?
  - d. What percentage of these children plays 8 or more times a week?
10. consider the following frequency distribution, representing the weights of 36 students of a class:



Weights (in kg)	Number of students
40 → 50	9
50 → 60	6
60 → 70	15
70 → 80	5
80 → 90	1
<b>Total</b>	<b>36</b>

- Draw the histogram, polygon and ogive for the above table.
- How many students have weights less than 70 Kg?

11. The following table gives the life times of 400 neon lamps:

N.o.C.	Lifetime (in hours)	Number of lamps
1	200 → 300	14
2	300 → 400	56
3	400 → 500	60
4	500 → 600	76
5	600 → 700	64
6	700 → 800	52
7	800 → 900	40
8	900 → 1000	38

- Represent the given information with the help of a histogram.
- How many lamps have a life time of more than 700 hours?

12. The following table gives the distribution of students of two sections according to the marks obtained by them:

Marks	Section A Frequency	Section B Frequency
0 → 10	3	5
10 → 20	9	19
20 → 30	17	15
30 → 40	12	10
40 → 50	9	1

Represent the marks of the students of both the sections on the same graph by two frequency polygons. From the two polygons compare the performance of the two sections.

13. Consider the following frequency distribution, representing the degree of an examination of 50 students of a class:

## EXERCISES

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency (ACF)
2 - 6			6		
7 - 11				0.24	
12 - 16					36
17 - 21				0.12	
22 - 26			8		
<b>Total</b>			<b>50</b>		

Then:

- Complete the above frequency distribution table.
  - Draw the histogram, polygon and ogive for this frequency distribution table.
  - Calculate the mean, median and mod for the above frequency distribution table.
  - Calculate the standard deviation of the above frequency distribution table.
14. The points scored by a team in a series of matches are as follows:  
17, 2, 7, 27, 15, 5, 14, 8, 10, 24, 48, 10, 8, 7, 18, 28

Then:

- Calculate the mean and standard deviation of the given data.
  - Calculate the standard score of the value (7) in the given data.
  - Calculate the coefficient of variation for the given data.
  - Calculate  $Q_1$ ,  $Q_2$  and  $Q_3$ .
15. Consider the marks obtained (out of 100 marks) by 30 students of Class X of a school:

10	20	36	92	95	40	50	56	60	70
92	88	80	70	72	70	36	40	36	40
92	40	50	50	56	60	70	60	60	88
92	88	80	70	72	70	36	40	36	40
92	40	50	50	56	60	70	60	60	88

Then:

- Calculate the mean and standard deviation of the given data.
  - Calculate  $P_{10}$ ,  $P_{50}$  and  $P_{93}$ .
  - Calculate  $D_3$ ,  $D_5$  and  $D_8$ .
  - Calculate  $Q_1$ ,  $Q_2$  and  $Q_3$ .
16. Consider a small unit of a factory where there are 5 employees: a supervisor and four laborers. The laborers draw a salary of \$ 5,000 per month each while the supervisor gets \$ 15,000 per month. Calculate the mean, median and mode of the salaries of this unit of the factory.

17. The daily sale of sugar (kg) in a certain grocery shop is given below:

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
75	120	12	50	70.5	140.5

- Calculate the average daily sale.
  - Calculate the variance and the standard deviation of the above data.
  - Determine the coefficient of variation.
18. Let the following data be marks obtained (out of 100) by 10 students in a test:
- 45, 45, 63, 76, 67, 84, 75, 48, 62, 65
- Then:
- Calculate  $Q_1$ ,  $Q_2$  and  $Q_3$ .
  - Calculate the *IQR*.
  - Have the given data extreme values?
  - Construct the box plot for the given data.
19. Consider the following data:
- 40, 40, 40, 60, 65, 65, 70, 70, 75, 75, 75, 80, 85, 90, 90, 150
- Then:
- Calculate  $Q_1$ ,  $Q_2$  and  $Q_3$ .
  - Calculate the *IQR*.
  - Have the given data extreme values?
  - Construct the box plot for the given data.
20. Consider the following data:
- 40, 45, 55, 65, ?, ?, 75, 75, 78, 83
- Then use the suitable measure to calculate the average and dispersion for the given data.
21. Consider the following data:
- 15, 20, 40, 50, 65, 65, 70, 73, 75, 137
- Have the given data extreme values?
  - Use the suitable measure to calculate the average and dispersion for the given data.
22. The following data give the number of computer keyboards assembled at a company for a sample of 25 days

## EXERCISES

45	52	48	41	56
46	44	42	48	53
51	53	51	48	46
43	52	50	54	47
44	47	50	49	52

Prepare a box-plot and then comment the skewness of these data.

- 23.** Create a dot plot for the following data set

1	2	0	5	1
1	3	2	0	5
2	1	2	1	2
0	1	3	1	2

- 24.** The mean age of six persons is 49 years. The ages of five of these six persons are 55, 39, 44, 51, and 45 years respectively. Find the age of the sixth person.

- 25.** The following data give the masses in grams to the nearest gram, of 10 eggs.

46, 51, 48, 62, 54, 56, 58, 60, 71, 75

- Calculate the mean, median, and standard deviation of this data.
  - Calculate the five numbers summary and construct the box plot of this data.
- 26.** The following observations have been arranged in ascending order.

29, 32, 48, 50,  $x$ ,  $x + 2$ , 72, 78, 84, 95

Now, if the median of the data is 63, then:

- Calculate the value of  $x$ .
  - Calculate the mean, and standard deviation of this data.
  - Find the five numbers summary and construct the box plot of this data.
- 27.** Consider the following two data sets.

Data set I: 12, 25, 37, 8, 41

Data set II: 19, 32, 44, 15, 48

Notice that each value of the second data set is obtained by adding 7 to the corresponding value of the first data set. Then:

- Calculate the mean for each of these two data sets. Comment on the relationship between the two means.
- Calculate the standard deviation for each of these two data sets. Comment on the relationship between the two standard deviations.

- c. Calculate the standard score of the value (37) in data set I.
- d. Calculate the coefficient of variation for each of these two data sets, then compare them.

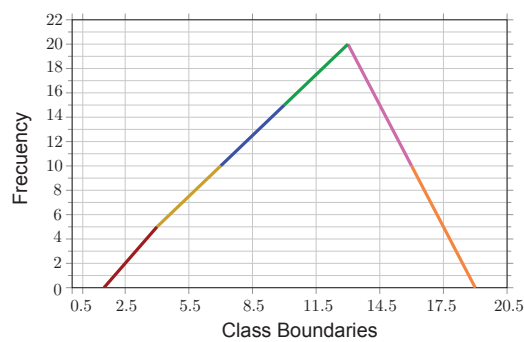
28. Consider the following two data sets.

Data set I: 4, 8, 15, 9, 11

Data set II: 12, 24, 45, 27, 33

Notice that each value of the second data set is obtained by multiplying the corresponding value of the first data set by 3.

- a. Calculate the mean for each of these two data sets. Comment on the relationship between the two means.
  - b. Calculate the standard deviation for each of these two data sets. Comment on the relationship between the two standard deviations.
  - c. Calculate the standard score of the value (27) in data set II.
  - d. Calculate the coefficient of variation for each of these two data sets, then compare them.
29. Using the following data set 15, 15, 15, 15, 15, 15. Then:
- a. Calculate the standard deviation.
  - b. Is its value of the standard deviation equal to zero? If yes, why?
30. Using the following data set {7}. Then:
- c. Calculate the standard deviation of this data set.
  - d. Is its value of the standard deviation equal to zero? If yes, why?
31. Consider the following polygon of grouped data, representing the degree of an examination of 60 students:



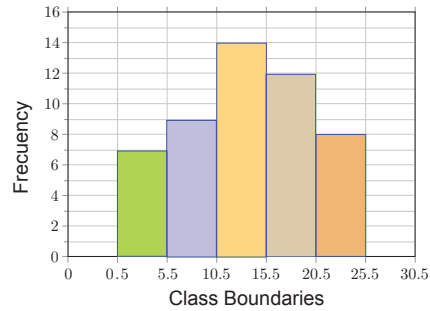
Then:

- a. Prepare the frequency distribution table of this data.
- b. Draw the histogram ogive for this table.

## EXERCISES

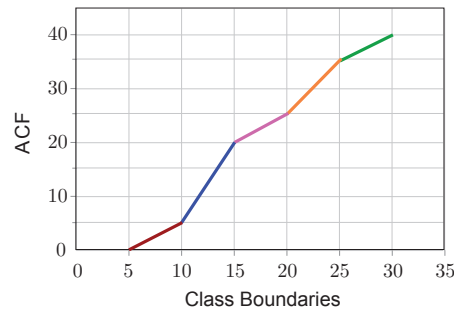
c. Calculate the mean, median and mod for this table.

- 32.** Consider the following histogram of grouped data, representing the temperatures in 50 cities of Europe:



Then:

- a. Prepare the frequency distribution table of this data.  
b. Draw the polygon and ogive for this table.  
c. Calculate the standard deviation for this table.
- 33.** Consider the following ogive of grouped data, representing the weight to 30 fruit boxes:



- a. Prepare the frequency distribution table of this data.  
b. Draw the histogram and polygon for this table.

# CHAPTER 2

## PROBABILITY



### LEARNING OBJECTIVES

After completing this chapter, you should be able to:

1. Define and explain the terms sample space, event, mutually exclusive, and Venn diagram.
2. Counting techniques: multiplicative rule, permutation and combination.
3. Different approaches to assigning probabilities.
4. Explain what is meant by marginal & conditional probability.
5. Contrast independent and dependent events.
6. Total probability and Bayes' theorem.

### INTRODUCTION

The techniques and methods covered in this chapter will be useful in covering the material in Chapter 3. In probability and statistics, we sometimes need to count the number of ways that a phenomenon can occur. In addition, probabilities express the degree of certainty in managerial decision making or any situation involving uncertainty. Assigning probabilities to future events allows us to analyze decision options in a rational way. Probability does not tell us exactly what will happen, it is just a guide.

- SECTION 2.1 MATHEMATICAL CONCEPTS
- SECTION 2.2 DEFINITIONS AND CONCEPTS IN PROBABILITY CALCULAS
- SECTION 2.3 CONCEPT OF PROBABILITY FUNCTION
- SECTION 2.4 CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

## Section 2.1

# MATHEMATICAL CONCEPTS

Below we present some mathematical concepts needed in the study of probability. The first concept we will give is known as the fundamental principle of counting.

### THE FUNDAMENTAL PRINCIPLE OF COUNTING

The fundamental principle of counting includes two rules:

#### Multiplicative Rule

If we have  $k$  phenomenon  $O_1, O_2, \dots$  and  $O_k$ . So that these phenomenon occurs in  $n_1, n_2, \dots$  and  $n_k$  ways respectively. Then the number of ways that all these phenomenon occurs in one time is  $n_1 \times n_2 \times \dots \times n_k$  ways.

#### Addition Rule

If we have  $k$  phenomenons  $O_1, O_2, \dots$  and  $O_k$ . So that these phenomenons occurs in  $n_1$  or  $n_2$  or  $\dots$  or  $n_k$  ways respectively. Then the number of ways that all these phenomenons occurs in one time is  $n_1 + n_2 + \dots + n_k$  ways.

#### ► EXAMPLES 2.1.1

1. At a restaurant, for a fixed price a person may choose from one of four salads, one of five entrées, and one desert. How many different meals are possible, if the person must select one salad, one entrée, and one dessert?

**Solution:** One possible meal a person may have is the first path through the tree, " $S_1 E_1 D$ ". By the multiplication rule, we obtain that the total number of possible meals is:

$$4 \times 5 \times 1 = 20.$$

2. How many 6-digits zip codes are possible if:
  - a. digits can be repeated?
  - b. digits cannot be repeated?

**Solutions:** For the item:

- a. If digits can be repeated. Then by the multiplication rule, we obtain that the number of 6-digits zip codes equal to:

$$10 \times 10 \times 10 \times 10 \times 10 \times 10 = 10^6 = 1000000 \text{ codes.}$$



- b. If digits cannot be repeated. Then by the multiplication rule, we obtain that the number of 6-digits zip codes equal to:

$$10 \times 9 \times 8 \times 7 \times 6 \times 5 = 151200 \text{ codes.}$$

3. How many book can we index if we use Arabic characters or English characters?

**Solutions:** By the addition rule, we can index  $28 + 26 = 54$  books.

#### DEFINITION 2.1.1 (Factorial Notation)

If  $n$  is a whole number, which means  $n$  is an element of the set  $\{0, 1, 2, 3, 4, 5, 6, \dots\}$ , then  $n$  factorial, written as  $(n!)$ , and is defined by the following relation:

$$n! = n \times (n - 1) \times (n - 2) \times (n - 3) \times \dots \times 3 \times 2 \times 1.$$

Note that we have (as result of the gamma function) the following:

$$0! = 1 \text{ and } 1! = 1$$

#### DEFINITION 2.1.2 (Permutations)

Any an ordered arrangement of  $r$  distinct objects, from a set of  $n$  different objects called permutation.

We use the notation  $nPr$  to represent the total number of different permutations of size  $r$  that can be selected from  $n$  distinct objects.

The general formula for computing the total number of permutations of size  $r$  selected from  $n$  distinct objects is

$$nPr = \frac{n!}{(n-r)!} \quad ; 0 \leq r \leq n$$

► **EXAMPLE 2.1.2** How many ways one can arrange in order any three of the first 8 letters of the alphabet  $l, m, o, p, q, r, s, t$ .

**Solution:** The permutation formula gives:

$${}_8P_3 = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{8 \times 7 \times 6 \times (5!)}{5!} = 8 \times 7 \times 6 = 336.$$

**DEFINITION 2.1.3 (Combinations)**

Any an unordered group of  $r$  distinct objects, from a set of  $n$  different objects is called combination. We use the notation  $nCr$  to represent the total number of different combinations of size  $r$  that can be selected from  $n$  distinct objects and read as " $n$  choose  $r$ ".

The general formula for computing the number of combinations of size  $r$  selected from  $n$  distinct objects is:

$$nCr = \frac{n!}{r!(n-r)!} \quad ; 0 \leq r \leq n$$

**REMARK 2.1.1**

The symbol  $\binom{n}{r}$  can be used instead of  $\frac{n!}{r!(n-r)!}$ .

► **EXAMPLE 2.1.3** How many different unordered groups of any three of the six letters  $l, m, n, o, p$  and  $q$  exist?

**Solution:** With  $n = 6$  and  $r = 3$ , one has

$${}^6C_3 = \frac{6!}{3!(6-3)!} = \frac{6 \times 5 \times 4 \times (3!)}{3 \times 2 \times (3!)} = 20$$

► **EXAMPLE 2.1.4** From a class of 10 students, a group of 3 will be chosen to do a job. How many different groups of students are possible?

**Solution:** The number of different groups of students possibly doing the job is:

$${}^{10}C_3 = \frac{10!}{3!(10-3)!} = \frac{10 \times 9 \times 8 \times (7!)}{(3 \times 2 \times 1) \times (7!)} = 120$$

**DEFINITION 2.1.4 (Cardinal Number of a Set)**

Let  $\Omega$  be a given set, then  $|\Omega|$  denote the number of all elements in  $\Omega$ , and this number is called the cardinal number of  $\Omega$ .

**REMARKS 2.1.2**

1. If the set  $\Omega$  is infinite, and we can number its elements by the number of natural number  $\mathbb{N}$ , then we say that the set  $\Omega$  is a countable set, and we write  $|\Omega| = \infty$ .

2. If the set  $\Omega$  is infinite, and we cannot number its elements by the number of natural number  $\mathbb{N}$ , then we say that the set  $\Omega$  is a un countable set, and we write  $|\Omega| = \varphi$ , we say that the set  $\Omega$  has continuous capacity.

## Section 2.2

# DEFINITIONS AND CONCEPTS IN PROBABILITY CALCULAC

### TYPES OF EXPERIMENTS

Experiments performed by a person are usually divided into two types:

1. **Regular (or Systematic) experiments**, which we know the results of it in advance and with precision.
2. **Random (or Stochastically) experiments**, which we don't know its exact outcome in advance, but we can determine the set of all its possible results only.

► **EXAMPLES 2.2.1** The following situations are regular experiments:

- a. The reaction of concentrated chlorinated water with pure aluminum.
- b. Concentrated sulfuric acid reaction with pure copper.

The following situations are random experiments:

- a. Tossing a coin has two possibilities, head ( $H$ ) or Tail ( $T$ ).
- b. Tossing a die and observe the number appears on top.
- c. Tossing two dice and observe the number appears on top.
- d. A football team plays two games and in each game either wins ( $W$ ) or be equal ( $D$ ) or losses ( $L$ ).

### REMARKS 2.2.1

1. For a random experiment, each possible outcome is called an elementary event.
2. Two or more outcomes that have the same chance (appearance) of occurrence are said to be equally likely outcomes.

### DEFINITION 2.2.1 (Probability Science)

Probability Science is a branch of mathematics that deals with theoretical mathematical models of random experiments.

The question here is: *What is the theoretical mathematical model of a random experiment?*

The theoretical mathematical model (and is called **probability space** also) of any random experiment is a triple have the form  $[\Omega, \mathcal{A}, P]$ , where:

$\Omega$  is the set of all possible results of the random experiment.

$\mathcal{A}$  is called the algebra of events.

$P$  is called a probability measure (or probability function).

Through this chapter we will quickly and succinctly identify these three important terms in probability theory.

Below we will explain the first element of probability space of a random experiment.

When we specify all possible outcomes in a random experiment or a stochastic study, we are stating the space of elementary events.

#### DEFINITION 2.2.2 (Space of Elementary Events)

Suppose that we have a random experiment. Then the space of elementary events is the collection of all outcomes of this random experiment, and denoted by  $\Omega$ . Moreover, any outcome is called an elementary event.

► **EXAMPLE 2.2.2** Let us conduct a set during which we first flip a coin and then roll a die. For example, one of the outcomes of the random experiment is getting a head ( $H$ ) from flipping the coin and then getting a “1” from rolling the die.

Then the set of all the possible outcomes in this experiment of tossing a coin followed by rolling a die is:

$$\Omega = \{(H,1), (H,2), (H,3), (H,4), (H,5), (H,6), (T,1), (T,2), (T,3), (T,4), (T,5), (T,6)\}$$

So we have  $|\Omega| = 12$  outcomes (elementary events).

► **EXAMPLE 2.2.3** Refer to Example 2.2.1 determine spaces of elementary events of those events.

**Solution:** The space of elementary events for the previous random experiments in Example 2.2.1 is:

**For a)** we have  $\Omega = \{H, T\}$ , so we have  $|\Omega| = 2$  outcomes.

**For b)** we have  $\Omega = \{1, 2, 3, \dots, 6\}$ , so we have  $|\Omega| = 6$  outcomes.

**For c)** we have  $\Omega = \{(1,1), (1,2), \dots, (1,6), (2,1), \dots, (2,6), \dots, (6,6)\}$ , so we have  $|\Omega| = 36$  outcomes.

**For d)** we have:

$$\Omega = \{(W,W), (W,D), (W,L), (D,W), (D,D), (D,L), (L,L), (L,D), (L,W)\}$$

So we have  $|\Omega| = 10$  outcomes.

► **EXAMPLE 2.2.4** Tossing a coin three times and observe the sequence of heads ( $H$ ) and tails ( $T$ ) that appears. Write down the space of elementary events for this random experiment and determine the following events.

$E_1$ : is the event that two heads only occurs.

$E_2$ : is the event that at least two heads occurs.

$E_3$ : is the event that at most two heads occurs.

$E_4$ : is the event that a heads is the first toss.

**Solution:** For the given random experiment, we have:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

$$E_1 = \{HHT, HTH, THH\}$$

$$E_2 = \{HHT, HTH, THH, HHH\}$$

$$E_3 = \{HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

$$E_4 = \{HHH, HHT, HTH, HTT\}$$

### TYPES OF SPACE OF ELEMENTARY EVENTS

The space of elementary events can be classified into two basic types.

#### DEFINITION 2.2.3 (Discrete Space)

If a space of elementary events  $\Omega$  is either finite or countable infinite, then it is called a discrete space.

#### DEFINITION 2.2.4 (Continuous Space)

If the space of elementary events  $\Omega$  consists uncountable number of outcomes, then it is called a continuous space.

Now we will present the concept of algebra for events.

#### DEFINITION 2.2.5 (Algebra of Events)

Suppose that we have a random experiment with a space of elementary events  $\Omega$ . Then a collection  $\mathcal{A}$  of subset of  $\Omega$  is said to be an algebra on  $\Omega$  if and only if the following condition are verified:

1.  $\Omega \in \mathcal{A}$ .
2. For any two element  $A$  and  $B \in \mathcal{A}$ . Then  $A \cup B \in \mathcal{A}$ .
3. For any element  $A \in \mathcal{A}$ . Then  $\bar{A} \in \mathcal{A}$ .

**REMARKS 2.2.2:**

1. If algebra  $\mathcal{A}$  fulfills the following condition:

For any sequence  $A_1, A_2, \dots, A_n, \dots \in \mathcal{A}$ . Then  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ .

Then  $\mathcal{A}$  is said to be an  $\sigma$  – algebra.

2. The elements of  $\mathcal{A}$  (as an algebra or an  $\sigma$  – algebra) are called events.
3. In the triple  $[\Omega, \mathcal{A}, P]$  must  $\mathcal{A}$  to be an  $\sigma$  – algebra.
4. When  $\Omega$  is finite or countable infinite ( $\Omega$  is a discrete space), then any subset of  $\Omega$  is an event. But this statement is not true if  $\Omega$  is a continuous space. For this latter case, there are specially studies of its.
5. In our next study (as an illustration) we will take  $\Omega$  a finite, and in this case one can prove that any algebra on  $\Omega$  is an  $\sigma$  – algebra on  $\Omega$  also.
6. When we talk about events, we will always assume that we have a given random experiment.

**DEFINITION 2.2.6 (Compound Event and Simple Event)**

In the general case, an event consists one or more elementary event, and these elementary event have a common characteristic. Now if an event  $A$  contain only one elementary event (outcome), then  $A$  is called a simple event. Except that, the event  $A$  is called a compound event.

- **EXAMPLE 2.2.5** Tossing a single die once, so the space of elementary events is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Then we find that the events  $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$  are simple events. ◀

- **EXAMPLE 2.2.6** Tossing a coin twice, so the space of elementary events is  $\Omega = \{HH, HT, TH, TT\}$ .

Then we find that the events  $\{HH, HT\}, \{HH, TH\}, \{TH, TT\}, \{HH, HT, TT\}$  are compound events. ◀

- **EXAMPLE 2.2.7** A box contains a few red and a few green balls. If two balls are randomly drawn (one after the other) and the colors of these balls are observed, how many total outcomes are possible?

## SECTION 2.2 DEFINITIONS AND CONCEPTS IN PROBABILITY CALCULAS

List all the outcomes included in each of the following events. Indicate which are simple and which compound events are.

- $E_1$ : both balls are of different colors.
- $E_2$ : at least one ball in red.
- $E_3$ : at most one ball is red.
- $E_4$ : the first ball is green and the second is red.

**Solution:** We have  $\Omega = \{RR, RG, GR, GG\}$ , and we suppose that  $R$  is the event that a red ball is selected, and  $G$  is the event that a green ball is selected. Then we find:

**For a)** we have  $E_1 = \{RG, GR\}$ , so  $E_1$  is compound event.

**For b)** we have  $E_2 = \{RG, GR, RR\}$ , so  $E_2$  is compound event.

**For c)** we have  $E_3 = \{RG, GR, GG\}$ , so  $E_3$  is compound event.

**For d)** we have  $E_4 = \{GR\}$ , so  $E_4$  is simple event.

### REMARK 2.2.3

1. If  $\Omega$  is a set, then we denote the set of all subset in  $\Omega$  by  $2^\Omega$ . One can see that the number of all elements in  $2^\Omega$  equal to  $2^{|\Omega|}$ . Knowing that  $|\Omega|$  is the cardinal number of  $\Omega$ . For example, if the we have a set  $\Omega = \{a, b, c\}$ , then the elements of  $2^\Omega$  are:

$$\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\} = \Omega \quad \Rightarrow \quad |2^\Omega| = 2^{|\Omega|} = 2^3 = 8$$

2. If  $\Omega$  is a space of elementary events of a random experiment, and  $\Omega$  is finite or countable infinite, then  $2^\Omega$  is an  $\sigma$ -algebra on  $\Omega$ , therefore the elements of  $2^\Omega$  are events. But if  $\Omega$  is uncountable, as  $\Omega = \mathbb{R}$ , then can be exist elements of  $2^\Omega$  which not events. In this case one denote the  $\sigma$ -algebra on  $\Omega$  by  $\mathfrak{R}$  ( $\mathfrak{R}$  is called Boreal field).
3. In the following study we will take  $\Omega$  a finite or countable (for simplicity), therefore the elements of  $2^\Omega$  are events. Therefore (to avoid repetition), when we write  $A, B$  and ... are events, we mean it that we have a random experiment with space of elementary events  $\Omega$  and  $\sigma$ -algebra  $2^\Omega$ , and the events  $A, B$  and ... are from  $2^\Omega$ .



**SOME OPERATIONS ON EVENTS**

Since a space of elementary events  $\Omega$  is a set while, an event  $E$  is a subset of  $\Omega$ . We can form new events by using the usual operations of set theory.

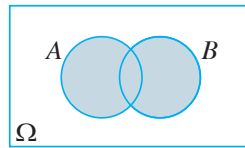
For some of these operations some different terminology is used in probability theory than in set theory.

**Union of Two Events (This Expression is Metaphorical)**

The union of the two events  $A$  and  $B$  denoted by  $A \cup B$ , is an event containing all the elementary events that belong to  $A$  **or**  $B$  or to both. This means, the event  $A \cup B$  is the occurrence of at least one of the two events.

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$$

where  $\omega$  refer to any element belong to  $A$  or  $B$  or to both.



**Figure 2.2.1** (Shaded region represents the event  $A \cup B$ )

► **EXAMPLE 2.2.8** Let  $\Omega = \{1, 2, 3, 4, 5, 6\}$  be space of elementary events,  $A = \{1, 2\}$  and  $B = \{2, 5, 6\}$ , then we have  $A \cup B = \{1, 2, 5, 6\}$ . ◀

► **EXAMPLE 2.2.9** In the random experiment of tossing a coin twice, let the event  $A$  is to get a head in the first toss and  $B$  is to get a head in the second toss, so we have  $\Omega = \{HH, HT, TH, TT\}$  is the space of elementary events, and  $A = \{HH, HT\}$ ,  $B = \{HH, TH\}$ . Therefore, we have: ◀

$$A \cup B = \{HH, HT, TH\}$$

**Intersection of Two Events (This Expression is Metaphorical)**

The intersection of the two events  $A$  and  $B$ , denoted  $A \cap B$ , is an event containing all elementary events that are common to  $A$  **and**  $B$ .

If  $A$  and  $B$  are any two events of  $2^\Omega$ , then  $A \cap B$  is the event of occurring both  $A$  and  $B$  together. This means:

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$$

Both the two events  $A$  and  $B$  occurs.

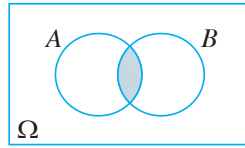


Figure 2.2.2 (Shaded region represents the event  $A \cap B$ )

► **EXAMPLE 2.2.10** In the random experiment of tossing a coin twice, let the event  $A$  is to get a tail in the first toss and  $B$  is to get tail in the second toss, we have:

$$\Omega = \{HH, HT, TH, TT\}, A = \{TH, TT\}, B = \{HT, TT\}$$

And then we have  $A \cap B = \{TT\}$ .

### Complement of an Event

Let  $A$  be an event of  $2^\Omega$ , the complement of  $A$  with respect to  $\Omega$ , is an event that occurs if  $A$  does not occur, and is denoted by  $\bar{A}$ . So we have:

$$\bar{A} = \{\omega : \omega \in S, \omega \notin A\}$$

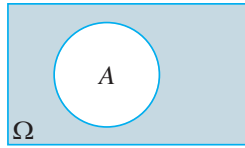


Figure 2.2.3 (Shaded region represents the event  $\bar{A}$ )

For every event  $A$  their corresponds the event  $\bar{A}$  is called the complement of  $A$ , consisting of all elementary events of  $\Omega$  which are not in  $A$ .

► **EXAMPLE 2.2.11** Tossing a die and observe the number that appears on top. Determine the following events

- $E_1$ : is the event that an even or a prime number occurs.
- $E_2$ : is the event that a prime number occurs.
- $E_3$ : is the event that a prime number does not occur.

**Solution:** We have  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , and then we have:

**For a)** the event  $E_1 = \{2, 3, 4, 5, 6\}$ ,

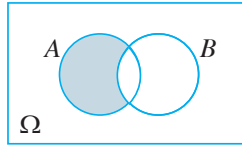
**For b)** the event  $E_2 = \{2, 3, 5\}$ ,

**For c)** the event  $E_3 = \{1, 4, 6\}$ .

**Difference Between Two Events (This Expression is Metaphorical)**

If  $A$  and  $B$  are two events from  $\Omega$ , then  $A \setminus B$  or  $A \cap \bar{B}$  means the event of the occurrence of  $A$  but not  $B$ , i.e.  $A$  occurs and  $B$  does not occur or only  $A$  must occur.

$$A \setminus B \equiv A \cap \bar{B} = \{\omega : \omega \in A \text{ and } \omega \notin B\}$$



**Figure 2.2.4** (Shaded region represents the event  $A \setminus B$ )

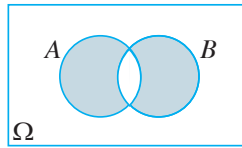
**Exactly One of Event**

If  $A$  and  $B$  are two events from  $\Omega$  then:

$$(A \setminus B) \cup (B \setminus A) \text{ or } (A \cap \bar{B}) \cup (B \cap \bar{A})$$

means  $A$  or  $B$  occurs:

$$A \Delta B = \{x : x \in A \cap \bar{B} \text{ or } x \in B \cap \bar{A}\}$$



**Figure 2.2.5** (Shaded region represents the event  $A \Delta B$ )

**DEFINITION 2.2.7 (Impossible Event)**

For an event  $A \in 2^\Omega$  we know that it is impossible an outcome of the experiment belong to  $A \cap \bar{A}$ . Therefore, the event  $A \cap \bar{A}$  is called an impossible event, and since in set theory  $A \cap \bar{A} = \emptyset$ , so one denotes the impossible event by  $\emptyset$  also.

**▶ EXAMPLES 2.2.12**

1. In the random experiment of tossing a die once, the event getting a number, which is divisible by 7, is an impossible event.
2. In the random experiment of tossing a coin 3 times, the event that 5 heads appear is an impossible event.

**DEFINITION 2.2.8 (Certain Event)**

For an event  $A \in 2^\Omega$  we know that it is surly an outcome of the experiment belong to  $A \cup \bar{A}$ . Therefore, the event  $A \cup \bar{A}$  is called a certain event, and since in set theory  $A \cup \bar{A} = \Omega$ , so one denotes the certain event by  $\Omega$  also.

► **EXAMPLES 2.2.13**

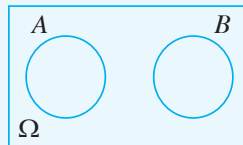
1. In the random experiment of tossing a single die once, the event of getting a number less than 7 is a certain event.
2. If the space of elementary events of a random experiment is  $\Omega = \{1, 2, 3\}$ , then the elements of  $2^\Omega$  are:

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\} \Rightarrow |2^\Omega| = 2^3 = 8$$

We note that  $\emptyset$  is the impossible event,  $\Omega$  is the certain event,  $\{1\}$ ,  $\{2\}$ ,  $\{3\}$  are simple events,  $\{1, 2\}$ ,  $\{1, 3\}$ ,  $\{2, 3\}$  are compound events.

**DEFINITION 2.2.9 (Mutually Exclusive Event)**

Two events  $A$  and  $B$  of  $2^\Omega$  are called mutually exclusive events if  $A \cap B = \emptyset$ . That is  $A$  and  $B$  have no elementary events in common.  $A$  and  $B$  are mutually exclusive if they cannot occur simultaneously.



**Figure 2.2.6** ( $A$  and  $B$  are mutually exclusive events)

**DEFINITION 2.2.10 (Pair Wise Mutually Exclusive Event)**

Events  $A_1, A_2, A_3, \dots$  of  $2^\Omega$  are said to be pair wise mutually exclusive if:

$$A_i \cap A_j = \emptyset \quad ; \forall i \neq j$$

► **EXAMPLE 2.2.14** One throws a single die and observes the number that appears on top.

List each of the following events:

- a.  $A$ : appearance of an odd number,
- b.  $B$ : appearance of an even number,
- c.  $C$ : appearance of a prime number, and
- d. Find the events  $A \cap B$ ,  $A \cap C$ ,  $B \cap C$  and  $A \cap B \cap C$ .

**Solution:** We have  $\Omega = \{1, 2, 3, 4, 5, 6\}$

**For a)** We have  $A = \{1, 3, 5\}$ .

**For b)** We have  $B = \{2, 4, 6\}$ .

**For c)** We have  $C = \{2, 3, 5\}$ .

**For d)** We have:  $A \cap B = \emptyset$ ,  $A \cap C = \{3, 5\}$ ,  $B \cap C = \{2\}$  and  $A \cap B \cap C = \emptyset$ .

## Section 2.3

# CONCEPT OF PROBABILITY FUNCTION

Below we will discuss the third element in probability space  $[\Omega, \mathcal{A}, P]$ , that is the probability measure or probability function. But one can asking: What is the probability?

### DEFINITION 2.3.1 (Probability)

Probability is numerical measure of the likelihood that a specific event will occur.

Now, to assign the probability function, several attempts were made to present the appropriate formula, and the results ranged from approximation to accuracy. But the exact condition for the probability function was in 1933 by the Russian mathematician Kolmogorov.

The first attempt to define the probability function was through the relative frequency of an event.

### DEFINITION 2.3.2 (Relative Frequency of Event)

If  $n(A)$  represents the number of times (trials)' that event  $A$  occurs among  $N$  trials of a given experiment, then  $f_A = \frac{n(A)}{N}$  represent the relative frequency of occurrence of  $A$  on these trials of the random experiment.

### RELATIVE FREQUENCY AS AN APPROXIMATION OF PROBABILITY

If an experiment is repeated  $N$  times and an even  $A$  is observed  $n(A)$  times, then according to the relative frequency concept, supposed to be the probability of  $A$  equal to:

$$P(A) = \frac{n(A)}{N}$$

► **EXAMPLE 2.3.1** Ten of the 500 randomly selected ciao manufactured at a certain auto factory are found to be defective. What is the probability that the next car manufactured at this auto factory is a defective?

**Solution:** Assume that the event under study is  $B$ , then we have  $N = 500$   $n(A) = 10$ , therefore, we can write:

$$P(\text{next car is a defective}) = P(B) = \frac{10}{500} = 0.02$$

## SECTION 2.3 CONCEPT OF PROBABILITY FUNCTION

► **EXAMPLE 2.3.2** In a random experiment of tossing a coin, the chance of head and tail are equal, thus this could be interpreted in terms of the relative frequency with which ahead is obtained on repeated tosses.

### REMARK 2.3.1

Relative frequencies are not probabilities but approximate probabilities. However, if the experiment is repeated again and again, this approximate probability of an outcomes obtained from the relative frequency will approach actual probability of that outcomes this is called the law of large numbers, and when  $n$  approaches infinity, the probability of an event  $A$  can be given by:

$$P(A) = \lim_{N \rightarrow \infty} \frac{n(A)}{N} \quad \text{if this limit exists.}$$

The next attempt to define the probability function was to take advantage of the homogeneity of the material on which the random experiment was applicate.

### CLASSICAL CONCEPT OF PROBABILITY (LAPLACE'S CONCEPT OF POSSIBILITY)

The classical probability rule is applied to compute the probabilities of events for an experiment all of whose outcomes are equally likely. According to classical probability rule, the probability of a simple event is equal to one divided by the total number of outcomes for the random experiment.

On the other hand, for a random experiment with space of elementary events  $\Omega$ , which all its elements have the same chance in appearance, the probability of a compound event  $A$  is equal to the number of outcomes favorable to event (equals to  $|A|$ ) divided by the total number of outcomes for the experiment (equals to  $|\Omega|$ ). This rule in probability calculation is known as the **Laplace Principle of Probability** (or the classical definition of probability).

$$\begin{aligned} \text{For simple event } E : \quad & P(E) = \frac{1}{|\Omega|} \\ \text{For compound event } A : \quad & P(A) = \frac{|A|}{|\Omega|} \end{aligned}$$

► **EXAMPLE 2.3.3** Calculate the probability of obtaining a head and the probability of obtaining a tail for one toss of a fair coin.

**Solution:** We have  $\Omega = \{H, T\}$ . Then the two outcomes, head and tail, are equal likely outcomes, therefore:

$$P(\{H\}) = \frac{1}{|\Omega|} = \frac{1}{2}, \quad P(\{T\}) = \frac{1}{2}$$

► **EXAMPLE 2.3.4** The probability of obtaining an even number in one roll of a fair die  $\Omega = \{1, 2, 3, 4, 5, 6\}$  and event  $A = \{2, 4, 6\}$  includes three outcomes, hence:

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

Finally, the correct definition of the probability function was through the axioms of probability space of the mathematician Kolmogorov.

#### DEFINITION 2.3.3 (Probability Measure)

Let  $\Omega$  be a space of elementary events of a random experiment, and  $\mathcal{A}$  is an  $\sigma$ -algebra on  $\Omega$ . Furthermore, we suppose that  $P$  a real set function on  $\mathcal{A}$  with the following properties:

1. We have  $P(\emptyset) = 0$ .
2. For any sequence  $A_1, A_2, \dots, A_n, \dots \in 2^\Omega$  with  $A_i \cap A_j = \emptyset$ , then:

$$P\left(\bigcup_{n=1}^{\infty} A_i\right) = \sum_{n=1}^{\infty} P(A_i)$$

3. We have  $P(\Omega) = 1$ .

Then one say that  $P$  a probability measure (or a probability function).

#### THEOREM 2.3.1

For a random experiment with space of elementary events  $\Omega$ , we have:

1. For any event  $A$  of  $2^\Omega$ , the events  $A$  and  $\emptyset$  are disjoint (mutually exclusive events).
2. For any event  $A$  of  $2^\Omega$  we have  $P(\bar{A}) = 1 - P(A)$ .
3. For any two events  $A$  and  $B$  of  $2^\Omega$  with  $A \subset B$ , then  $P(A) \leq P(B)$  (monotone property of the function  $P$ ).

#### THEOREM 2.3.2

If the space of elementary events  $\Omega$  is finite, and spouse  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ . Then we can calculate the probability of any event  $A \in 2^\Omega$  by the following relation:

$$P(A) = \sum_{i; \omega_i \in A} P(\{\omega_i\})$$

## SECTION 2.3 CONCEPT OF PROBABILITY FUNCTION

Note here that it is not necessary the elementary events to have the same probability.

### REMARK 2.3.2

In order to study the following helpful present important relations, they are the so called "**De-Morgan's laws**":

- i.  $\overline{A \cup B} = \bar{A} \cap \bar{B}$
- ii.  $\overline{A \cap B} = \bar{A} \cup \bar{B}$

### ADDITIVE RULE

For a random experiment with space of elementary events  $\Omega$ , one can prove the following statements.

1. If  $A$  and  $B$  are any two events of  $2^\Omega$ , then

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

2. If  $A$  and  $B$  are "mutually exclusive" events of  $2^\Omega$ , then:

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

► **EXAMPLE 2.3.5** Let  $A$  and  $B$  be events with  $P(A) = \frac{3}{8}$ ,  $P(B) = \frac{1}{2}$ , and  $P(A \cap B) = \frac{1}{4}$ .

Find:

- |                              |                             |                              |
|------------------------------|-----------------------------|------------------------------|
| a. $P(A \cup B)$             | b. $P(\bar{A}), P(\bar{B})$ | c. $P(\bar{A} \cap \bar{B})$ |
| d. $P(\bar{A} \cup \bar{B})$ | e. $P(A \cap \bar{B})$      | f. $P(\bar{A} \cap B)$       |

**Solution:** We have:

**For a)**  $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{8} + \frac{1}{2} - \frac{1}{4} = \frac{5}{8}$ .

**For b)**  $P(\bar{A}) = 1 - P(A) = 1 - \frac{3}{8} = \frac{5}{8}$ ,  $P(\bar{B}) = 1 - P(B) = 1 - \frac{1}{2} = \frac{1}{2}$ .

**For c)**  $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \frac{5}{8} = \frac{3}{8}$ .

**For d)**  $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - \frac{1}{4} = \frac{3}{4}$

**For e)**  $P(A \cap \bar{B}) = P(A) - P(A \cap B) = \frac{3}{8} - \frac{1}{4} = \frac{1}{8}$

**For f)**  $P(\bar{A} \cap B) = P(B) - P(A \cap B) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$ .



► **EXAMPLE 2.3.6** Three students  $A$ ,  $B$  and  $C$  are in a swimming race.  $A$  and  $B$  have the same probability of winning and each is twice as likely to win as  $C$ . Find the probability that  $B$  or  $C$  wins.

**Solution:** If we put  $P(C) = x$ , then we have  $P(A) = P(B) = 2x$ .

Because of the sum of all probabilities must equal to one, we have:

$$P(A) + P(B) + P(C) = 2x + 2x + x = 1$$

Thus,  $P(C) = \frac{1}{5}$  and  $P(A) = P(B) = \frac{2}{5}$ .

The probability that  $B$  or  $C$  wins is:

$$P(B \cup C) = P(B) + P(C) = \frac{2}{5} + \frac{1}{5} = \frac{3}{5}$$

► **EXAMPLE 2.3.7** Let a die is weighted so that the even number have the same chance of appearing, the odd number have the same chance of appearing, and each even number is twice as likely to odd appear. Then find the probability that:

- An even number appears
- An odd number appears
- A prime number appears
- An odd number occurs but not prime number.

**Solution:** We have:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\omega_1 \quad \omega_2 \quad \omega_3 \quad \omega_4 \quad \omega_5 \quad \omega_6'$$

Now we assume that:

$$A = \{2, 4, 6\}, B = \{1, 3, 5\} \text{ and } C = \{2, 3, 5\}$$

And Since  $P(\{2\}) = P(\{4\}) = P(\{6\}) = 2 P(\{1\}) = 2P(\{3\}) = 2P(\{5\})$ ,

If we put

$$P(\{i\}) = P(\{\omega_i\}) = x; \quad i = 1, 3, 5$$

Then we have:

$$P(\{j\}) = P(\{\omega_j\}) = 2x; \quad j = 2, 4, 6.$$

Because of the sum of all probabilities must equal to one, we have:

$$P(\{2\}) + P(\{4\}) + P(\{6\}) + 2 P(\{1\}) + 2P(\{3\}) + 2P(\{5\}) = 1$$

$$\Rightarrow x + x + x + 2x + 2x + 2x = 9x = 1 \Rightarrow x = \frac{1}{9}$$

## SECTION 2.3 CONCEPT OF PROBABILITY FUNCTION

and hence:

$$P(\{1\}) = P(\{3\}) = P(\{5\}) = \frac{1}{9}, \text{ and } P(\{2\}) = P(\{4\}) = P(\{6\}) = \frac{2}{9}$$

Therefore, we have:

$$\text{For a) } P(A) = \sum_{i; \omega_i \in A} P(\{\omega_i\}) = P(\{\omega_2\}) + P(\{\omega_4\}) + P(\{\omega_6\}) = \frac{6}{9}$$

$$\text{For b) } P(B) = \sum_{i; \omega_i \in B} P(\{\omega_i\}) = P(\{\omega_1\}) + P(\{\omega_3\}) + P(\{\omega_5\}) = \frac{3}{9}$$

$$\text{For c) } P(C) = \sum_{i; \omega_i \in A} P(\{\omega_i\}) = P(\{\omega_2\}) + P(\{\omega_3\}) + P(\{\omega_5\}) = \frac{4}{9}$$

$$\text{For d) } P(B \cap \bar{C}) = P(B) - P(B \cap C) = \frac{3}{9} - \frac{2}{9} = \frac{1}{9}.$$

► **EXAMPLE 2.3.8** If  $A$  and  $B$  are two events in a space of elementary events  $\Omega$ , and we suppose that  $P(A) = 0.8$ ,  $P(B) = 0.55$  and  $P(A \cup B) = 0.9$ . Then calculate the probability of:

- Occurrence of  $A$  and  $B$ ,
- Occurrence of only  $A$  and not  $B$ ,
- Non occurrence of  $A$  and  $B$ ,
- Occurrence only  $A$  or only  $B$ .

**Solution:** We have:

$$\text{For a) } P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.8 + 0.55 - 0.90 = 0.45$$

$$\text{For b) } P(A \cap \bar{B}) = P(A) - P(A \cap B) = 0.8 - 0.45 = 0.35$$

$$\text{For c) } P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.90 = 0.10$$

$$\text{For d) } P((A \cap \bar{B}) \cup (\bar{A} \cap B)) = P(A \cap \bar{B}) + P(\bar{A} \cap B) = 0.35 + 0.10 = 0.45$$

► **EXAMPLE 2.3.9** Let  $A$ ,  $B$  and  $C$  are events of  $P(A) = \frac{1}{2}$ ,  $P(B) = \frac{1}{8}$  and  $P(C) = \frac{1}{4}$ , where  $A$ ,  $B$  and  $C$  are mutually exclusive. Then calculate:

- $P(A \cup B \cup C)$
- $P(\bar{A} \cap \bar{B} \cap \bar{C})$

**Solution:** We have:

$$\text{For a)} \quad P(A \cup B \cup C) = P(A) + P(B) + P(C) = \frac{1}{2} + \frac{1}{8} + \frac{1}{4} = \frac{7}{8}.$$

$$\text{For b)} \quad P(\overline{A \cap B \cap C}) = P(\overline{A \cup B \cup C}) = 1 - P(A \cup B \cup C) = 1 - \frac{7}{8} = \frac{1}{8}.$$

► **EXAMPLE 2.3.10** A certain family owns two television sets, one color and black and white set. Let  $A$  be the event "the color set is on" and  $B$  the event "the black and white is on". If  $P(A) = 0.4$ ,  $P(B) = 0.3$  and  $P(A \cup B) = 0.5$ . Find:

- both sets are on,
- the color set is on and the other is off,
- exactly one set is on,
- Neither set is on.

**Solution:** We have:

$$\text{For a)} \quad P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.4 + 0.3 - 0.5 = 0.2$$

$$\text{For b)} \quad P(A \cap \overline{B}) = P(A) - P(A \cap B) = 0.4 - 0.2 = 0.2$$

$$\text{For c)} \quad P((A \cap \overline{B}) \cup (\overline{A} \cap B)) = P(A \cap \overline{B}) + P(\overline{A} \cap B) = 0.2 + 0.1 = 0.3$$

$$\text{For d)} \quad P(\overline{A \cap B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.5 = 0.5$$

► **EXAMPLE 2.3.11** Let  $P(A) = 0.4$  and  $P(A \cup B) = 0.6$ . Find the value of  $P(B)$  which make  $A$  and  $B$  mutually exclusive events.

**Solution:** Since  $A, B$  are mutually exclusive events, then  $P(A \cap B) = 0$  and:

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

So we become  $P(B) = 0.2$ .

## Section 2.4

# CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

Suppose that we have a random experiment with a space of elementary events  $\Omega$ .

Let  $B \in 2^\Omega$  an arbitrary event with  $P(B) > 0$ . The probability that an event  $A$  occurs given the event  $B$  occurs, written as  $P(A | B)$  is defined as:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

### REMARKS 2.4.1

1. The conditional probability  $P(A | B)$  can be read as follows: the conditional probability of  $A$  given  $B$ .
2. If  $\Omega$  is finite and all elementary events has the same chance in appearance, then we can write the previous relation as follow:

$$P(A | B) = \frac{|A \cap B|}{|B|}$$

This tells us that of the fraction of times  $B$  has occurred, what fraction of the time has  $A$  occurred. To find this, we no longer need to look for the fraction of times  $A$  occurs in our space of elementary events, but only the portion of  $A$  that is in  $B$ , i.e.  $A \cap B$ , and compare this to the fraction of times that  $B$  occurs. We notice that the conditional probabilities enjoy all the usual properties of probability on the reduced space of elementary events.

### CONDITIONAL PROBABILITIES

1. If two events  $A_1$  and  $A_2$  are mutually exclusive, then:

$$P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B)$$

This result can be extended as follows:

When we consider a sequence of mutually exclusive events  $B_1, B_2, \dots, B_n, \dots$  we have:

$$\left( \bigcup_{i=1}^{\infty} B_i \right) \cap A = \bigcup_{i=1}^{\infty} (B_i \cap A)$$

See the Venn diagram (or primary diagram – In the relation to the English Scientist John Venn (1834 – 1923)) below for illustration of finite version of the result.

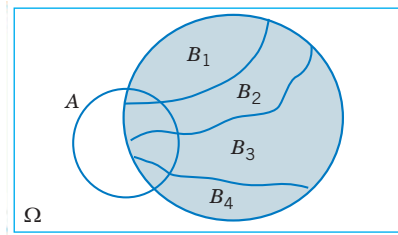


Figure 2.4.1

Note that the sequence  $B_i \cap A$ ,  $i = 1, 2, \dots$  is also sequence of disjoint events. Therefore, we get the following results:

$$P\left(\left(\bigcup_{i=1}^{\infty} B_i\right) \cap A\right) = \sum_{i=1}^{\infty} P(B_i \cap A)$$

$$P\left(\bigcup_{i=1}^{\infty} B_i \mid A\right) = \sum_{i=1}^{\infty} P(B_i \mid A)$$

2. If  $P(B) > 0$ . Then by using the formula of conditional probability we get the following result:

$$P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

3. If  $B \subseteq A$ , then we have we have  $A \cap B = B$ , therefore:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)}{P(A)} \Rightarrow P(B) = P(A)P(B \mid A)$$

4. Chain rule of probability:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \dots P(A_n \mid A_1 \cap A_2 \dots \cap A_{n-1})$$

► **EXAMPLE 2.4.1** A student is randomly selected from a class where 35% of the class is left-handed and 50% are sophomores. We further know that 5% of the class consists of left-handed sophomores. Given that a randomly selected student is a sophomore, what is the probability that he/she is left-handed?

**Solution:** We define:

$A$  = event that a randomly selected student is left-handed,

$B$  = event that a randomly selected student is a sophomore.

Then we have:

## SECTION 2.4 CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

$$P(A) = 0.35, P(B) = 0.5, P(A \cap B) = 0.05$$

What we want:  $P(A | B)$  by the definition,

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.5} = 0.1$$

► **EXAMPLE 2.4.2** A certain system can experience 2 different types of defects. Let  $A_i$ ,  $i = 1, 2$  denote the event that the system has a defect of type  $i$ . Suppose that:

$$P(A_1) = 0.15, P(A_2) = 0.10, \text{ and } P(A_1 \cap A_2) = 0.08.$$

If the system has a type 1 defect, what is the probability that it has a type 2 defect?

**Solution:** What we want to calculate:

$$P(A_2 | A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{0.08}{0.15} = 0.53$$

### MULTIPLICATION IN THE PROBABILITY CALCULATION

Suppose that we have a random experiment with a space of elementary events  $\Omega$ . And  $A, B \in 2^\Omega$  with  $P(B) > 0$ . Then by using the formula of conditional probability:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

We become the following relation:

$$P(A \cap B) = P(A)P(B | A)$$

If we have  $P(A) > 0$ . Then one can write:

$$P(A \cap B) = P(B)P(A | B)$$

Any of the above two relations is called the rule of multiplication in probability.

### REMARKS 2.4.2

The general formula for the rule of multiplication in probability as follows:

If we have  $A_1, A_2, \dots$  and  $A_n \in 2^\Omega$  with  $P(A_1 A_2 \dots A_{n-1}) > 0$ . Then we can write:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 A_2) \cdot \dots \\ \dots \cdot P(A_{n-1} | A_1 A_2 \dots A_{n-2}) \cdot P(A_n | A_1 A_2 \dots A_{n-1})$$

**INDEPENDENT EVENTS**

With conditional probability, we can see that the occurrence of one event can affect the probability of another event occurring. One question we might be interested in is whether or not this true for any two events?

To illustrate this, we will take the following example

► **EXAMPLE 2.4.3** Let's look at two events:

*A*: the event you scrape your knee on the sidewalk

*B*: the event you take a shower today

Does the fact that you took a shower today influence whether you will scrape your knee on the sidewalk? Experience would tell us that the two events are unrelated to each other. When the occurrence of one event does NOT affect the occurrence of another event, we say the two events are independent of each other.

In terms of the two events that we defined above, mathematically we can convey the idea of independent events by saying:

$$P(A | B) = P(A)$$

This last equation says that the probability of your scraping your knee is unaffected by the occurrence of taking a shower today. From our previous discussion of conditional probability, we can also rewrite the above equation as:

$$P(A \cap B) = P(A | B)P(B) = P(A)P(B)$$

This last equation gives us another way of looking at independent events. It basically says that if two events *K* and *S* are independent then the product of their individual probabilities is equal to the probability of their intersection.

Let's look at independent events with an example.

► **EXAMPLE 2.4.4** Suppose that you flip a coin twice. Let say that *A* represents the event you get a head on the first flip, *B* is the event you get a head on the second flip, and so on. What is the probability that you would get a head on the second flip given that you had a head on the first flip?

**Solution:** In terms of the events we've defined, we are asking for  $P(B | A)$ . Now looking at this problem from personal experience, we could argue that each flip is unaffected by what took place on the previous flip. In other words:

$$P(B | A) = P(B)$$

Now we can actually verify this mathematically by using the fact that two events are independent if the product of their individual probabilities is equal to the probability of their intersection. In other words:

$$P(A \cap B) = P(A)P(B)$$

Looking at the space of elementary events for this problem, we have:

$$\Omega = \{HH, HT, TH, TT\}$$

Then we have:

$$P(A \cap B) = P(\{HH\}) = \frac{1}{4}$$

Comparing this quantity with:

$$P(A)P(B) = P(\{HH, HT\})P(\{HH, TH\}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

We see that in fact  $A$  and  $B$  are independent. ◀

► **EXAMPLE 2.4.5** Let us extend Example 2.4.4 a little further and flip three coins now. What's the probability of getting a head on the third flip given that the previous two flips were heads?

**Solution:** Let say that  $C$  represents the event you get a head on the third flip, Then mathematically, we want  $P(C | (A \cap B))$ . Again, we should expect to see that getting a head on the third flip is unaffected by having two heads on the previous two flips. In other words:

$$P(C | (A \cap B)) = P(C)$$

To show this is in fact the case we need to show that:

$$P(A \cap B \cap C) = P(C)P(A \cap B)$$

From the previous calculation, we already showed that:

$$P(A \cap B) = P(A)P(B)$$

Therefore, for this calculation, we need to show:

$$P(A \cap B \cap C) = P(C)P(A \cap B) = P(A)P(B)P(C)$$

The space of elementary events for flipping a coin three times is:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

From this, we have:



$$P(A \cap B \cap C) = P(\{HHH\}) = \frac{1}{8}.$$

Likewise, we find that:

$$\begin{aligned} P(A)P(B)P(C) &= P(\{HHH, HHT, HTT\})P(\{HHH, THH, THT\})P(\{HHH, TTH, THH\}) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \end{aligned}$$

So this shows that getting a head on the third flip is unaffected by having heads on the previous two flips. In other words,  $A$ ,  $B$ , and  $C$  are all independent of each other. ◀

### MUTUALLY EXCLUSIVE AND INDEPENDENT EVENTS

It is important that you're able to distinguish between events that are mutually exclusive and events that are independent. The two are not the same. If we return to the Example 2.4.4, first introduced for taking a shower and scraping your knee, we can see the two events were independent but not mutually exclusive since it is possible for you to scrape your knee and take a shower. For two events  $A$  and  $B$ , the following table will help to distinguish between mutually exclusive and independent events:

	Probabilities	Verbal Description
<b>Mutually Exclusive</b>	$P(A \cap B) = 0$	Both events cannot happen
<b>Independent</b>	$P(A   B) = P(A)$ $P(A \cap B) = P(A)P(B)$	The occurrence of B does not affect the occurrence of A

► **EXAMPLE 2.4.6** The next table represents the condition of the dice used and the acceptability of 450 wafers. 'Particles' means that particles were on the die that produced a given wafer. A wafer can be classified as 'Good' or 'Bad.'

Quality	Condition of Die		Total
	No Particles	Particles	
Good	320	14	334
Bad	80	36	116
<b>Total</b>	<b>400</b>	<b>50</b>	<b>450</b>

- If a wafer is bad, what is the probability that it was produced from a die that had particles?
- If a wafer is good, what is the probability that it was produced from a die that had particles?

## SECTION 2.4 CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

- c. Are the two events ‘Particles’ and ‘Good’ independent?

**Solution:** Define the following events:

$Y$  “Yes, there are particles.”

$N$  “There are no particles.”

$G$  “The wafer is good.”

$B$  “The wafer is bad.”

If we use the events above, we get the next table divide by 450:

	$N$	$Y$	
$G$	0.7111	0.0311	0.7422
$B$	0.1778	0.0800	0.2578
	0.8889	0.1111	1.0000

Therefore, we have:

**For a)**  $P(Y | B) = 36/116 = 0.3103$

**For b)**  $P(Y | G) = 14/334 = 0.0419$

**For c)**  $P(N | G) = 320/334 = 0.9581$

$$P(N) = 400/450 = 0.8889$$

Since  $P(N | G) \neq P(N)$ , “a good wafer” and “a die with no particle” are not statistically independent. ◀

► **EXAMPLE 2.4.7** In the next table, 2000 community members were sampled with the following results:

Drives to Work	Homeowner	Tenant	Total
Yes	824	681	1505
No	176	319	495
<b>Total</b>	<b>1000</b>	<b>1000</b>	<b>2000</b>

Then answer the following questions:

- What is the probability that someone who drives to work is a homeowner;
- What is the probability that a homeowner drives to work;
- What is the difference between (a) and (b), and
- Are the whether driving to work and being a homeowner are independent.

**Solution:** This table was made into a joint probability table by dividing through by 2000. Use  $H$  for homeowner,  $T$  for tenant and  $D$  for “drives to work.”

	$H$	$T$	
$D$	0.4120	0.3405	0.7525
$\bar{D}$	0.0880	0.1559	0.2475
	0.5000	0.5000	1.0000

So, for example  $P(H \cap D) = 0.4120$ . This is the probability that a respondent is both a homeowner and drives to work. Therefore, we have:

**For a)**  $P(\text{a homeowner} \mid \text{drives to work}) = 824/1505 = 0.5475$  or

$$P(H \mid D) = \frac{P(H \cap D)}{P(D)} = \frac{0.4120}{0.7525} = 0.5475$$

**For b)**  $P(\text{drives to work} \mid \text{a homeowner}) = 824/1000 = 0.8240$  or

$$P(D \mid H) = \frac{P(D \cap H)}{P(H)} = \frac{0.4120}{0.5000} = 0.8240$$

**For c)** The conditional events are reversed.

**For d)** Since  $P(\text{a homeowner}) = 1000/2000 = 0.50$  is not equal to:

$$P(\text{a homeowner} \mid \text{drives to work}) = 824 / 1505 = 0.5475,$$

driving to work and whether the respondent is a homeowner or a tenant are not statistically independent. ◀

► **EXAMPLE 2.4.8** Given that a student studied, the probability of passing a certain quiz is 0.99. Given that the student did not study. The probability of passing the quiz is 0.05. Assume that the probability of studying is 0.7. A student flunks the quiz, what is the probability that he or she did not study?

**Solution:** From the given data we have

$$P(\text{pass} \mid \text{studied}) = 0.99$$

$$P(\text{pass} \mid \text{no study}) = 0.05$$

$$P(\text{study}) = 0.7$$

$$P(\text{no study} \mid \text{no pass}) = ?$$

Let  $A$  = Event of passing quiz

$B$  = Event of studying

$$P(A \mid B) = 0.99, \quad P(A \mid \bar{B}) = 0.05, \quad P(B) = 0.7 \quad P(\bar{B} \mid \bar{A})$$

## SECTION 2.4 CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

$$P(A \cap B) = P(B) P(A | B) = 0.7 \times 0.99 = 0.693$$

and

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(A) - P(A \cap B)}{P(\bar{B})} \Rightarrow 0.05 = \frac{P(A) - 0.693}{0.3}$$

Hence we become  $P(A) = 0.708$ .

And then we have:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.708 + 0.3 - 0.693 = 0.715$$

Therefore, we have:

$$\begin{aligned} P(\bar{B}|\bar{A}) &= \frac{P(\bar{B} \cap \bar{A})}{P(\bar{A})} = \frac{P(\overline{B \cup A})}{P(\bar{A})} = \frac{1 - P(A \cup B)}{P(\bar{A})} \\ &= \frac{1 - 0.715}{1 - 0.708} = \frac{0.285}{0.292} = 0.976 \end{aligned}$$

► **EXAMPLE 2.4.9** In a certain college, 0.25 of the students failed mathematics, 0.15 of the students failed chemistry and 0.10 of the students failed both mathematics and chemistry. A student is selected at random:

- If he failed chemistry, what is the probability that he failed mathematics?
- If he failed mathematics, what is the probability that he failed chemistry?
- What is the probability that he failed mathematics or chemistry?

**Solution:**

Let:

$M$  = (student who failed mathematics)

$C$  = (student who failed chemistry)

$$P(M) = 0.25, P(C) = 0.15,$$

**For a)** We have:  $P(M|C) = \frac{P(M \cap C)}{P(C)} = \frac{0.10}{0.15} = \frac{2}{3}$ .

**For b)** We have:  $P(C|M) = \frac{P(C \cap M)}{P(M)} = \frac{0.10}{0.25} = \frac{2}{5}$ .

**For c)** We have:  $P(M \cup C) = P(M) + P(C) - P(M \cap C) = 0.30$

**TOTAL PROBABILITY****DEFINITION 2.4.1**

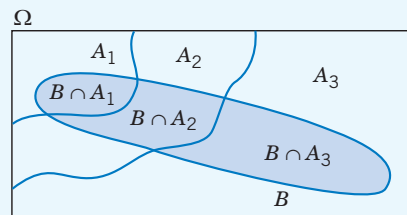
The events  $Z_1, Z_2, \dots, Z_n$  constitute a partition of the space of elementary events  $\Omega$ , if

- $Z_i \neq \emptyset$  for all  $i$
- $Z_i \cap Z_j = \emptyset$  for all  $i \neq j$
- $\bigcup_{i=1}^n Z_i = \Omega$

**THEOREM 2.4.1 (Total Probability Formula)**

If the events  $Z_1, Z_2, \dots, Z_n$  constitute a partition of the space of elementary events  $\Omega$  such that  $P(Z_k) \neq 0$  for  $k = 1, 2, \dots, n$ , then for any event  $A$ :

$$P(B) = \sum_{k=1}^n P(Z_k) P(B|Z_k)$$



**Figure 2.4.2**

The above relation is known as total probability formula.

**BAYES' THEOREM**

Bayes' theorem is a basic element of probability theory first discovered or codified by the British statistician, Thomas Bayes. At the most basic level, Bayes' theorem is an equation that relates two conditional probabilities,  $P(B | A)$  and  $P(A | B)$ .

**Importance of Bayes' Theorem**

It could rightly be said that today the field of statistics is being gradually transformed by Bayes' theorem, producing a new field of Bayesian statistics.

This revolution not only has immense potential for scientific research, but for fundamentally changing how probabilistic thinking occurs in human culture. Bayesian statistics:

- Is arguably a superior way of thinking about and using probabilities?

## SECTION 2.4 CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

- Has the potential to transform every discipline that draws inferences from uncertain evidence (medicine, law, quality analysis...)
- Helps us to see that many important inferences are only probabilistic, not certain.
- Requires us to identify and explicitly quantify the components of probabilistic inference.
- Makes every day probabilistic inferences less vague and subjective

### THEOREM 2.4.2 (Bayes' Rule)

If the events  $Z_1, Z_2, \dots, Z_n$  constitute a partition of the space of elementary events  $\Omega$  such that  $P(Z_k) \neq 0$  for  $k = 1, 2, \dots, n$ , then for any event  $b$  such that  $P(B) \neq 0$ :

$$P(Z_i | B) = \frac{P(Z_i) P(B | Z_i)}{\sum_{i=1}^n P(Z_i) P(B | Z_i)}$$

### PROOF

From the definition of conditional probability and the total probability formula we find for any natural number  $k \leq n$ :

$$P(Z_k | B) = \frac{P(B \cap Z_k)}{P(B)} = \frac{P(Z_k) \cdot P(B | Z_k)}{P(B)} = \frac{P(Z_k) \cdot P(B | Z_k)}{\sum_{i \in I} P(Z_i) \cdot P(B | Z_i)}$$

### THE DERIVATION OF BAYES' THEOREM FOR TWO EVENTS

Let  $A$  and  $B$  be two events on  $\Omega$  with  $P(B) > 0$ . Then we can write the following relation:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B | A)}{P(A)P(B | A) + P(\bar{A})P(B | \bar{A})}$$

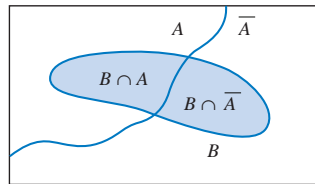


Figure 2.4.3

We want to know:

$$P(A | B) = \frac{P(B \cap A)}{P(B)}.$$

From the multiplication rule we have:

$$P(B \cap A) = P(A) P(B | A),$$

In addition, in view of the above Venn diagram we notice that:

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap \bar{A}) \\ &= P(A)P(B | A) + P(\bar{A})P(B | \bar{A}). \end{aligned}$$

Thus,

$$P(A | B) = \frac{P(B \cap A)}{P(B)} = \frac{P(A)P(B | A)}{P(A)P(B | A) + P(\bar{A})P(B | \bar{A})}$$

► **EXAMPLE 2.4.10** In a recent survey in a Statistics class, it was determined that only 60% of the students attend class on Thursday. From past data it was noted that 98% of those who went to class on Thursday pass the course, while only 20% of those who did not go to class on Thursday passed the course.

- What percentage of students is expected to pass the course?
- Given that a student passes the course, what is the probability that he/she attended classes on Thursday?

**Solution:** We defined the following events:

$A_1$ : the students attend class on Thursday.

$A_2$ : the students do not attend class on Thursday.

$B_1$ : the students pass the course.

$B_2$ : the students do not pass the course.

**For a)** From the text of the problem we have:

$$P(A_1) = 0.6, P(A_2) = 1 - P(A_1) = 0.4, P(B_1 | A_1) = 0.98, P(B_1 | A_2) = 0.2$$

And then we have:

$$\begin{aligned} P(B_1) &= P(B_1 \cap A_1) + P(B_1 \cap A_2) = P(B_1 | A_1)P(A_1) + P(B_1 | A_2)P(A_2) \\ &= 0.6 \times 0.98 + 0.4 \times 0.2 = 0.668 \end{aligned}$$

**For b)** By Bayes' theorem, we have:

$$\begin{aligned} P(A_1 | B_1) &= \frac{P(B_1 | A_1)P(A_1)}{P(B_1 | A_1)P(A_1) + P(B_1 | A_2)P(A_2)} \\ &= \frac{0.6 \times 0.98}{0.6 \times 0.98 + 0.4 \times 0.2} = 0.854 \end{aligned}$$

## SECTION 2.4 CONDITIONAL PROBABILITY AND INDEPENDENCE OF EVENT

► **EXAMPLE 2.4.11** An automobile dealer has kept records on the customers who visited his showroom. Forty percent of the people who visited his showroom were women. Furthermore, his records show that 37% of the women who visited his showroom purchased an automobile, while 21% of the men who visited his dealership purchased an automobile.

- What is the probability that a customer entering the showroom will buy an automobile?
- Suppose a customer visited the showroom and purchased a car, what is the probability that the customer was a woman?
- Suppose a customer visited the showroom but did not purchase a car, what is the probability that the customer was a man?

**Solution:** Define the following events:

$A_1$ : Customer is a woman,

$A_2$ : Customer is a man

$B$ : Customer purchases an automobile

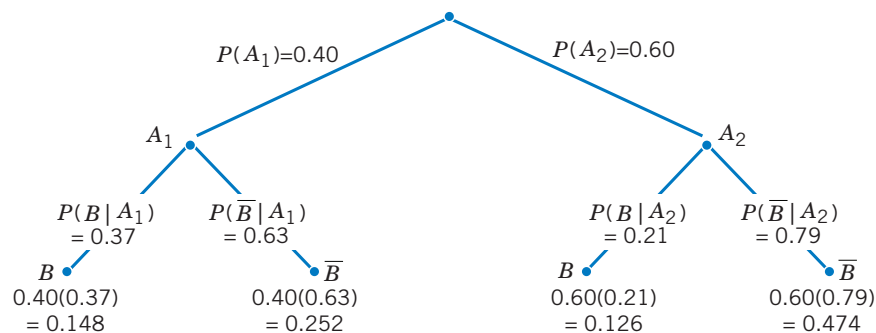
$\bar{B}$ : Customer does not purchase an automobile

Express the given information and question in probability notation:

- Forty percent of the people who visited his showroom were women:  $P(A_1) = 0.40$ .
- this statement also tells us that 60% of the customers must be men:  $P(A_2) = 0.60$ .
- 37% of the women who visited his showroom purchased an automobile:  
 $P(B | A_1) = 0.37$
- 21% of the men who visited his showroom purchased an automobile:  $P(B | A_2) = 0.21$ .

“What is the probability that a customer entering the showroom will buy an automobile:  
 $P(B) = ?$ ”

Create a tree diagram:



**Figure 2.4.4**

- Use your tree diagram and the Law of Total Probability to answer the question:

$$P(B) = 0.274$$



- b. Suppose a customer visited the showroom and purchased a car, what is the probability that the customer was a woman?

Express the question in probability notation:

We can rewrite the question as, “What is the probability that the customer was a woman, given that the customer purchased an automobile.” That is, we want to find  $P(A_1 | B)$ .

Use Bayes’ Theorem and your tree diagram to answer the question:

$$P(A_1 | B) = \frac{P(B | A_1)P(A_1)}{P(B | A_1)P(A_1) + P(B | A_2)P(A_2)} = \frac{0.148}{0.148 + 0.126} = 0.54$$

- c. Suppose a customer visited the showroom but did not purchase a car, what is the probability that the customer was a man?

Use Bayes’ Theorem and your tree diagram to answer the question:

$$P(A_2 | \bar{B}) = \frac{P(\bar{B} | A_2)P(A_2)}{P(\bar{B} | A_2)P(A_2) + P(\bar{B} | A_1)P(A_1)} = \frac{0.474}{0.474 + 0.252} = 0.653$$



1. List the space of elementary events for the outcomes for all possible sums you can get by rolling two dice by using set notation.
2. What is the space of elementary events for rolling a die and then flipping a coin? Represent this with a tree diagram, or with set notation.
3. Noura will choose either orange juice (o) or grapefruit juice (g) for breakfast. Then she will choose either eggs (e), pancakes (p), or cereal (c). Using these letter versions, create the space of elementary events for all breakfasts Belinda can have using set notation.
4. How many ways can you arrange 4 out of 7 books on a shelf?
5. How many possible different hands of 5 cards each can be dealt from a standard deck of 52 cards?
6. If a man owns 5 pairs of pants, 7 shirts, and four pairs of shoes, how many outfits can be assembled?
7. If an automobile license plate must consist of three letters followed by three single-digit numbers, how many different license plates are possible?
8. The science club has challenged the math club to a friendly competition. Each club's team should be comprised of 2 boys and 3 girls. There are 20 boys and 15 girls in the science club and 25 boys and 30 girls in the math club.
9. A manager must choose five secretaries from among 12 applicants and assign them to different stations. How many different arrangements are possible?
10. A health inspector has time to visit seven of the 20 restaurants on a list. How many different routes are?
11. Nadia is a bit forgetful, and if she doesn't make a "to do" list, the probability that she forgets something she is supposed to do is .1. Tomorrow she intends to run three errands, and she fails to write them on her list.
  - a. What is the probability that Nadia forgets all three errands?
  - b. What is the probability that Nadia remembers at least one of the three errands?
  - c. What is the probability that Nadia remembers the first errand but not the second or third?

12. New spark plugs have just been installed in a small airplane with a four-cylinder engine. For each spark plug, the probability that it is defective and will fail during its first 20 minutes of flight is  $1/10,000$ , independent of the other spark plugs.
- For any given spark plug, what is the probability that it will not fail during the first 20 minutes
  - What is the probability that none of the four spark plugs will fail during the first 20 minutes of flight?
  - What is the probability that at least one of the spark plugs will fail?
  - If a plane rental company has 25 of these small airplanes, what is the probability that at least one of the spark plugs will fail?
13. Fatin and Ali both drive older cars that don't always work in the wintertime. Suppose that Fatin's car works 80% of the time and Ali's works 65% of the time. What is the probability their cars will both be working on any given winter morning?
14. Call a household better off if its income exceeds Saudi Riyals (SR) 100,000. Call the household educated if the householder completed college. Select a Saudi household at random, and let  $A$  be the event that the selected household is better off and  $B$  the event that the household is educated. According an earlier studies,  $P(A) = 0.134$ ,  $P(B) = 0.254$ , and  $P(A \text{ and } B) = 0.080$ . What is the probability that the household selected is better off or educated,  $P(A \text{ or } B)$ ?
15. For problem 21, create a Venn diagram that shows the relation between events  $A$  and  $B$  with the respective probabilities. Describe in words the four joint events and indicate the probabilities associated with each.
16. Many fire stations handle emergency calls for medical assistance as well as calls requesting fire-fighting equipment. A particular station says that the probability that an incoming call is for medical assistance is .85.
- What is the probability that a call is not for medical assistance?
  - Assuming that successive calls are independent, what is the probability that both of two successive calls will be for medical assistance?
  - What is the probability that three consecutive calls are not for medical assistance?
  - What is the probability that of the next 10 calls, at least one is for medical assistance?

## EXERCISES

17. Hypothetical records indicate that the probability that a freshman in high school has maturity issues or family issues is 0.71. The probability that a freshman has maturity issues is 0.48. We also know that the chance of a freshman having family issues is 0.34. What is the probability that a freshman has both maturity and family issues?
18. If you eat at Star Café there's a 40% chance that your food will be cold and a 30% chance your food will taste bad. Assume that these two events are independent.
- What is the probability that both will occur, your food is cold and it tastes bad?
  - What is the probability that your food is cold or it tastes bad?
19. When spot-checked for safety, automobiles are found to defective tires 15% of the time, defective lights 25% of the time, and both defective tires and lights 8% of the time. Find the probability that a randomly chosen car has defective lights or that its tires are found to be defective.
20. A construction firm has bid on two different contracts. Let  $B_1$  be the event that the first bid is successful and  $B_2$ , that the second bid is successful. Suppose that  $P(B_1) = 0.4$ ,  $P(B_2) = 0.6$  and that the bids are independent. What is the probability that:
- Both bids are successful?
  - Neither bid is successful?
  - Is successful in at least one of the bids.
21. There are two traffic lights on the route used by Pickup Andropov to go from home to work. Let  $E$  denote the event that Pickup must stop at the first light and  $F$  in a similar manner for the second light. Suppose that  $P(E) = 0.4$ ,  $P(F) = 0.3$  and  $P(E \cap F) = 0.15$ . What is the probability that he:
- Must stop for at least one light?
  - Doesn't stop at either light?
  - Must stop at exactly one light?
  - Must stop just at the first light?

22. There are 100 students enrolled in various AP courses at American High School. There are 31 students in AP European History, 52 students in AP Calculus, and 15 students in AP Spanish. Ten students study both AP European History and AP Calculus, five students study both AP European History and AP Spanish, eight students study both AP Calculus and AP Spanish, and three students study all three. What is the probability that a student takes an AP course other than these three?
23. Data on marital status of U.S. adults can be found in Current Population Reports, a publication of the U.S. Bureau of Census. The table provides a joint probability distribution for the marital status of U.S. adults by gender.

	Never Married $M_1$	Married $M_1$	Widowed $M_3$	Divorced $M_4$	$P(GM)$
Male ( $G_1$ )	0.129	0.298	0.013	0.040	<b>0.480</b>
Female ( $G_2$ )	0.104	0.305	0.057	0.054	<b>0.520</b>
$P(MG)$	<b>0.233</b>	<b>0.603</b>	<b>0.070</b>	<b>0.095</b>	<b>1.00</b>

- Determine the probability that the adult selected is divorced.
  - Determine the probability that the adult selected is male.
  - Determine the probability that the adult selected is divorced and male.
  - Determine the probability that the adult selected is divorced, given that the adult selected is male.
  - Determine the probability that the adult selected is male, given that the adult is divorced.
24. Suppose that 23% of adults, in a particular population, smoke cigarettes. It's known that 57% of smokers and 13% of non-smokers develop a certain lung condition by the age of 60. What is the probability that a randomly selected 60-year-old, of that population, has this lung condition?



# CHAPTER 3

## RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS



### LEARNING OBJECTIVES

After completing this chapter, you should be able to:

1. Define and explain the terms discrete and continuous random, probability distributions, expectation and sample distributions.
2. Distinguish various probability distributions.
3. Understand relations among probability models. Explain what is meant by expectation and sample distributions.
4. Contrast new distributions.

### INTRODUCTION

In this Chapter we have come to know that there are mainly two common types of random variables, namely, discrete random variable and continuous random variable. Discrete random variable is a quantity assumes either a finite number of values or an infinite sequence of countable values, such as  $0, 1, 2, \dots, n$  or value  $x \in \mathbb{Q}$ . These values represent, for example, number of whole time units or number of manufactured items, the number of distance units. In the other hand, continuous random variable is a quantity assumes any numerical value in an interval or collection of intervals of  $\mathbb{R}$ , such as time, weight, distance, and temperature. Continuous random variables can take any whole values or with decimal within the interval of their possible intervals.

- SECTION 3.1 CONCEPT OF RANDOM VARIABLES AND THEIR DISTRIBUTIONS
- SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS
- SECTION 3.3 CONTINUOUS RANDOM VARIABLES AND THEIR DISTRIBUTIONS

## Section 3.1

# CONCEPT OF RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Let us first explain the concept of random variable by the following example.

A coin is tossed two times, then the space of elementary events  $\Omega$  contains four outcomes:

$$\omega_1 = HH, \omega_2 = HT, \omega_3 = TH \text{ and } \omega_4 = TT$$

Now we suppose that a real map (here it is a function)  $X$  defined on the space of elementary events  $\Omega$  as follows:

$$X : \Omega = \underbrace{\{HH\}}_{\omega_1}, \underbrace{\{HT\}}_{\omega_2}, \underbrace{\{TH\}}_{\omega_3}, \underbrace{\{TT\}}_{\omega_4} \longrightarrow \mathbb{R}$$
$$\omega \mapsto X(\omega) = \begin{cases} 0 & \text{for } \omega = TT \\ 1 & \text{for } \omega = TH, HT \\ 2 & \text{for } \omega = HH \end{cases}$$

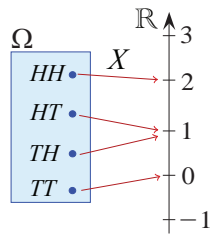


Figure 3.1.1

So we note that this map has three real values 0, 1 and 2. These are random values because the outcomes:

$$\omega_1 = HH, \omega_2 = HT, \omega_3 = TH \text{ and } \omega_4 = TT$$

are random outcomes. Therefore, we can consider the set  $\Omega^* = \{0,1,2\}$  as the results of a random experiment also. Now, if the inverse image of any event  $B$  of  $2^{\Omega^*}$  is an event belonging to  $2^{\Omega}$ , then one says that  $X$  is a random variable on  $\Omega$ .

Refer to the previous example we note that:

$$X^{-1}(\{0\}) = \{TT\} \text{ is an event of } 2^{\Omega}$$

$$X^{-1}(\{1\}) = \{HT, TH\} \text{ is an event of } 2^{\Omega}$$

$$X^{-1}(\{2\}) = \{HH\} \text{ is an event of } 2^{\Omega}$$

So the map  $X$  is a random variable on  $\Omega$ .



However, the correct definition of the random variable as follow.

**DEFINITION 3.1.1 (Random Variable)**

Let  $[\Omega, \mathcal{A}, P]$  a probability space, and  $X$  a map (or a function) defined on a space of elementary events  $\Omega$  with values in  $\mathbf{R}$ . Now if the inverse image of any event  $B$  of  $\mathfrak{R}$  is an event in  $\mathcal{A}$ . This means:

$$X^{-1}(B) \in \mathcal{A} \quad ; \forall B \in \mathfrak{R}$$

Then one say that  $X$  is a random variable on the probability space  $[\Omega, \mathcal{A}, P]$ .

**REMARKS 3.1.1**

1. Usually random variable (**r.v.**) is denoted by capital letters  $X, Y, Z, \dots$ , whereas their values are denoted by small letters such  $x, y, z, \dots$ etc.
2. In fact, checking that  $X$  is a random variable is difficult by using the definition, so the following condition can be used as an alternative to the condition in the definition:

$X$  is a random variable on probability space  $[\Omega, \mathcal{A}, P]$  if and only if:

$$\{\omega \in \Omega ; X(\omega) \leq x\} \in \mathcal{A} \quad ; \forall x \in (-\infty, +\infty) = \mathbb{R}$$

For example, if we go back to the previous example, we find:

$$\{\omega \in \Omega ; X(\omega) \leq x\} = \begin{cases} \emptyset & \text{for } x < 0 \\ \{TT\} & \text{for } 0 \leq x < 1 \\ \{TT, TH, HT\} & \text{for } 1 \leq x < 2 \\ \{TT, HT, TH, HH\} = \Omega & \text{for } x \geq 2 \end{cases}$$

But we know that  $\emptyset, \{TT\}, \{TT, TH, HT\}$  and  $\Omega$  are elements of  $\mathcal{A} = 2^\Omega$ .

Therefore, the map  $X$  is a random variable on the probability space  $[\Omega, \mathcal{A}, P]$ .

3. For a random variable  $X$  we write as a shortcut  $P(X = x), P(X < x), P(X \leq x), P(X > x), P(X \geq x)$  instead of  $P(\{\omega \in \Omega ; X(\omega) = x\}), P(\{\omega \in \Omega ; X(\omega) < x\}), P(\{\omega \in \Omega ; X(\omega) \leq x\}), P(\{\omega \in \Omega ; X(\omega) > x\}), P(\{\omega \in \Omega ; X(\omega) \geq x\})$  respectively.
4. Random variables are classified as either discrete or continuous, according to the assumed values of  $X$ . The distinguish between discrete and continuous random variables is very important in probability theory because different techniques are used to describe their distributions.

**DEFINITION 3.1.2 (Distribution Function of a r.v.)**

Let  $X$  be a Random variable on a probability space  $[\Omega, \mathcal{A}, P]$ . Now for this random variable we define a real function  $F_X$  on  $\mathbf{R}$  as follow:

$$F_X : \mathbf{R} \longrightarrow \mathbf{R} ; x \mapsto F_X(x) := P(X \leq x)$$

The function  $F_X$  is called distribution function (D.F.) of  $X$ .

**REMARKS 3.1.2**

1. For any value  $x \in \mathbf{R}$  the value  $F_X(x)$  between 0 and 1, e.g.

$$0 \leq F_X(x) \leq 1 \quad \text{for any } x \in \mathbf{R}$$

2. One can prove that:

a. The distribution function  $F_X$  is a non-decreasing function, e.g.  $F_X(x) \leq F_X(y)$  for any  $x < y$ .

b.  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  and  $\lim_{x \rightarrow \infty} F_X(x) = 1$ .

c. The distribution function  $F_X$  is right continuous function at each point  $x \in \mathbf{R}$ .

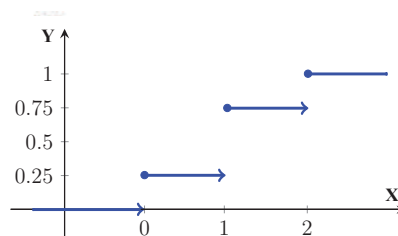
3. One can proof that  $P(a < X \leq b) = F_X(b) - F_X(a)$ .

► **EXAMPLE 3.1.1** Return to the previous example (a coin is tossing two times). If the coin is fair, then we find that the distribution function (D.F.) of  $X$  has the following form:

$$F_X(x) = P(\{\omega \in \Omega ; X(\omega) \leq x\}) = \begin{cases} P(\emptyset) = 0 & \text{for } x < 0 \\ P(\{TT\}) = 1/4 & \text{for } 0 \leq x < 1 \\ P(\{TT, TH, HT\}) = 3/4 & \text{for } 1 \leq x < 2 \\ P(\{TT, HT, TH, HH\}) = P(\Omega) = 4/4 & \text{for } x \geq 2 \end{cases}$$

This relation can be written and draw graphically as follow:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.25 & \text{for } 0 \leq x < 1 \\ 0.75 & \text{for } 1 \leq x < 2 \\ 1 & \text{for } x \geq 2 \end{cases}$$



**Figure 3.1.2** the graph of the distribution function of  $X$

## Section 3.2

# DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

In the following is a study on a special type of random variables:

### DEFINITION 3.2.1 (Discrete Random Variable)

A discrete random variable is a random variable that takes a finite or countable infinite number of values.

► **EXAMPLES 3.2.1** Note that the following statements are examples of discrete random variables:

1. Number of cars passing through a street during one hour of the day.
2. Number of phone calls received in a day by a police station.
3. Number of heads in 5 tosses of a coin.
4. Number of rainy days during a year in Riyadh.
5. Number of defective parts in a shipment.

► **EXAMPLE 3.2.2** A coin is tossed three times, the number of heads obtained can be 0, 1, 2 or 3. The probabilities of each of these possibilities or events can be tabulated as shown:

Table 3.2.1

Number of heads	0	1	2	3
The Probability of Appearing	1/8	3/8	3/8	1/8

The space of elementary events here is:

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

The corresponding random variable is discrete, since it can only take a countable number of values. In this example, the number of heads can only take 4 values (0, 1, 2, 3). ◀

### REMARKS 3.2.1

In Example 3.3, one may associate a function to represent the probability of certain event occurring. The usual notation for this function is  $p_{\bullet} := P(X = \bullet)$ . Here  $\{\omega \in \Omega ; X(\omega) = \bullet\}$  is the event that we are considering. Therefore,  $X$  observes the number of heads that we obtain in throw the coin three times. Also,  $P(X = 0)$  means

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

"the probability that no heads has appeared". Here,  $p_0 = P(X = 0) = 1/8$  (the probability that we obtain no heads is  $1/8$ ). Similarly,  $P(X = 1)$  means "the probability that one head has appeared". Here,  $p_1 = P(X = 1) = 3/8$  (the probability that we obtain one head is  $3/8$ ). These and other results are given in the following table:

**Table 3.2.2**

$x$ the values of the r.v. $X$	0	1	2	3
$P_x = P(X = x)$	1/8	3/8	3/8	1/8

### DEFINITION 3.2.2 (Probability Mass Function)

Let  $X$  be a discrete random variable on a probability space  $[\Omega, \mathcal{A}, P]$ . Then the function:

$$p_{\bullet} : \mathbb{R} \longrightarrow [0, 1] \quad ; x \mapsto p_x := P(\{\omega \in \Omega : X(\omega) = x\})$$

such that:

- i)  $p_x = P(X = x) \geq 0$
- ii)  $\sum_x p_x = \sum_x P(X = x) = 1$

is called a probability mass function, and denoted by **p.m.f.**

### REMARK 3.2.2

- For a discrete r.v.  $X$  it is sometimes  $p_{\bullet} = P(X = \bullet)$  called the probability density function (**p.d.f**) (or as a shortcut, density function (**d.f.**)).
- A discrete random variable  $X$  with values  $x_1, x_2, \dots, x_n$  (the range of  $X$  is finite set). Then can  $X$  be represented in a table. This table is called the probability distribution table for  $X$ , and it is as follows:

**Table 3.2.3**

$x_i$ the values of the r.v. $X$	$x_1$	$x_2$	.....	$x_n$
$p_i = P(X = x_i)$	$p_1$	$p_2$	.....	$p_n$

- Assuming that  $X$  is a discrete random variable on a space of elementary events  $\Omega$  with values  $x_1, x_2, \dots, x_n$ . Then can this random variable be graphically represented by drawing two orthogonal axes  $XoY$ , and then drawing a column  $p_i$  (called the random variable jump) at an altitude equal to the value  $x_i$  (called the jump position of the random variable) for all possible values for  $i$ .

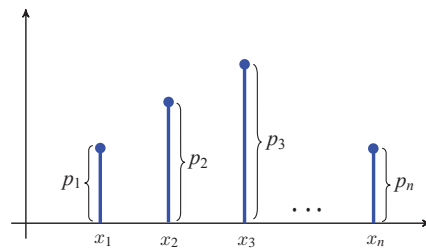


Figure 3.2.1

Note that tabular and graphical representation for a discrete random variable are useless if the number of values is large.

4. Some statistical programs provide graphical representation of discrete random variables in the form of rectangular columns above the values of this random variable as follow.

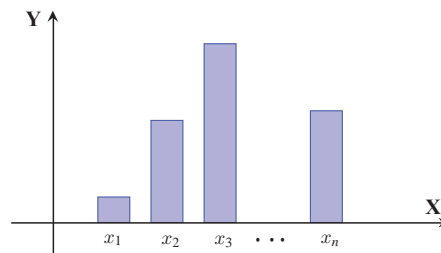


Figure 3.2.2

5. Usually, when  $x$  is an integer number, we use the letter  $k$  or  $j$  or  $i$  in the form  $P(X = x)$  instead of  $x$  as an illustration. That means we write  $P(X = k)$  or  $P(X = j)$  or  $P(X = i)$  instead of  $P(X = x)$ .

► **EXAMPLE 3.2.3** A fair die, cube with each of its faces showing different number of dots from 1 to 6, is thrown repeatedly until a 6, at the top, is obtained.

- a. Determine the **p.m.f.** for the number times one throws the die.
- b. Show that  $p_{\bullet} = P(X = \bullet)$  is a probability mass function.
- c. Calculate the probability of getting 6 in not more than two trials.
- d. Calculate the probability of getting 6 in more than two trials.

**Solution:** We have:

**For a)** Let  $X$  be the random variable representing the number of times one throws the die to get 6. Thus  $P(X = 1) = 1 / 6$ , if one only throws the die once and gets a 6. To get six in the second trial for the first time, means to get other than 6 in the first trial that is to get 1, 2, 3, 4 or 5 with probability  $5/6$ , thus  $P(X = 2) = (1 / 6) (5 / 6) = 5 / 36$ . Hence, one gets 6 in the third trial if he gets other than 6 in the first two trial and 6 in the third and so on.

Table 3.2.4

$k$ the values of the r.v. $X$	1	2	3	4	...	$i$
$p_k = P(X = k)$	1/6	5/6 <sup>2</sup>	5 <sup>2</sup> /6 <sup>3</sup>	5 <sup>3</sup> /6 <sup>4</sup>	...	5 <sup><math>i-1</math></sup> / 6 <sup><math>i</math></sup>

**For b)** To show that the probability mass function, one has to prove that it is non-negative for any value and the summation of all possible probabilities is equal to one. Note that:

$$P(X = k) = \frac{5^{k-1}}{6^k} \geq 0 \quad ; \text{for } k = 1, 2, 3, \dots$$

Also

$$\sum_{k=1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} \frac{5^{k-1}}{6^k} = \frac{1}{5} \sum_{k=1}^{\infty} \left(\frac{5}{6}\right)^k = \frac{1}{5} \left( \frac{5/6}{1 - (5/6)} \right) = 1$$

**For c)** Now if one wants to find that the probability of getting 6 in not more than two trials. This can be expressed as:

$$P(X \leq 2) = P(X = 1) + P(X = 2) = \frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} = \frac{11}{36}$$

**For d)** Similarly, the probability of getting 6 in more than 2 trials is:

$$\begin{aligned} P(X \geq 3) &= \sum_{k=3}^{\infty} P(X = k) = 1 - \sum_{k=1}^2 P(X = k) \\ &= 1 - [P(X = 1) + P(X = 2)] = 1 - \left[ \frac{1}{6} + \frac{5}{36} \right] = 1 - \frac{11}{36} = \frac{25}{36} \end{aligned}$$

### DEFINITION 3.2.3: (Distribution Function of a Discrete r.v.)

Let  $X$  be a discrete random variable on a probability space  $[\Omega, \mathcal{A}, P]$ . Then the distribution function of  $X$  (denoted by  $F_X$ ) is a real function on  $\mathbb{R}$  given as follow:

$$\begin{aligned} F_X : \mathbb{R} &\longrightarrow [0, 1] \\ x &\mapsto F_X(x) := \sum_{k=1}^x P(X = k) = \sum_{k=1}^x p_k \end{aligned}$$

### REMARKS 3.2.3:

- From the distribution function definition we observe that  $0 \leq F_X(x) \leq 1$  always.
- For each value  $x$  of the random variable  $X$ , the distribution function  $F_X$  has a jump up.
- The distribution function  $F_X$  has constant values between the values of the random variable  $X$ .

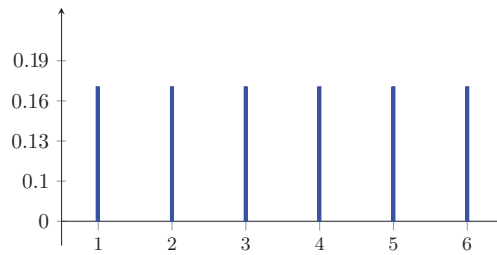
► **EXAMPLE 3.2.4** Consider a random variable  $X$ , assumes the values  $x_1, x_2, x_3, x_4, x_5, x_6$  with equal probabilities. We will represent this r.v. tabular and graphical, and then we determine the probability mass function and a distribution function.

**Solution:** We have the tabular representation of  $X$  as in the following table:

**Table 3.2.5**

$k$ the values of the r.v. $X$	1	2	3	4	5	6
$p_k = P(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The graphical representation of  $X$  is as in the following Figure.



**Figure 3.2.3**

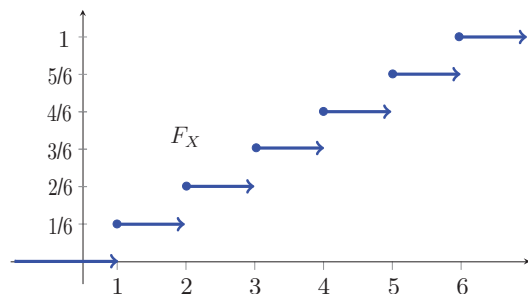
The probability mass function of this random variable is:

$$P(X = k) = \frac{1}{6} \quad \text{for any } k = 1, 2, 3, 4, 5, 6$$

The distribution function of this random variable is:

$$F_X(x) = \sum_{k=1}^x P(X = k) = \begin{cases} 0 & \text{for } x < 1 \\ \frac{1}{6} & \text{for } 1 \leq x < 2 \\ \frac{2}{6} & \text{for } 2 \leq x < 3 \\ \frac{3}{6} & \text{for } 3 \leq x < 4 \\ \frac{4}{6} & \text{for } 4 \leq x < 5 \\ \frac{5}{6} & \text{for } 5 \leq x < 6 \\ \frac{6}{6} = 1 & \text{for } x \geq 6 \end{cases}$$

Then the graph of the distribution function  $F_X$  has the following form:



**Figure 3.2.4**

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

This random variable is a famous random variable, and it is called a **discrete uniform random variable** (or uniformly distributed **r.v.**) with parameter  $n = 6$ .

### REMARK 3.2.4

One can represent the distribution definition of a discrete random variable by

**Table 3.2.6**

$k$	$x_1$	$x_2$	.....	$x_n$	.....
$F_X(x) = \sum_{k=1}^x P(X = x_k)$	$p_1$	$p_1 + p_2$	.....	$\sum_{j=1}^n p_j$	.....

► **EXAMPLE 3.2.5** Consider tossing four fair coins. Let  $X$  a random variable observes the number of heads on all four coins. Now

- Determine the probability space  $[\Omega, \mathcal{A}, P]$  for the experiment.
- What are the possible values for  $X$ ?
- Is the random variable  $X$  continuous or discrete?
- Construct the probability distribution table for this experiment.
- Give the tabular and graphical representation of  $X$ .

**Solution:** We have:

**For a)** The probability space  $[\Omega, \mathcal{A}, P]$  for this experiment is:

$$\Omega = \left\{ \begin{array}{l} \overbrace{HHHH}^{\omega_1}, \overbrace{HHHT}^{\omega_2}, HHTH, HTHH, THHH, HHTT, HTHT, THHT, \\ HTTH, TTHH, THTH, HTTT, THTT, TTHT, \underbrace{TTTH}_{\omega_{15}}, \underbrace{TTTT}_{\omega_{16}} \end{array} \right\}$$

$$\mathcal{A} = 2^\Omega \quad \text{and} \quad P(A) = \sum_{\omega; \omega \in A} P(\{\omega\}) \quad ; \forall A \in \mathcal{A}$$

**For b)** The possible values for  $X$  are  $k = 0, 1, 2, 3, 4$ .

**For c)** This random variable  $X$  is discrete since it takes a finite countable number of values.

**For d)** The tabular representation of  $X$  and its probability distribution function is given in the following table:

**Table 3.2.7-a**

$k$	0	1	2	3	4
$P_k = P(X = k)$	$1/2^4$	$4/2^4$	$6/2^4$	$4/2^4$	$1/2^4$
$F_X(x) = \sum_{k=1}^x P(X = x_k)$	$1/2^4$	$5/2^4$	$11/2^4$	$15/2^4$	1

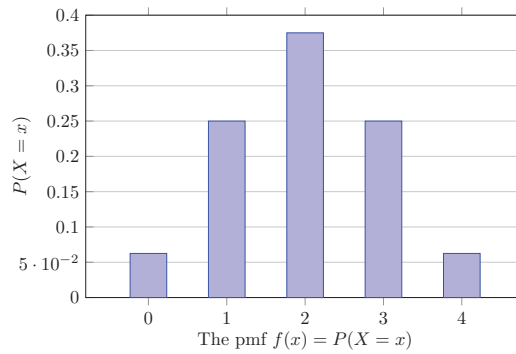


Or by the following table:

**Table 3.2.7-b**

$k$	0	1	2	3	4
$P_k = P(X = k)$	0.0625	0.25	0.375	0.25	0.0625
$F_X(x) = \sum_{k=1}^x P(X = x_k)$	0.0625	0.3125	0.6875	0.9375	1

The graphical representation of this random variable  $X$  is as follow:



**Figure 3.2.5**

► **EXAMPLE 3.2.6** Consider tossing a fair coins four times. Let  $X$  a random variable observes the number of heads on all four coins. Now

- a. Determine the space of elementary events for this experiment.
- b. what do you note?

**Solution:** We have:

**For a)** The space of elementary events for this experiment is

$$\Omega = \left\{ \begin{array}{l} \overbrace{HHHH}^{\omega_1}, \overbrace{HHHT}^{\omega_2}, HHHT, HHTH, HTHH, THHH, HHHT, HTHT, THHT, \\ HTTH, TTHH, THTH, HTTT, THTT, TTHT, \underbrace{TTTH}_{\omega_{15}}, \underbrace{TTTT}_{\omega_{16}} \end{array} \right\}$$

We note that this random variable is discrete with values  $k = 0, 1, 2, 3, 4$ . And the probability distribution table for this experiment is the same as in the table 3.7-a.

**For b)** We note that the results of this example does not change from the results of the previous example. ◀

► **EXAMPLE 3.2.7** It was found that, in maternity hospital in Saudi Arabia, the number of twin births is approximately 1 in 95. Let  $X$  be the number of births in that hospital until the

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

first twins are born. Determine the probability mass function and the distribution function of the r.v.  $X$ .

**Solution:** Denote twin births by  $T$  and single births by  $\Omega$ . Then  $X$  is a real-valued function defined on the space of elementary events  $\Omega = \{T, ST, SST, SSST, \dots\}$ .

Now the  $j^{\text{th}}$  twin birth means

$$X(\underbrace{SSS\dots ST}_{j-1}) = j.$$

Thus the twin birth can occur on the first birth or the second birth or the third birth, and so on, this leads to the possible values of  $X$  as  $\{1, 2, 3, \dots\}$ . Therefore, we have:

$$P(X = k) = P(\underbrace{SSS\dots ST}_{k-1}) = \left(\frac{94}{95}\right)^{j-1} \left(\frac{1}{95}\right).$$

**Table 3.2.8**

$k$ the values of the r.v. $X$	1	2	3	...	$k$	...
$P_k = P(X = k)$	$\left(\frac{1}{95}\right)$	$\left(\frac{94}{95}\right)\left(\frac{1}{95}\right)$	$\left(\frac{94}{95}\right)^2\left(\frac{1}{95}\right)$	...	$\left(\frac{94}{95}\right)^{j-1}\left(\frac{1}{95}\right)$	...
$F_X(x) = \sum_{k=1}^x p_k$	$\left(\frac{1}{95}\right)$	$\left(\frac{1}{95}\right) + \left(\frac{94}{95}\right)\left(\frac{1}{95}\right)$	$\left(\frac{1}{95}\right) + \left(\frac{94}{95}\right)\left(\frac{1}{95}\right) + \left(\frac{94}{95}\right)^2\left(\frac{1}{95}\right)$	...	$\sum_{i=1}^j \left(\frac{94}{95}\right)^{i-1} \left(\frac{1}{95}\right)$	...

It is noted that:

$$P(X = k) = \left(\frac{94}{95}\right)^{k-1} \left(\frac{1}{95}\right) \geq 0 \quad \text{for all } i = 2, 3, 4, \dots, j, \dots, \dots$$

$$\sum_{k=1}^{\infty} p_k = \sum_{k=1}^{\infty} \left(\frac{94}{95}\right)^{k-1} \left(\frac{1}{95}\right) = \frac{1}{95} \sum_{k=0}^{\infty} \left(\frac{94}{95}\right)^k = \left(\frac{1}{95}\right) \left[1 / \left(1 - \frac{94}{95}\right)\right] = 1$$

This sum is for a geometric series with base  $\left(\frac{1}{95}\right)$  and progression  $\left(\frac{94}{95}\right)$ .

### EXPECTATION, MEAN AND VARIANCE for D.R.V.

Mathematical expectation and variance of a random variable are two special cases of the moments of random variables.

**DEFINITIONS 3.2.4 (Moments of a Discrete r.v.):**

Let  $X$  be a discrete random variable on a probability space  $[\Omega, \mathcal{A}, P]$  with values set  $X = \{x_i : i \in I\}$ . Then the moment of order  $k$  (denoted by  $E(X^k)$ ) of this random variable given by:

$$E(X^k) = \sum_{i \in I} x_i^k P(X = x_i)$$

**Special Cases**

- If  $k = 1$ , then  $E(X)$  is called the expected value of  $X$  (and can be denoted by  $\mu$  also). That means the expected value of  $X$  is:

$$\mu = E(X) = \sum_{i \in I} x_i P(X = x_i)$$

- If  $k = 2$ , then  $E[X - E(X)]^2$  is called the variance of this random variable  $X$  (denoted by  $\sigma^2$  or  $Var(X)$ ). That means the variance of this random variable  $X$ :

$$\sigma^2 = Var(X) = E[X - E(X)]^2 = \sum_{i \in I} (x_i - \mu)^2 P(X = x_i)$$

**DEFINITIONS 3.2.5 (Factorial Moments of a Discrete r.v.):**

Let  $X$  be a discrete random variable on a probability space  $[\Omega, \mathcal{A}, P]$  with values set  $X = \{x_i : i \in I\}$ . Then the factorial moment of order  $k$  (we denoted it by  $\mathcal{F}_k(X)$ ) of this random variable given by:

$$\begin{aligned} \mathcal{F}_k(X) &= E[X(X-1)(X-2)\dots(X-k+1)] \\ &= \sum_{i \in I} [x_i(x_i-1)(x_i-2)\dots(x_i-k+1)] P(X = x_i) \end{aligned}$$

**REMARK 3.2.5**

We know that the positive square root of the variance is called the standard deviation ( $sd$ ) (denoted by  $\sigma$ ), thus we have:

$$\sigma = +\sqrt{Var(X)}$$

**PROPERTIES OF EXPECTED VALUE AND VARIANCE**

Let  $a$  and  $b$  be constants and  $X$  is a discrete random variable, then:

1.  $E(aX + b) = aE(X) + b$
2.  $Var(X) = E(X^2) - \mu^2$

SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

- 3.  $Var(a) = 0$  for any constant  $a$
- 4.  $Var(aX + b) = a^2Var(X)$

Using these properties one can get:

$$\begin{aligned} Var(X) &= E\left[(X - E(X))^2\right] = E\left[X^2 - 2X E(X) + (E(X))^2\right] \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 = E(X^2) - 2E(X) \cdot E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

This last relation is known as the **Steiner Formula**.

► **EXAMPLE 3.2.8** Calculate the mean, variance and standard deviation of the random variable in Example 3.2.5.

**Solution:** The probability mass function in Example 3.2.5 is given in the table:

**Table 3.2.9-a**

$k$	0	1	2	3	4
$P_k = P(X = k)$	$1/2^4$	$4/2^4$	$6/2^4$	$4/2^4$	$1/2^4$

Now we compute  $\mu = E(X)$  and  $\mu = E(X^2)$  in the following table:

**Table 3.2.9-b**

$k$	0	1	2	3	4
$P_k = P(X = k)$	$1/2^4$	$4/2^4$	$6/2^4$	$4/2^4$	$1/2^4$
$\mu = E(X)$	0	$4/2^4$	$12/2^4$	$12/2^4$	$4/2^4$
$E(X^2)$	0	$4/2^4$	$24/2^4$	$36/2^4$	$16/2^4$

Hence the mean of this distribution is:

$$\begin{aligned} \mu = E(X) &= \sum_{i=1}^{\infty} x_i P(X = x_i) = \sum_{k=1}^{\infty} k P(X = k) \\ &= 0 \times (1/2^4) + 1 \times (4/2^4) + 2 \times (6/2^4) + 3 \times (4/2^4) + 4 \times (1/2^4) = \frac{32}{16} = 2 \end{aligned}$$

where the expected value of  $X^2$  (is called second moment of  $X$ ) is

$$\begin{aligned} E(X^2) &= \sum_{i=1}^{\infty} x_i^2 P(X = x_i) = \sum_{k=1}^{\infty} k^2 P(X = k) \\ &= 0^2 \times (1/2^4) + 1^2 \times (4/2^4) + 2^2 \times (6/2^4) + 3^2 \times (4/2^4) + 4^2 \times (1/2^4) = \frac{80}{16} = 5 \end{aligned}$$

Now the variance of  $X$  is:

$$\sigma^2 = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2 = 5 - 2^2 = 1$$

The standard deviation for this random variable is  $\sigma = +\sqrt{1} = 1$ .

► **EXAMPLE 3.2.9** Let  $f$  be a probability mass function for a discrete random variable  $X$ , it takes integer numbers, and defined by the following relation:

$$f(k) = \begin{cases} 2c & \text{for } k = 10 \\ c & \text{for } k = 20 \\ c - 0.2 & \text{for } k = 30, \\ 0 & \text{otherwise} \end{cases}$$

where  $c$  is a real constant. Then:

- Determine the constant  $c$ .
- Determine the corresponding distribution function.
- Calculate the mean, the variance and standard deviation for  $X$ .
- Calculate  $E(5X + 9)$  and  $Var(3X + 14)$

**Solution:** We have

**For a)** According to the definition of a discrete random variable with density function  $f$ , we can write:

$$\begin{aligned} 1 &= \sum_{k \in \mathbb{Z}} f(k) = \sum_{k \in \mathbb{Z}} P(X = k) \\ &= P(X = 10) + P(X = 20) + P(X = 30) + P(X = k \in \mathbb{Z}; k \neq 10, 20, 30) \\ &= 2c + c + c - 0.2 + 0 \end{aligned}$$

Thus, we find  $c = 0.3$ .

**For b)** The corresponding distribution function is:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for } x < 10 \\ 0.6 & \text{for } 10 \leq x < 20 \\ 0.9 & \text{for } 20 \leq x < 30 \\ 1 & \text{for } x \geq 30 \end{cases}$$

**For c)** The mean of this random variable is:

$$\begin{aligned} \mu = E(X) &= \sum_x x \cdot P(X = x) \\ &= 10P(X = 10) + 20P(X = 20) + 30P(X = 30) \\ &= 10 \cdot 0.6 + 20 \cdot 0.3 + 30 \cdot 0.1 = 15 \end{aligned}$$

Where the variance can be obtained by  $E(X^2)$  (the moment of order 2 of  $X$ ) as follows:

$$\begin{aligned} E(X^2) &= \sum_x x^2 \cdot P(X = x) \\ &= 10^2 P(X = 10) + 20^2 P(X = 20) + 30^2 P(X = 30) \\ &= (100 \times 0.6) + (400 \times 0.3) + (900 \times 0.1) = 270 \end{aligned}$$

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Hence the variance is:

$$\sigma^2 = \text{Var}(X) = E(X^2) - \mu^2 = 270 - 225 = 45$$

Hence the standard deviation is:

$$\sigma = +\sqrt{\text{Var}(X)} = \sqrt{45} = 6.71$$

Now, by using the properties of the expectation, one has:

$$E(5X + 9) = 5 E(X) + 9 = (5 \times 15) + 9 = 84$$

**For d)** Also,

$$\text{Var}(3X + 14) = 3^2 \text{Var}(X) + \text{Var}(14) = (9 \times 45) + 0 = 405$$

► **EXAMPLE 3.2.10** Let  $X$  be a discrete random variable representing the number of hours' university student spending on practicing sports per week. The probability mass function for  $X$  is given by the form:

$$P(X = k) = \frac{k^3}{9c}, \quad k = 1, 2, 3, \text{ where } c \text{ is a constant.}$$

- Determine the value of  $c$ .
- Determine the distribution function of the random variable  $X$ .
- Calculate the mean and the variance.

**Solution:** We have:

**For a)** The value of  $c$  is the quantity that make  $\sum_k P(X = k) = 1$ , hence:

$$\begin{aligned} 1 &= \sum_k P(X = k) = P(X = 1) + P(X = 2) + P(X = 3) \\ &= \frac{1^3}{9c} + \frac{2^3}{9c} + \frac{3^3}{9c} = \frac{36}{9c} = \frac{4}{c} \end{aligned}$$

Thus  $c = 4$ .

**For b)** The corresponding distribution function is:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{for } x < 1 \\ \frac{1}{36} & \text{for } 1 \leq x < 2 \\ \frac{9}{36} & \text{for } 2 \leq x < 3 \\ \frac{9 + 27}{36} = 1 & \text{for } x \geq 3 \end{cases}$$

**For c)** The mean for this discrete random variable is:

$$\begin{aligned}
 \mu = E(X) &= \sum_{i=1}^{\infty} x_i P(X = x_i) = \sum_{k=1}^3 k P(X = k) \\
 &= 1P(X = 1) + 2P(X = 2) + 3P(X = 3) \\
 &= 1 \times \frac{1}{36} + 2 \times \frac{8}{36} + 3 \times \frac{27}{36} = \frac{98}{36} = \frac{49}{18} = 2.7\bar{2}
 \end{aligned}$$

and for the variance, one find:

$$\begin{aligned}
 E(X^2) &= \sum_{i=1}^{\infty} x_i^2 P(X = x_i) = \sum_{k=1}^3 k^2 P(X = k) = 1^2 P(X = 1) + 2^2 P(X = 2) + 3^2 P(X = 3) \\
 &= \frac{1}{36} + 4 \cdot \frac{8}{36} + 9 \cdot \frac{27}{36} = \frac{276}{36} = \frac{46}{6} = 7.667
 \end{aligned}$$

Hence the variance is:

$$\sigma_X^2 = Var(X) = E(X^2) - \mu^2 = \frac{46}{6} - \left(\frac{49}{18}\right)^2 = 0.2561$$

► **EXAMPLE 3.2.11** Suppose an insurance company pays the amount of Saudi Riyals, SR 2000 for lost luggage on an airplane trip. From past experience, it is known that the company pays this amount in 1 out of 400 insurance policies it sells. What premium should the insurance company charge to every policy to break even?

**Solution:** We define the r.v.  $X$  as follows:

$X = 0$  if no loss occurs, which happens with probability  $1 - (1/400) = 0.9975$

$X = -2000$  with probability  $\frac{1}{400} = 0.0025$ .

Now the table representation for this discrete random variable is:

**Table 3.2.10**

$k$	0	-2000	Total
$p_k = P(X = k)$	0.9975	0.0025	1

The expected loss to the company is:

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i) = \sum_{k=1}^2 k P(X = k) = 0 \times 0.995 + (-2000 \times 0.0025) = -5 \text{ SR}$$

Thus, the company must charge 5 SR to break even. Note that the insurance can make profit only if it sells any policy more than 5 SR.

This random variable is one of the famous random variables, we will present it in the next paragraph.

## SOME SPECIAL DISTRIBUTIONS

Below we will present some of the famous discrete probability distributions

## THE BINOMIAL DISTRIBUTION

There are some discrete distributions occur with such regularity in real-life situations that they have got their own names and it is worth studying their properties. One of these well-known distributions is the binomial.

The binomial distribution applies in many real life situations where an experiment is performed and the outcomes are only two possibilities with fixed probabilities. The experiment is repeated  $n$  times and one is interested the number of either possibility to occur. In brief the binomial distribution arises in the following situation:

- a. The outcome of any trial can only take on two possible values, say success and failure. This type of experiment is called **Bernoulli experiments**.
- b. There is a constant probability  $1 > p > 0$  of success on each trial, hence the probability of failure is  $q = 1 - p$ .
- c. The experiment is repeated  $n$  times.
- d. The trials are assumed to be statistically independent.

Now note that  $x$  equals the number of successes in the  $n$  independent trials.

► **EXAMPLE 3.2.12** The following represents binomial random variables:

- a. The number of good edited transcripts out of  $n$  transcripts that is either in compliance with procedures or it is not.
- b. The number of correct guesses at 30 true-false questions when you randomly guess all answers
- c. The number of heads out of  $n$  times flipping a coin that lands on head with probability  $p$  and lands on tail with probability  $q = 1 - p$ .

**DEFINITION 3.2.6 (Binomial Distribution)**

Let  $X$  be a discrete random variable on a probability space  $[\Omega, \mathcal{A}, P]$ . Then one says that  $X$  has a **binomial distribution** with parameters  $n \in \mathbf{N}$  and  $1 > p > 0$ , if the probability mass function given by:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, 2, \dots, n$$



**REMARKS 3.2.7**

1. We denote the binomial distribution with parameters  $n$  and  $p$ , by  $B(x; n, p)$ .
2. Let  $X$  be a binomial distributed **r.v.** with parameters  $n$  and  $p$ . Then the distribution function of  $X$  given by:

$$F(x) = P(X \leq x) = \sum_{k=0}^x P(X = k) = \sum_{k=0}^x \binom{n}{k} p^k (1-p)^{n-k}$$

3. According to Newton's binomial expansion we note that:

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1$$

**The Mean and The Variance of The Binomial Distribution**

The mean of the binomial distribution is derived as follows:

$$\begin{aligned} \mu = E(X) &= \sum_{k=0}^n k P(X = k) = \sum_{k=0}^n k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= n p \sum_{i=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\ &= n p (1-p) \sum_{k=1}^{n-1} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k-1} \\ &= n p (1-p) \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = n p [p + (1-p)]^{n-1} = n p \end{aligned}$$

To obtain the variance, we calculate the factorial moment of order 2, so we find:

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^n k(k-1) P(X = k) = \sum_{k=2}^n k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= n(n-1) p^2 \underbrace{\sum_{k=2}^{n-2} \frac{(n-2)!}{(k-2)! [(n-2)-(k-2)]!} p^{k-2} (1-p)^{(n-2)-(k-2)}}_{=1} \\ &= n(n-1) p^2 \end{aligned}$$

Thus, the variance of the binomial is (by using **Steiner** formula):

$$\begin{aligned} \sigma^2 &= E(X^2) - (E(X))^2 = E[X(X-1)] + E(X) - (E(X))^2 \\ &= n(n-1)p^2 + n p - (n p)^2 = n p (1-p) \end{aligned}$$

Therefore, the mean and the variance of binomial distribution  $B(x; n, p)$  is given by:

$$E(X) = n p \quad \text{and} \quad \text{Var}(X) = n p (1-p)$$

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

► **EXAMPLE 3.2.13** If the mean and the variance of a binomial distribution are 16 and 8 respectively, then:

- Determine the probability mass function.
- Calculate the probability  $P(X = 0)$ .
- Calculate the probability  $P(X \geq 2)$ .

**Solution:** We have:

**For a)** Note that:

$$\text{Mean} = np = 16 \Rightarrow n = \frac{16}{p}$$

$$\text{Variance} = np(1-p) = 8 \Rightarrow \frac{16}{p}p(1-p) = 8 \Rightarrow p = \frac{1}{2}$$

And from there we find  $n = 32$

The probability mass function is:

$$P(X = k) = \frac{32!}{k!(32-k)!} (1/2)^k (1/2)^{32-k}$$

**For b)** We have:

$$P(X = 0) = \binom{32}{0} (1/2)^0 (1/2)^{32-0} = (1/2)^{32}$$

**For c)** We have:

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) \\ &= 1 - (1/2)^{32} - 32(1/2)(1/2)^{32-1} \\ &= 1 - (1/2)^{32} - 32(1/2)^{32} = 1 - 33(1/2)^{32} \end{aligned}$$

In order to simplify understanding the graph of the binomial random variable, we give the following example.

► **EXAMPLE 3.2.14** Let  $X$  be a binomial distributed **r.v.** with parameters:

- $n = 12$ ,  $p = 1/2$
- $n = 8$ ,  $p = 1/3$

We will draw the graph for  $X$  for the two cases.

**For a)** Note that for  $n = 12$ ,  $p = 1/2$  we have for any  $k = 0, 1, \dots, 12$  the following:

$$P(X = k) = \binom{12}{k} (1/2)^k (1/2)^{12-k} = \frac{12!}{k!(12-k)!} (1/2)^k (1/2)^{12-k}$$

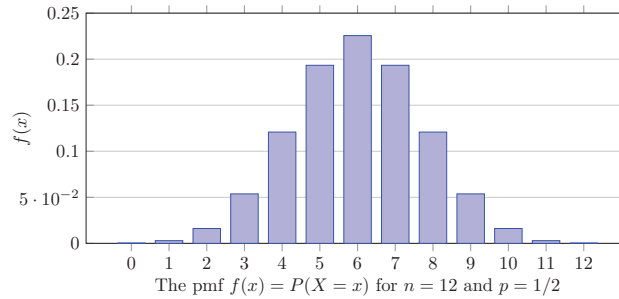


Figure 3.2.6

We note that the **p.m.f.** is symmetric around the mean.

Whereas the distribution function of  $X$  is:

$$F_X(x) = P(X \leq x) = \sum_{k=0}^x \binom{12}{k} (1/2)^k (1/2)^{12-k} \quad ; x \in \mathbb{R}$$

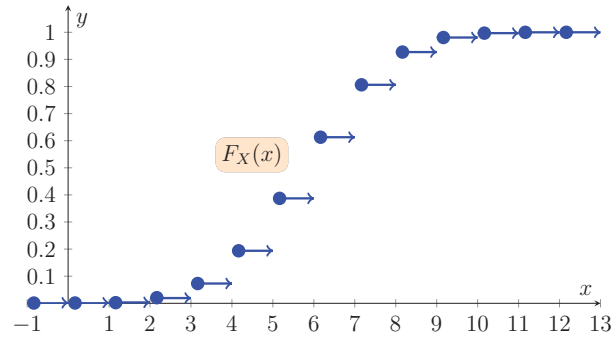


Figure 3.2.7

Note that the distribution function is a non-decreasing step function.

**For b)** Now for  $n = 8$ ,  $p = 1/3$  one has any  $k = 0, 1, \dots, 8$ :

$$P(X = k) = \binom{8}{k} (1/3)^k (2/3)^{8-k} = \frac{8!}{k!(8-k)!} (1/3)^k (2/3)^{8-k}$$

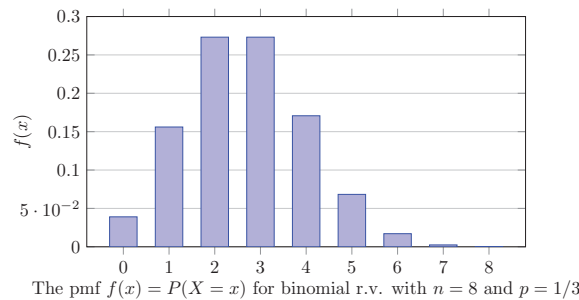


Figure 3.2.8

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Note here the binomial is slightly skewed toward the left side of the first half of the random variable values.

Whereas the distribution function of  $X$  is

$$F_X(x) = P(X \leq x) = \sum_{k=0}^x \binom{8}{k} (1/3)^k (2/3)^{8-k} \quad ; x \in \mathbb{R}$$

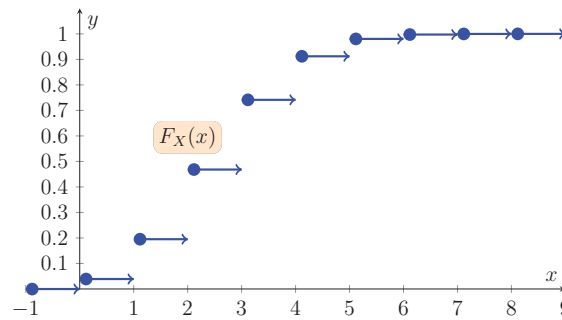


Figure 3.2.9

Where we note that the distribution function  $F_X$  is a step function. ◀

### THE POISSON DISTRIBUTION

The Poisson distribution is a discrete probability distribution. It is usually that represents the probability of a given number of phenomenon occurring in a fixed interval of time. They occur with a known average rate and are independently of each other. This distribution might also be used for the number of phenomenon in other specified intervals such as distance, area or volume. The name of the distribution refers to The French Mathematician Siméon Poisson (1781-1840).

One can use the Poisson distribution if the following assumptions have been fulfilled:

1. The phenomenon occurs one-at-a-time, not simultaneously, or in different sub-areas of observation.
2. The probability of a phenomenon occurrence is constant.
3. The probability of each phenomenon is independent of all other phenomenon.

► **EXAMPLE 3.2.15** The following situations represent Poisson random variables:

- a. Number of phone calls arriving at a call center within a time unit.
- b. Number of customers arriving at airline ticket purchasing office during one day.
- c. Number of cars arriving at a traffic light during a rush hour.
- d. Number of patients arriving to a medical clinic during a day. ◀

### The Form of Poisson Distribution

The Poisson probability distribution is determined by only one parameter  $\lambda > 0$ , this parameter represents the average number of phenomenon occurring per time unit.

#### DEFINITION 3.2.7 (Poisson Distribution)

Let  $X$  be a discrete random variable on a on a probability space  $[\Omega, \mathcal{A}, P]$ . Then one says that  $X$  has a **Poisson distribution** with parameter  $\lambda > 0$ , if the probability mass function given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad ; k = 0, 1, 2, \dots$$

It is obvious that the probability mass function of the Poisson distribution satisfies the following:

If we know that for  $\lambda > 0$  we have (the series)  $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^\lambda$ . Then we have:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \left[ 1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \dots \right] = e^{-\lambda} e^\lambda = 1$$

This is because we have Taylor expansion of the function:

$$f(x) = e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots$$

### The Mean and The Variance of The Poisson Distribution

Let  $X$  be a Poisson distributed random variable with parameter  $\lambda > 0$ , then the mean of the Poisson distribution is derived as follows:

$$\mu = E(X) = \sum_{k=1}^{\infty} k P(X = k) = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda$$

To obtain the variance, one finds:

$$\begin{aligned} E(X(X-1)) &= \sum_{k=1}^{\infty} k(k-1) P(X = k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k(k-1)\lambda^k}{k!} \\ &= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda} \lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2 \end{aligned}$$

as the variance of the binomial is:

$$\sigma^2 = E(X^2) - [E(X)]^2 = E(X(X-1)) + E(X) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

## SECTION 3.2 DISCRETE RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Therefore, the mean and the variance of Poisson distribution with parameter  $\lambda > 0$  is given by

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

► **EXAMPLE 3.2.16** In a large oil exploration firm company, engineering accidents occur independently at the mean of three per month  $\lambda = 3$ . The probability mass function is:

$$P(X = k) = \frac{3^k e^{-3}}{k!} \quad ; k = 0, 1, 2, \dots$$

Thus the instance probability of occurrence of 4 accidents is:

$$P(X = 4) = \frac{3^4 e^{-3}}{4!} = 0.168$$

► **EXAMPLE 3.2.17** Suppose the number of flaws in a 100-foot roll of paper is a Poisson random variable with  $\lambda = 10$ .

- Calculate the probability that there are eight flaws in a 100-foot roll.
- Calculate the mean and the variance of the random variable.

**Solution:** We have:

**For a)** The required probability:

$$P(X = 8) = \frac{10^8 e^{-10}}{8!} = \frac{(100\,000\,000)(2.71828)^{-10}}{40320} = 0.1126$$

**For b)** It is known that  $E(X) = \lambda = 10$  and  $\text{Var}(X) = \lambda = 10$ .

## Section 3.3

# CONTINUOUS RANDOM VARIABLES AND THEIR DISTRIBUTIONS

### The Concept of Continuous Random Variable

The following is a study of another type of random variables. These random variables are characterized by continuous distribution functions on  $\mathbb{R}$ .

#### DEFINITION 3.3.1 (Continuous Random Variable)

A continuous random variable is a random variable whose set of its possible values is uncountable set (some interval(s) of the real numbers).

► **EXAMPLE 3.3.1** The following are some examples of continuous random variables:

1. The variable that measures the life length of an electric lamp.
2. The variable that measures temperature in a city.
3. The variable that measures blood pressure of a person.
4. The variable that measures electric voltage in an electrical circuit.

All these variables can take any value of interval(s) of real numbers, in spite of the fact that they are sometimes approximated by discrete values. ◀

#### THEOREM 3.3.1

A random variable  $X$  is a continuous random variable on a probability space  $[\Omega, \mathcal{A}, P]$  if exist a real non-negative function  $f \geq 0$  on  $\mathbb{R}$ , for it the following relation is realized on any interval  $(a, b) \subseteq \mathbb{R}$ :

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

#### REMARK 3.3.1

Since each continuous random variable  $X$  has a function  $f$  (if exist) with previous properties, then denoted this function by  $f_X$ , and is called the probability density function (**p.d.f.**) of  $X$ . Therefore, the previous relation is written as follows:

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

**DEFINITION 3.3.2 (Distribution Function of a Continuous Random Variable)**

Let  $X$  be a continuous random variable on a probability space  $[\Omega, \mathcal{A}, P]$  with density function  $f_X$ . Then the **distribution function** (DF) of  $X$  (denoted by  $F_X$ ) is real function defined on  $\mathbb{R}$  by the following relation:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt \quad ; \forall x \in \mathbb{R}$$

**REMARKS 3.3.2**

1. By using properties of  $F_X$  we see that the density function  $f_X$  has the following property:

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

Because we have:

$$1 = \lim_{x \rightarrow +\infty} F_X(x) = \lim_{x \rightarrow +\infty} \int_{-\infty}^x f_X(t) dt = \int_{-\infty}^{+\infty} f_X(x) dx$$

This relation means that the area under its curve of  $f_X$  and for all values of  $x$  equal to one.

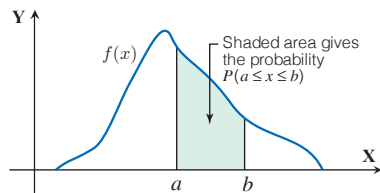


Figure 3.3.1-a

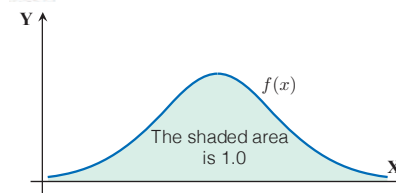


Figure 3.3.1-b

2. One can prove that  $P(X = x) = 0$  for any  $x \in \mathbb{R}$ . This property is equivalent to the following statement:

*The distribution function  $F_X$  is continuous on  $\mathbb{R}$ .*

3. It is easy to verify that  $P(X > x) = 1 - F_X(x)$ .
4. It is easy to verify that:

$$f_X(x) = \frac{d}{dx} F_X(x) = F_X'(x)$$

5. Because of the previous property (2) we find that for any continuous random variable we have:

$$P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)$$



► **EXAMPLE 3.3.2** Let  $X$  be a continuous random variable with distribution function  $F_X$ . Then we will find the pdf of  $X$ .

- a. If the distribution function  $F_X$  has the form  $F_X(x) = \frac{x^4}{16} \quad ; 0 < x < 2$
- b. If the distribution function  $F_X$  has the form  $F_X(x) = 1 - e^{-5x} \quad ; x \geq 0$

**Solution:** It was given that the relation of  $F_X$  and its pdf  $f_X$ , assumed to be exist, is given by

$$f_X(x) = \frac{d}{dx} F_X(x) = F'_X(x)$$

Thus we have:

**For a)** We have for any  $2 > x > 0$ :

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \left( \frac{x^4}{16} \right) = \frac{x^3}{4}$$

**For b)** We have for any  $x \geq 0$ :

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} (1 - e^{-5x}) = 5e^{-5x}$$

► **EXAMPLE 3.3.3** Let  $X$  be a continuous random variable with density function  $f_X$  given by the following relation:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

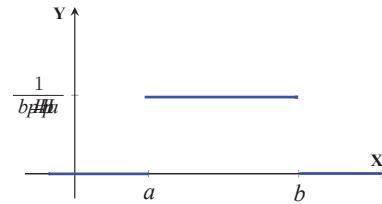


Figure 3.3.2

This is known by the **p.d.f** of a **uniform random variable** defined on  $[a, b]$ . The distribution function for this **r.v.**  $X$  is:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^a \underbrace{f_X(t)}_0 dt + \int_a^x f_X(t) dt = \begin{cases} \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

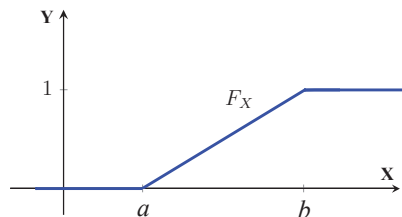


Figure 3.3.3

**EXPECTATION, MEAN AND VARIANCE for C.R.V.**

We have studied earlier that the mean and the variance of a discrete random variable. Now we will study the mean and the variance of a continuous random variable.

Let  $X$  be a continuous random variable on a probability space  $[\Omega, \mathcal{A}, P]$  with density function  $f_X$ . Then the expected value or mean is defined by:

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$

Whereas, the variance of  $X$  is given by:

$$\sigma^2 = Var(X) = E(X - E(X))^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx$$

Using the Steiner formula, we get:

$$\sigma^2 = Var(X) = \int_{-\infty}^{+\infty} x^2 f_X(x) dx + \left[ \int_{-\infty}^{+\infty} x f_X(x) dx \right]^2$$

► **EXAMPLE 3.3.4** Refer to the Example 3.3.3 we will calculate the mean, the variance and the standard deviation of the uniform distribution.

**Solution:** Using the pdf of uniform distribution, we find that the mean for  $X$  is:

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \left( \frac{x^2}{2} \Big|_a^b \right) = \frac{a+b}{2}$$

The corresponding variance for this variable is obtained by using Steiner formula, therefore, we must calculate the moment of order 2 of  $X$ .

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \frac{1}{b-a} \int_a^b x^2 dx \\ &= \frac{b^3 - a^3}{3(b-a)} = \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} = \frac{(b^2 + ab + a^2)}{3} \end{aligned}$$

Therefore, the variance of  $X$  is:

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{b^2 + ab + a^2}{3} - \frac{b^2 - 2ab + a^2}{4} = \frac{(b-a)^2}{12}$$

The standard deviation for this variable is:

$$\sigma = \sqrt{Var(X)} = \frac{b-a}{\sqrt{12}}$$

► **EXAMPLE 3.3.5** Let  $X$  be a continuous random variable with distribution function  $F_X$  given by:

$$F_X(x) = 1 - e^{-x}(1+x) \quad ; x \geq 0$$

We will determine the density function of  $X$ , and then we calculate  $P(1 \leq X < 2)$ .

**Solution:** It has been dated that  $f_X(x) = \frac{d}{dx} F_X(x)$ . Therefore, we have:

$$f_X(x) = \frac{d}{dx} [1 - e^{-x}(1+x)] = e^{-x}(1+x) - e^{-x} = x e^{-x}$$

Now to calculate  $P(1 \leq X < 2)$  we have:

$$\begin{aligned} P(1 \leq X < 2) &= F_X(2) - F_X(1) = [1 - e^{-2}(1+2)] - [1 - e^{-1}(1+1)] \\ &= 2e^{-1} - 3e^{-2} = 0.735 - 0.406 = 0.329 \end{aligned}$$

► **EXAMPLE 3.3.6** Let  $X$  be a continuous random variable with density function  $f_X$  given by:

$$f_X(x) = 2x \quad ; 0 \leq x \leq 1$$

Now we will to answer the following:

- Prove that  $f_X$  is a density function.
- Calculate the probability  $P(-1/2 \leq X < 1/2)$ .
- Determine the distribution function  $F_X$ .
- Calculate the mean, the variance and the standard deviation of  $X$ .
- Draw the graph of the of the density function  $f_X$ .

**Solution:** We have:

**For a)** We always assume that  $f_X(x) = 0$  for all values of  $X$  not mentioned in its definition, i.e.  $f_X(x) = 0$  for  $x < 0$  and  $x > 1$ .

Now for  $0 \leq x \leq 1$  we note that  $f_X(x) > 0$ . And more so we have:

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^0 0 dx + \int_0^1 2x dx + \int_1^{\infty} 0 dx = 0 + \left( x^2 \Big|_0^1 \right) + 0 = 1$$

Therefore,  $f_X$  satisfies the properties of the density function.

**For b)** To find  $P(-1/2 \leq X < 1/2)$ , we have:

$$P\left(-\frac{1}{2} < X < \frac{1}{2}\right) = \int_{-1/2}^{1/2} f_X(x) dx = \int_{-1/2}^0 0 dx + \int_0^{1/2} 2x dx = 0 + \left( x^2 \Big|_0^{1/2} \right) = \frac{1}{4}$$

**For c)** For the distribution function we have  $F_X(x) = 0$  for  $x < 0$

$$\text{For } 0 \leq x < 1 \text{ we have } F_X(x) = \frac{1}{2}(x)(2x) = x^2$$

$$\text{Whereas, for } x \geq 1 \text{ we have } F_X(x) = \frac{1}{2}(1)(2) = 1$$

Thus we get:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^2 & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

**For d)** Now the mean of  $X$  is

$$E(X) = \int_0^1 x f_X(x) dx = \int_0^1 x(2x) dx = \int_0^1 2x^2 dx = \left( \frac{2}{3} - 0 \right) = \frac{2}{3}$$

The moment of order 2 for  $X$  is:

$$E(X^2) = \int_0^1 x^2 f_X(x) dx = \int_0^1 x^2(2x) dx = \int_0^1 2x^3 dx = 2 \left( \frac{1^4}{4} - 0 \right) = \frac{2}{4} = \frac{1}{2}$$

Therefore, by using Steiner formula we find the variance equal to:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \left( \frac{1}{2} \right) - \left( \frac{2}{3} \right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} = 0.056$$

The standard deviation is:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{0.056} = 0.24$$

**For e)** The graph of the density function  $f_X(x) = 2x$  for  $0 \leq x \leq 1$  as in in the following Figure.

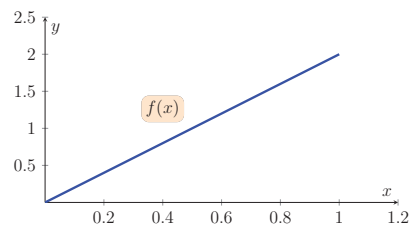


Figure 3.3.4

## SOME SPECIAL DISTRIBUTIONS

Below we will present some of the famous continuous probability distributions

**THE EXPONENTIAL DISTRIBUTION**

The exponential distribution arises in many probability and statistical applications rates (failure analysis). Some applications are worth mentioning:

1. Time between arrivals of cars at bridge or traffic light.
2. Times between failures of mechanical system service.
3. Life lengths of some electronic devices.

The exponential distribution is related to Poisson distribution, for instance, if the Poisson distribution represents the number of events that occur in a specified time period, then the exponential distribution represents times between events occurring, or time until the next event.

The exponential distribution is determined by a single parameter,  $\lambda$ .

**DEFINITION 3.3.3 (Exponential Distributed Random Variable)**

A continuous random variable  $X$  on a probability space  $[\Omega, \mathcal{A}, P]$  is said to have an **exponential distribution** with parameter  $\lambda > 0$  (or it is exponentially distributed) if its density function is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Usually the parameter  $\lambda$  represents the number of arrival or/events in a given unit of time.

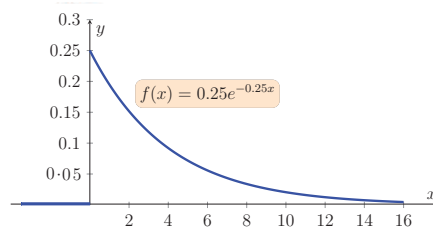
Hence  $f_X$  is the probability density function for the length of time between events.

**DEFINITION 3.3.4 (Distribution Function Of An Exponential Random Variable)**

Let  $X$  be an exponential random variable with parameter  $\lambda > 0$ . Then the distribution function of  $X$  given by the following relation:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

One may draw, for example, the graph of the density and the distribution function for the exponential distribution with mean  $\mu = 4$  or  $\lambda = 0.25$  are as follows:



**Figure 3.3.5**

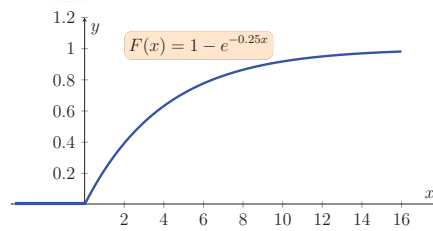


Figure 3.3.6

### Mean, Variance And Standard Deviation of Exponential R.V.

Let  $X$  be an exponential random variable with parameter  $\lambda > 0$ . Then the mean (or the average time between events) of  $X$  is:

$$E(X) = \int_{-\infty}^0 x f_X(x) dx + \int_0^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

To evaluate the variance, we first calculate  $E(X^2)$ , that is:

$$E(X^2) = \int_{-\infty}^0 x^2 f_X(x) dx + \int_0^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

Thus the variance and the standard deviation are:

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

And

$$\sigma = +\sqrt{Var(X)} = \frac{1}{\lambda}$$

► **EXAMPLE 3.3.7** If the life length of a refrigerator follows the exponential distribution, and let  $X$  represents the life length of a refrigerator. Suppose the average life length for this type of refrigerator is 15 years. Answer the following:

- What is the probability that a refrigerator can be used for less than 6 years?
- What is the probability that this refrigerator can be used for more than 18 years?
- What is the variance and the standard deviation of this random variable?

**Solution:** The random variable  $X$  has an exponential distribution with mean  $\mu = \frac{1}{\lambda} = \frac{1}{15}$ .

Thus the corresponding pdf of the life length of these refrigerators is:

$$f_X(x) = \frac{1}{15} e^{-\frac{x}{15}} \quad \text{for } x \geq 0$$

**For a)** The probability to calculate is:

$$P(X \leq 6) = 1 - e^{\frac{-6}{15}} = 0.3297$$

**For b)** The probability to calculate is:

$$P(X \geq 18) = e^{\frac{-18}{15}} = 0.3012$$

**For c)** The variance of this random variable is:

$$\text{Var}(X) = \frac{1}{\lambda^2} = \frac{1}{15^2} = 0.0044$$

And from there we find, the standard deviation is:

$$\sigma = \frac{1}{\lambda} = \frac{1}{15} = 0.067$$

### THE NORMAL DISTRIBUTION

In this part of the book we presents the normal distribution and direct use of  $Z$ -Table for calculating the probabilities of the normal distribution, the normal distribution or sometimes known by the Gaussian distribution is a continuous probability distribution that frequently occurs in natural phenomena and in many real life situations. Importance of normal distribution also comes from the fact that random errors are often following a normal distribution.

Normally distributed represents, for instance, many natural variables and phenomena such as;

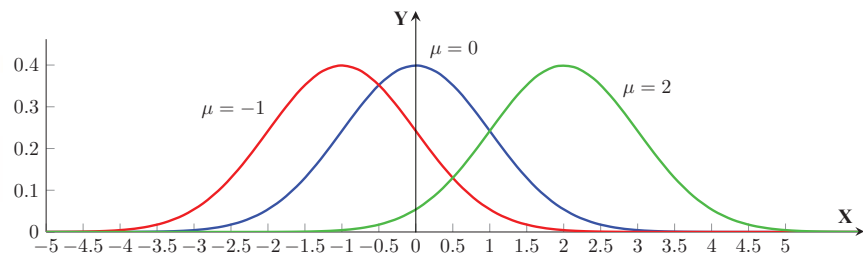
1. Human height, weights or age.
2. Human intelligence (IQ) in most communities.
3. Strength of a steel girder or steel bars.
4. No. of defective parts in a batch of manufactured items.

#### DEFINITION 3.3.5 (Normal Distributed Random Variable)

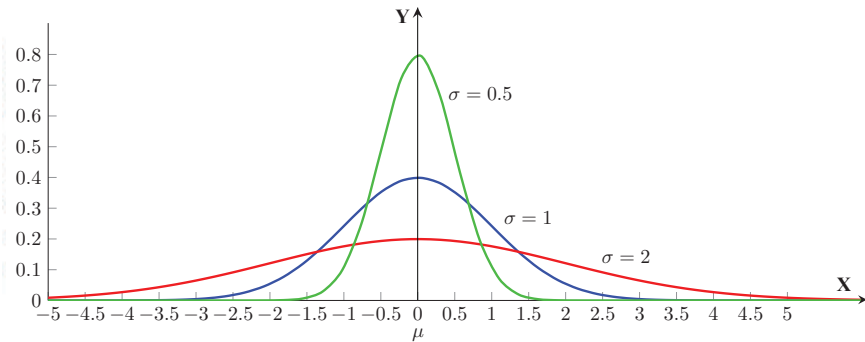
A continuous random variable  $X$  on a probability space  $[\Omega, \mathcal{A}, P]$  is said to have a **normal distribution** with parameters  $\mu \in \mathbb{R}$  (location parameter) and  $\sigma > 0$  (scale parameter) if its density function is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; x \in (-\infty, \infty)$$

The graph of the density and the distribution function for the normal distribution with several values for mean  $\mu$  and standard deviation  $\sigma$  are as follows:



**Figure 3.3.7** (The density function for a normal distribution for  $\sigma = 1$ )

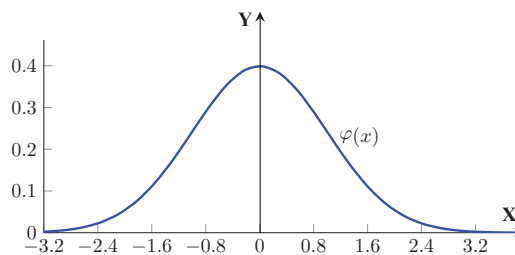


**Figure 3.3.8** (The density function for a normal distribution for  $\mu = 0$ )

### REMARKS 3.3.3

1. The normal pdf curve, for any values of the mean and the variance is bell-shaped.
2. The pdf has a single peak at the exact center of the pdf curve.
3. The mean, median, and mode of the distribution are equal and located at the mean.
4. Normal distribution can have zero mean and a unit standard deviation. In this case we have the so called "standard normal distribution". The pdf of the standard normal distribution (denoted by  $\varphi(x)$ ) has the following relation and graph:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad ; x \in (-\infty, \infty)$$



**Figure 3.3.9**

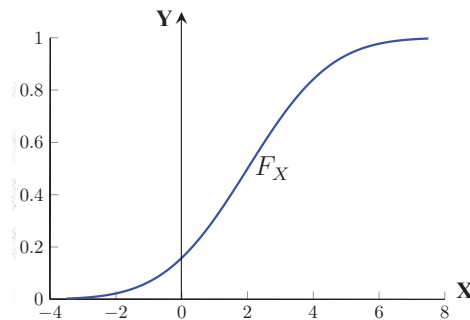


**DEFINITION 3.3.6 (Distribution Function of A Normal Random Variable)**

Let  $X$  be a normal random variable with parameters  $\mu \in \mathbb{R}$  and  $\sigma > 0$ . Then the distribution function of  $X$  given by the following relation:

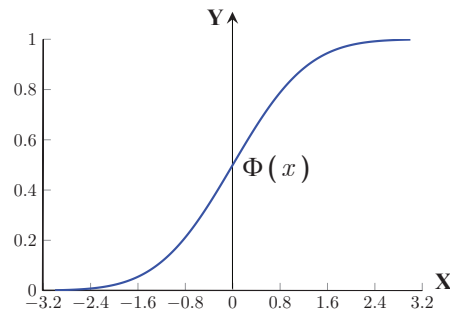
$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \quad ; x \in \mathbb{R}$$

The graph of the normal distribution function for  $\mu = 2$  and  $\sigma = 2$  as follow:



**Figure 3.3.10** (the normal distribution function for  $\mu = 2$  and  $\sigma = 2$ )

For the special case  $\mu = 0$  and  $\sigma = 1$  (standard normal distribution) one denoted the distribution function by  $\Phi(x)$ , and the graph for  $\Phi(x)$  is as follow:



**Figure 3.3.11** (The standard normal distribution function)

### STANDARDIZING NORMALLY DISTRIBUTED RANDOM VARIABLE

In fact, standardizing normal distribution simplifies the computation of areas under various regions of the curve. Such calculation of areas, which correspond to various probabilities of interest, is an essential task in solving many statistical problems.

Let  $X$  be a normal random variable with parameters  $\mu \in \mathbb{R}$  and  $\sigma > 0$ . Then one can transform this distribution to a standard normal distribution by sing the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$

## SECTION 3.3 CONTINUOUS RANDOM VARIABLES AND THEIR DISTRIBUTIONS

By standardizing a normal distribution, we eliminate the need to consider  $\mu$  and  $\sigma$ . Hence we have a standard frame of reference for all calculation of probabilities.

Note that the area under the pdf curve before and after the  $y$ -axis equals 0.5. The area under the curve of  $\varphi(x)$  is tabulated for different values of  $z$  ( $z$  is the value of  $Z$ ) and known by the normal table or the  $z$ -table, or the integration from  $-\infty$  to any value  $z$ .

The standard normal distribution function  $\Phi(z)$  is given by:

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{t^2}{2}} dt \quad ; z \in \mathbb{R}$$

This value  $\Phi(z)$  represents the area under the normal pdf curve until  $z$ , and is represented by the graph in Figure 3.22.

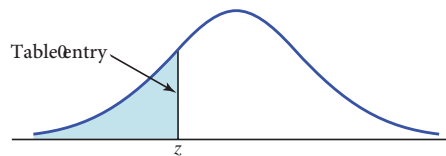


Figure 3.3.12

► **EXAMPLE 3.3.8** Assume that the student's scores in the General Aptitude Tests (GAT) of the National center for Assessment in Higher Education (NCAHE) of Saudi Arabia follow normal distribution with mean = 80 and standard deviation = 5.

- What proportion of GAT scores falls below 75?
- What proportion of GAT scores falls between 76 and 82?

**Solution:** We have:

**For a)** By standardizing the normal **r.v.**  $X$  we get:

$$\begin{aligned} P(X < 75) &= P\left(\frac{X - \mu}{\sigma} < \frac{75 - \mu}{\sigma}\right) = P\left(\frac{X - 80}{5} < \frac{75 - 80}{5}\right) \\ &= P\left(\frac{X - 80}{5} < \frac{75 - 80}{5}\right) = P\left(\frac{X - 80}{5} < \frac{75 - 80}{5}\right) \\ &= P(Z < -1) = \Phi(-1) = 0.1587 \end{aligned}$$

This value is obtained from the standard normal table or the  $z$ -table for  $z = -1$ , that is the second column adjacent to the  $z = -1$  value. This mean that 15.87 (of the scores lies below 75).

**For b)** By standardizing the normal r.v.  $X$  we get:

$$\begin{aligned} P(76 < X < 82) &= P\left(\underbrace{\frac{76-80}{5}}_{z_1 = -0.8} < \underbrace{\frac{X-\mu}{\sigma}}_Z < \underbrace{\frac{82-80}{5}}_{z_2 = 0.4}\right) = P(-0.8 < Z < 0.4) \\ &= P(Z < 0.4) - P(Z < -0.8) = \Phi(0.4) - \Phi(-0.8) \\ &= 0.6554 - 0.2119 = 0.4435 \end{aligned}$$

This is obtained by using the  $z$ -table for responding to the  $z = 0.4$  and  $z = -0.8$  respectively. Thus, the percentage of scores between 76 and 82 is 44.35. ◀

► **EXAMPLE 3.3.9** Use the standard normal table to find the  $z$ -value for the following probabilities:

- a.  $P(Z \leq z) = 0.4090$
- b.  $P(Z \leq z) = 0.80$
- c.  $P(Z > z) = 0.4090$

**Solution:** This example means finding the  $p^{\text{th}}$  percentile or the  $z$ -value that covers  $p$  of the area under the normal curve.

**For a)** We find for the  $z$ -value (or the percentile corresponds to the probability 0.4090) that  $P(Z \leq z) = 0.4090$ . We look for this probability in the  $z$ -table or the standard normal table and we note that this probability is; less than 0.5, so  $z$  has to be in the left half of the normal curve and it has to have negative value. This probability, in fact, lies in the fifth column where the  $z$ -value in the same row equals -0.2 and the fifth value of in the column gives the fraction part of the  $z$ -value that is 0.03. Thus  $z = -0.23$ .

**For b)** The  $z$ -value for  $P(Z \leq z) = 0.80$  means the value of  $z$  that keeps 0.80 of the normal curve behind it. We note that the nearest value in the table is larger value, in this case, 0.8023 which has  $z = 0.8$  of the first column and on the seventh column of the same  $z$ -row the value of  $z = 0.85$ . Thus  $z = 0.85$ .

**For c)** Also, for  $P(Z > z) = 0.4090$  one has:

$$P(Z \leq z) = 1 - P(Z > z) = 1 - 0.4090 = 0.5910$$

Thus the value is in the second half and hence  $z = 0.23$  ◀

► **EXAMPLE 3.3.10** It was found that the grades given for an airline company customer in each flight are normally distributed with mean 80 and standard deviation 10. The company considers that it is providing unsatisfactory service in any flight if it gets less than 65. Find the probability that it provides an unsatisfactory service in any flight.

**Solution:** Note that:

$$P(x < 65) = P\left(z = \frac{X - \mu}{\sigma} < \frac{65 - 80}{10} = -1.5\right)$$

From the table this corresponds to the probability of 0.0668.



## EXERCISES



1. Let  $X$  be a discrete **r.v.** representing the sum of the two numbers on throwing two balanced dice
  - a. Find the possible values of the **r.v.**  $X$ .
  - b. What is the probability mass function  $P(X = \bullet)$ ?
  - c. Calculate the cumulative distribution function  $F(x)$ .
  - d. Calculate the mean and variance for the **r.v.**  $X$ .
  
2. Let  $X$  be a discrete **r.v.** with probability mass function  $P(X = k) = c \frac{k}{7}$ ,  $k = 2, 3, 4, 5$ ,  
Then:
  - a. Determine the value of the constant  $c$  that make  $f$  probability density function.
  - b. Determine the distribution function of  $X$ .
  - c. Calculate the mean and variance of  $X$ .
  - d. Calculate  $E(X - 3)$ .
  
3. Consider rolling a balanced die twice and let the **r.v.**  $X$  be the maximum of the two numbers obtained. Then:
  - a. Determine the probability mass function and distribution function of  $X$
  - b. Sketch the functions in part (i)
  - c. Calculate the mean, the variance, and the standard deviation of  $X$ .
  
4. A discrete **r.v.**  $X$  has the following probability mass function:

$x$	-2	-1	0	1	2
$P_x = P(X = x)$	0.20	0.15	0.15	0.1	0.4

- a. Determine the cumulative distribution  $F_X$  of  $X$ .
  - b. Draw the density and distribution for this variable.
  - c. Calculate the mean, variance and standard deviation for the **r.v.**  $X$ .
  - d. Calculate  $E(3X + 2)$ ,  $Var(3X - a)$  for any real number  $a$ .
5. Consider the **Bernoulli** random variable with parameter  $1 > p > 0$ :

$$f(x) = \begin{cases} p & \text{for } x = 1 \\ q & \text{for } x = 0 \end{cases}$$

Then calculate the mean, variance and standard deviation of the **r.v.**  $X$ .

6. A box contains 10 equally likely chips numbered from 1 to 10. Let **r.v.**  $X$  be the total of two chips drawn at random and without replacement. Determine the probability mass function and distribution function of their total?

7. Let  $X$  be **r.v.** with density function (**p.m.f.**):

$$f(x) = c \left( \frac{4}{9} \right)^x \quad ; x = 1, 2, 3, \dots$$

Then:

- Determine the value of the constant  $c$  that make  $f$  probability density function.
  - Determine the distribution function of  $X$ .
  - Calculate the mean and variance of  $X$ .
8. A fair coin is tossed five times. Let  $X$  be a **r.v.** representing the number of heads. Then:
- Determine the set of all possible values of the random variable  $X$ .
  - Determine The probability mass function
  - Draw the graph of the probability mass function
  - Determine the distribution function  $F_X$ .
  - Calculate the mean and variance for this random variable.
9. Let  $X$  be **r.v.** with density function (**p.m.f.**)  $f$  given by:

$$f(x) = \begin{cases} 0, & x < 4, \\ 0.1, & x = 4 \\ 0.3, & x = 5 \\ 0.3 & x = 6 \\ 0.2, & x = 8 \\ 0.1, & x = 9 \\ 0, & x > 9 \end{cases}$$

Then:

- Draw the graph of  $f$ .
  - Calculate the probabilities  $P(X \leq 6.5)$ ,  $P(X > 8.1)$ ,  $P(5 < X < 8)$ .
  - Determine the distribution  $F_X$
  - Draw the graph of the distribution  $F_X$
  - Calculate the mean.
  - Calculate the variance.
  - If we define new **r.v.**  $W = 7 - 4X$ , then calculate  $E(W)$  and  $Var(W)$ .
10. The distribution function of a discrete **r.v.**  $X$  is given by

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ x / 4 & \text{for } 0 \leq x < 1 \\ 1 / 2 & \text{for } 1 \leq x < 2 \\ (x / 12) + (1 / 2) & \text{for } 2 \leq x < 3 \\ 1 & \text{for } x \geq 3 \end{cases}$$

Then:

- a. Determine the probability mass function  $P(X = \bullet)$
  - b. Calculate  $P(X < 2)$ ,  $P(X = 2)$ ,  $P(X = 3)$  and  $P(X = 2.5)$
  - c. Calculate the mean and variance of this **r.v.**
- 11.** Assume that the probability a baby born is a girl in a maternity hospital, is 0.51 and let the **r.v.**  $X$  be the number of births until the first boy is born. Then:
- a. Derive the probability mass function and the distribution function of  $X$ .
  - b. What is the probability that the third births are boys?
- 12.** Suppose that  $P(X = n) = \frac{1}{n(n+1)}$  for any finite integer  $n \geq 1$ . Then determine the distribution function  $F_X$  and calculate the mean  $E(X)$ .

- 13.** Let  $X$  be **r.v.** with the discrete distribution function:

$$F_X(x) = 1 - p^x \quad ; x = 0, 1, 2, \dots,$$

- a. Then determine the probability mass function  $P(X = \bullet)$ .
  - b. Calculate the mean, variance and the standard deviation of  $X$ .
- 14.** Consider the **Hyper Geometric** random variable with parameter  $N, M$  and  $n$  :

$$P(X = x) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} ; k = 0, 1, \dots, N;$$

and realize  $k \in \{\max(0, n + k - N), \dots, \min(n, k)\}$

where  $n = 1, 2, \dots, N$  and  $M = 0, 1, \dots, N ; N \in \mathbb{N}$ . Then calculate the mean, variance of the **r.v.**

- 15.** Let  $X$  be a **r.v.** representing head in an experiment of tossing fair coin three times.
- a. Find the possible values of  $X$  and the corresponding probabilities.
  - b. Derive the distribution of discrete random variable.
  - c. Calculate the mean of  $X$ .
  - d. Calculate the variance and the standard deviation of  $X$ .

16. A 3-digit integer is randomly selected from 000 to 999, inclusively. Let  $X$  be the integer selected on a particular day.
- Determine the **p.d.f.** of  $X$ .
  - Calculate the mean of  $X$ .
  - Calculate the variance of  $X$ .

17. Let  $X$  be a discrete uniform distributed random variable with the following density function

$$P(X = x) = 0.2 \text{ for } x = 5, 6, 7, 8, 9.$$

- Show that  $X$  has a true discrete density function.
  - Calculate the mean of  $X$ .
  - Calculate the variance and the standard deviation of  $X$ .
18. Let  $X$  be a **r.v.**  $X$  have a **Poisson** distribution with parameter  $\lambda > 0$ . Then:
- Calculate the mean.
  - Calculate the variance and standard deviation of  $X$ .

19. Consider a **r.v.**  $X$  with density function:

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then:

- Determine the distribution function  $F_X$ .
  - Draw the graph of density and the distribution for the **r.v.**  $X$ .
  - Calculate the mean and variance for this random variable.
20. Let  $X$  be a continuous random variable with **p.d.f.**

$$f_X(x) = \begin{cases} \frac{1}{2} + \frac{1}{4}(x-3) & \text{for } 1 \leq x < 3, \\ \frac{1}{2} - \frac{1}{4}(x-3) & \text{for } 3 < x \leq 5, \\ 0, & \text{otherwise} \end{cases}$$

Then:

- Draw the graph of **p.d.f.**
  - Determine the distribution function of  $X$ .
  - Calculate the mean and variance of  $X$ .
21. The life length of and an electronic device (in hour) is represented by the **r.v.**  $X$ . Previous experience has shown that:



$$P(X > x) = \left(1 - \frac{t}{99}\right) e^{-s/99} \quad ; x > 0.$$

Then:

- Determine the distribution function and probability density function of  $X$ .
- Calculate  $P(99 < X < 198)$ .

- 22.** Let the time for a student to finish the aptitude test of NCAHE (in hours) is a continuous r.v.  $X$  with

$$f_X(x) = \begin{cases} k(x-1)(2-x) & \text{for } 1 \leq x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Then:

- Calculate the value of the constant  $k$ .
- Derive the distribution function  $F_X$ .
- Calculate the mean and variance for  $X$ .
- What is the probability that a student can finish the test in 90 minutes?

- 23.** Determine which of the following is a distribution function:

$$F(x) = \begin{cases} \frac{1}{2} e^x & \text{for } x < 0 \\ 1 - \frac{3}{4} e^{-x} & \text{for } x \geq 0. \end{cases} \quad F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x}{1+x} & \text{for } x \geq 0. \end{cases}$$

- 24.** Let  $F_X$  be a distribution function of a continuous r.v.  $X$  such that:

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sqrt{x} & \text{for } 0 < x \leq 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$

Then:

- Determine the density function  $f_X$
- Calculate the mean and variance of  $X$ .
- Calculate  $P(0.125 < X < 0.25)$ .

- 25.** Let  $F_X$  be the distribution of a continuous r.v.  $X$  such that:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 0.5x^2 & \text{for } 0 < x \leq 1 \\ 1 - 0.5(1-x)^2 & \text{for } 1 \leq x \leq 2 \\ 1 & \text{for } x \geq 2 \end{cases}$$

Then:

- Determine the density function  $f_X$ .

- b. Calculate the mean and variance of  $X$ .
- c. Calculate  $P(0.225 < X < 0.64)$ .
- d. Calculate  $E(3X + 20)$  and  $\text{var}(5X + 20)$

**26.** A probability density function of a continuous **r.v.**  $X$  is given by:

$$f_X(x) = ax + b \quad ; 0 < x < 1$$

And verifying the equation  $P(X > 0.5) = 0.3$ . Then:

- a. Use the properties of  $f_X$  and the above probability to determine the values of  $a$  and  $b$ .
- b. Calculate  $P(0.2 < X < 0.9)$ .
- c. Calculate the mean of  $X$ .
- d. Calculate the variance and standard deviation of  $X$ .

**27.** Let  $X$  be a continuous uniform **r.v.** with parameter  $a = 5$  and  $b = 10$ , that means it has **p.d.f.:**

$$f_X(x) = \begin{cases} 0.2 & \text{for } 5 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases} .$$

- a. Then calculate the mean, variance and standard deviation.
- b. Calculate  $E(10X + 12)$ .

**28.** Let  $X$  be a continuous **r.v.** with **p.d.f.:**

$$f_X(x) = \frac{3}{x^4} \quad ; x > 1,$$

Then:

- a. Calculate the mean of  $X$ .
- b. Calculate  $E(3X + 2)$
- c. Calculate  $P(3 < X < 7)$  and  $P(X < 20)$  .

**29.** The distribution function for the inverse exponential **r.v.**  $X$  with parameter  $\lambda > 0$  is:

$$f(x) = \frac{\lambda}{x^2} e^{-\frac{\lambda}{x}} \quad ; x \geq 0$$

Then:

- a. Find the distribution  $F_X$  and calculate the mean of  $X$ .
- b. Draw the graph of the density and distribution for  $\lambda = 1$ .

# CHAPTER 4

## INTRODUCTION TO STATISTICAL INFERENCE



### INTRODUCTION

The statistical inference is the process of making judgment about a population based on the properties of a random sample from the population. There are two types of statistical inference, they are the estimation and hypotheses testing. The estimation techniques can be divided into two types they are the point estimation and interval estimation. For example, you may want to know the height of Saudi male adult (this piece of information may be useful for many purposes, such as social, health, economy, etc.). The hypotheses testing techniques is known as the decision making approach and can be applied in many real life problems. For example, we may wish to decide if the mean value of the home in Riyadh city is more than S.R.1000,000. Other example could be testing the ratio of diabetic people in Saudi Arabia, many other applications can be found in different real life situations.

In this chapter, we introduce the two types of inferences for the population mean and proportion. The standard normal distribution and central limit theorem (Chapter 3) play an important role in this study.

- SECTION 4.1      DEFINITIONS AND CONCEPTS
- SECTION 4.2      ESTIMATION OF THE POPULATION MEAN
- SECTION 4.3      ESTIMATION OF THE POPULATION PROPORTION
- SECTION 4.4      INTRODUCTION TO HYPOTHESES TESTING
- SECTION 4.5      HYPOTHESES TESTING FOR THE POPULATION MEAN
- SECTION 4.6      HYPOTHESES TESTING FOR THE POPULATION PROPORTION

## Section 4.1

# DEFINITIONS AND CONCEPTS

In the first chapter we have presented some basic concepts in statistics. For example: Descriptive Statistics, Inferential Statistics, Population, Sample, Parameter, Statistic, Variables and etc....

Now we will follow up on some of the other statistical concepts that inferential statistics need, and develop some other concepts to match the following study in the explanatory census.

A set representing all the elements of the population under study will be marked with  $\Omega$  also, and in this study of statistics one say that  $\Omega$  is a **sample space**.

Several times in this course we use the term “random sample”. Generally, the value of our data is only as good as the sample that produced it. For example, suppose we wish to estimate the proportion of all students at a large university who are females, which we denote by  $p$ . If we select 50 students at random and 27 of them are female, then a natural estimate is  $p \approx \hat{p} = \frac{27}{50} = 0.54$  or 54%. How much confidence we can place in this estimate depends not only on the size of the sample, but on its quality, whether or not it is truly random, or at least truly representative of the whole population. If all 50 students in our sample were drawn from a College of Nursing, then the proportion of female students in the sample is likely higher than that of the entire campus. If all 50 students were selected from a College of Engineering Sciences, then the proportion of students in the entire student body who are females could be underestimated. In either case, the estimate would be distorted or biased. In statistical practice an unbiased sampling scheme is important but in most cases not easy to produce. For this introductory course we assume that all samples are either random or at least representative.

In our future studies we will focus on the use of simple random samples, which means that we will assume that:

1. All elements of population have the same appearance (or choice).
2. All elements of population are independent of each other.

In other words, the elements of a simple random sample are independent of each other, and all samples watch have the same size, it will have the same probability of selection.

**REMARK 4.1.1**

Mathematically one denotes a simple random sample with a random vector of form:

$$\mathcal{X} = (X_1, X_2, \dots, X_n)$$

In this case, all random variables  $X_1, X_2, \dots$  and  $X_n$  independent of each other, as well as all of them have the same distribution of the random variable  $X$  which that describes the statistical population. For example, if the population is normal with mean  $\mu$  and standard deviation  $\sigma$ , then the random variable  $X$  will have a normal distribution  $N(\mu, \sigma^2)$ , so all random variables  $X_1, X_2, \dots$  and  $X_n$  will have the same normal distribution  $N(\mu, \sigma^2)$  also.

**DEFINITION 4.1.1 (Estimator)**

An estimator is a statistic (a function of a random sample  $\mathcal{X}$ ) whose value depends on the particular sample drawn.

For example if we have  $\mathcal{X} = (X_1, X_2, \dots, X_n)$  a random sample of a population  $\Omega$ . Then:

- a.  $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$  is an estimator (one denotes the value of  $\bar{x}$  by  $\bar{X}$ ).
- b.  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})^2$  is an estimator (one denotes the value of  $s^2$  by  $S^2$ ).

**REMARKS 4.1.2**

1. The value of the estimator is called the estimate and it is used to predict the value of a population parameter.
2. There are many ways to determine an estimator of a parameters of the statistical population, but we will not offer to these methods, and only provide some estimators for normal and Bernoulli populations. For example:
  - a. The mean  $\mu$  of the normal population has  $\bar{x}$  as an estimator.
  - b. The variance  $\sigma^2$  of the normal population has  $s^2$  as an estimator.
  - c. The proportion  $p$  of the Bernoulli population has  $\bar{x}$  as an estimator. For this case  $\bar{x}$  represents the relative frequency of successes in the sample  $\mathcal{X}$ .

Therefore, in this case one denote the value of  $\bar{x}$  by  $\hat{p}$ ,  $\hat{p} = \frac{k}{n}$ , where  $n$  is the sample size and  $k$  is the number of successes in the random sample.

## SECTION 4.1 DEFINITIONS AND CONCEPTS

The following chart shows the steps required to draw the statistical inference (estimation) of an unknown parameter.

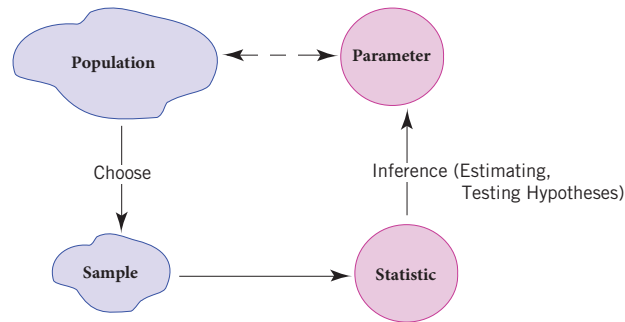


Figure 4.1.1

### DEFINITION 4.1.2 (Standard Normal Percentile (z-Value))

The **Standard Normal Percentile (or z-Value)**  $z_\alpha$  is that value on the real axis, which is on its left and under the curve of the density function of standard normal distribution, there is area equal to  $\alpha$ .

► **EXAMPLE 4.1.1** For  $z_\alpha = 0.57$  we find from the table of standard normal distribution that:

$$P(Z < z_\alpha) = \Phi(z_\alpha) = \Phi(0.57)$$

Table 4.1.1

$z$	0.00	0.01	...	0.07	0.08	0.09
0.0	0.5000	0.5040		0.5279	0.5391	0.5359
⋮						
0.5	0.6915	0.6950		0.7157	0.7190	0.7224
0.6	0.7257	0.7291		0.7486	0.7517	0.7549
0.7	0.7580	0.7611		0.7794	0.7823	0.7852

So we find  $P(Z < z_\alpha) = \Phi(0.57) = 0.7157$

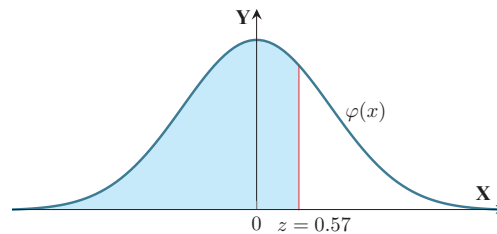


Figure 4.1.2

The standard normal distribution can be utilized in this chapter to draw some statistical inferences about the population mean and proportion. It is standard practice to identify the

area denoted by  $\alpha$  in the two tails of the standard distribution of when the middle part specified by the  $(1 - \alpha)$  is taken out. This is shown in Figure 4.1.3-a, drawn for the general situation, and in Figure 4.1.3-b, drawn for  $(1 - \alpha) = 0.95$ , remember from Chapter 3 that the standard normal distribution is a continuous random variables. The  $z$ -value that cuts off a right tail of area  $\alpha / 2$  is denoted  $z_{1-(\alpha/2)}$ , and the  $z$ -value that cuts off a left tail of area  $\alpha / 2$  is denoted  $-z_{1-(\alpha/2)}$ , i.e.  $z_{\alpha/2} = -z_{1-(\alpha/2)}$ .

Thus the numbers  $\pm 1.96$  in the example are  $\pm z_{1-(\alpha/2)} = \pm z_{1-0.025} = \pm z_{0.975}$ , which is for  $\alpha = 1 - 0.95 = 0.05$ .

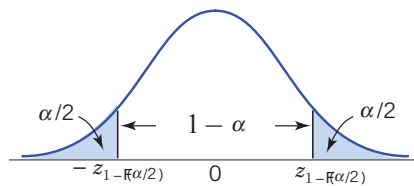


Figure 4.1.3-a (Areas under the two-tail Standard normal)

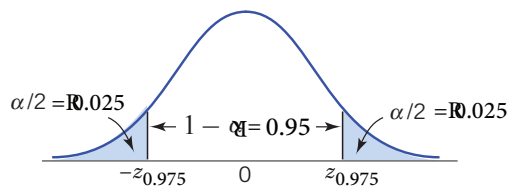


Figure 4.1.3-b (Areas of size 0.025 under the two-tail Standard normal)

Some common values of  $z_{1-(\alpha/2)}$  are given below:

Table 4.1.2

$100(1 - \alpha)\%$	$z_{1-(\alpha/2)}$
90%	1.65
95%	1.96
98%	2.33
99%	2.58

In most of statistical study population is the object of interest for which we would like to make inference. In fact, due to limited resource of time, funding or the nature of the population such as it is too large, it is sometimes impossible to know every aspect of the population. In these situations one instead, obtain a (random) sample from the population. Thus using the information from the sample to make inference about the population is mostly adopted by researchers.

Importance of sampling distributions comes from:

## SECTION 4.1 DEFINITIONS AND CONCEPTS

1. It gives the probability of getting a particular  $\bar{x}$  given the mean  $\mu$  and the standard deviation  $\sigma$  of the population.
2. It gives us estimates for population parameters.
3. It determines whether the sample mean differs from a known population mean only due to chance, or due to experimental treatment.

Next, for example, one naturally uses sample mean  $\bar{x}$  to estimate the population mean  $\mu$  and one can use the sample proportion to estimate the population proportion for any characteristic of the population.

It is known that  $\bar{x}$  is random variable, therefore, it has a probability distribution. This distribution is called the sampling distribution of  $\bar{x}$ .

Note that the probability distribution of this random variable is called sampling distribution.

Next, we present the following theorem.

### THEOREM 4.1.1 (The Central Limit Theorem)

Let  $\Omega$  be the sample space of a population, which described by a **r.v.**  $X$  with mean  $\mu$  and standard deviation  $\sigma$ , and  $\mathcal{X} = (X_1, X_2, \dots, X_n)$  is a random sample of  $\Omega$ . Then for sufficiently large sample size of  $X$ , the sampling distribution of  $\bar{x}$  follows normal distribution with mean  $\mu_{\bar{x}} = \mu$  and standard deviation  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  (it is called standard error also).

► **EXAMPLE 4.1.2** Let the number of defects in newly manufactured bulbs in each working shift is a **r.v.**  $X$  with normal distribution that has a mean  $\mu = 3$  and standard deviation  $\sigma = 2.5$  (here  $\mu$  and  $\sigma$  are known). Now, consider that in one of the working shifts a sample of 225 new bulbs were tested. Next find the following:

- a. The sampling distribution of  $\bar{x}$  based on samples of size 225.
- b. The mean and the standard deviation of  $\bar{x}$ .
- c. The probability that the sample average number of defects exceeds 3.

**Solution:** We have:

**For a)** Note that  $\bar{x}$ , based on samples of size 225, is normally distributed.

**For b)** Also, the mean of  $\bar{x}$  is  $\mu_{\bar{x}} = 3$  and the standard deviation (standard error) is:

$$\sigma_{\bar{x}} = \frac{2.5}{\sqrt{n}} = \frac{2.5}{\sqrt{225}} = 0.17$$



**For c)** The probability that the sample average number of major defects exceeds 3 is

$$P(\bar{x} \geq 3) = P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \geq \frac{3 - 3}{0.17}\right) = P(Z \geq 0) = 1 - P(Z < 0) = 0.5$$

### SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION $P$

Consider that the STAT 150 course has 200 Saudi (S) and Non-Saudi (NS). The proportion of NS is  $p$ , and is not known. Here we note that, we have a Bernoulli population with parameter  $0 < p < 1$  (unknown). Now assume 20 students were randomly selected from the section with replacement, and 6 students out of the 20 students were NS. In the sample, the proportion of NS is  $6 / 20 = 0.3$ . This quantity is known by the sample proportion.

Next, one is interested to find the sample distribution of the proportion  $p$ . One may consider a Bernoulli r.v such that  $X = 1$  if for NS and  $X = 0$  for S. Thus one may view  $p$  as the mean of 20  $X$ 's or the statistic  $\bar{x}$ . Thus by the central limit theorem and for large  $n$ , we get that  $p$  is normally distributed with mean  $\mu_{\bar{x}} = p$  and standard deviation:

$$\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}}$$

By large sample size we mean  $n \geq 30$  or  $np \geq 5$ ,  $n(1-p) \geq 5$ . If this condition is not satisfied, then the proportion  $p$  has unknown distribution with the same mean and variance.

#### REMARK 4.1.3

Let  $\mathcal{X} = (X_1, X_2, \dots, X_n)$  be a random sample of a Bernoulli population with sample space  $\Omega$ , and suppose that  $k$  of  $X_1, X_2, \dots, X_n$  satisfying the characteristic of interest, then the value  $k / n$  is  $\hat{p}$ . This value is used as an alternative (or as an estimate) for  $p$  (because  $p$  is unknown) in arithmetic operations.

► **EXAMPLE 4.1.3** In Example 4.1.2, suppose that the proportion of the population of defects bulbs is  $\hat{p} = 0.20$ , then calculate the following:

- The sampling distribution of  $p$  based on 225 observations,
- The mean and the standard error.
- The probability that  $p < 0.25$ .

## SECTION 4.1 DEFINITIONS AND CONCEPTS

**Solution:** We have:

**For a)** According to the central limit theorem we note that the sampling distribution of  $p$  approximate the normal distribution.

**For b)** According to the central limit theorem we have:

$$\mu_{\bar{x}} = \hat{p} = 0.20$$

And:

$$\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20(1-0.20)}{225}} = 0.0267$$

**For c)** The probability:

$$\begin{aligned} P(p < 0.25) &= P\left(\frac{p - \mu_{\bar{x}}}{\sigma_{\bar{x}}} < \frac{0.25 - 0.20}{0.0267}\right) \\ &= P(Z < 1.872659) \\ &\simeq P(Z < 1.87) \end{aligned}$$

By using the  $z$ -table, we get

$$P(p < 0.25) = P(Z < 1.87) = 0.9693.$$



## Section 4.2

# ESTIMATION OF THE POPULATION MEAN

### DEFINITION 4.2.1 (Point Estimation)

The point estimation is a summarization of the sample by a single number, and this number is an estimate of the population parameter.

► **EXAMPLES 4.2.1** The follow data represents the hemoglobin level in (g/dl) of a sample of 50 men in a certain city.

17.0	17.7	15.9	15.2	16.2	17.1	15.7	17.3	13.5	16.3
14.6	15.8	15.3	16.4	13.7	16.4	16.1	17.0	17.8	15.9
14.0	16.2	16.4	14.9	17.8	16.1	15.5	18.3	15.8	16.7
15.9	15.3	13.9	16.8	15.9	16.3	17.4	15.0	17.5	16.1
14.2	16.1	15.7	15.1	17.4	16.5	14.4	16.3	17.3	15.8

Find the point estimation of the overall mean of the hemoglobin level in the city.

**Solution:** It is easy to calculate the sample mean of this sample, so we get:

$$\bar{x} = \frac{1}{50} \sum_{i=1}^{50} x_i = \frac{1}{50} (17.0 + 17.7 + 15.9 + \dots + 17.3 + 15.8) = 16 \text{ g/dl.}$$

Then the point estimation of the population mean (overall mean of the hemoglobin level in the city) is  $\hat{\mu} = \bar{x} = 16 \text{ g/dl}$ .

When we estimate an unknown parameter by a point estimator, we do not expect the resulting estimator to exactly equal the parameter (mean), but we expect that it will be “close” to it. To be more specific, we sometimes try to find an interval about the point estimator in which we can be highly confident that the parameter lies. Such an interval is called an interval estimator of the population mean. It is simply the sample mean  $\pm$  margin error.

### DEFINITION 4.2.2 (Interval Estimation)

The interval estimation of a population parameter is an interval that is predicted to contain the parameter.

**DEFINITION 4.2.3 (Confidence Interval)**

A confidence interval is a range of values within which, we believe, the true value of population parameter lies in it with a specific probability (it is called the confidence level **or** the level of confidence).

A confidence interval is derived from the sampling distribution of the mean. When a sample is measured and a sample mean and standard deviation computed, a confidence interval can be computed. Whether the interval contains the population parameter is not known. However, the chosen level of confidence (for example, 95%) provides a probability statement that a certain percentage of samples (95%) will provide confidence intervals that include the population mean.

A confidence level  $100(1 - \alpha)\%$  refers to the percentage of all possible samples that can be expected to include the true population parameter. For example, suppose all possible samples were selected from the same population, and a confidence interval were computed for each sample. A 95% confidence level implies that 95% of the confidence intervals would include the true population parameter.

**REMARK 4.2.1**

The level of confidence of an interval estimation can be any number between 0 and 100%, but usually one chooses this probability value as large value (usually greater than or equal to 0.90, but does not preclude taking smaller values). The most common values are probably 90% ( $\alpha = 0.10$ ), 95% ( $\alpha = 0.05$ ), and 99% ( $\alpha = 0.01$ ).

**CONFIDENCE INTERVAL FOR A POPULATION MEAN**

To determine an interval estimation of a population parameter, we use the probability distribution of the point estimator of that parameter. Let us see how this works in the case of the interval estimator of a mean when the population standard deviation is assumed known.

Let  $\Omega$  the sample space of a normal population with mean  $\mu$  (is unknown) and standard deviation  $\sigma$  (is known and equal to  $\sigma$ ). Now let's take  $\mathcal{X} = (X_1, X_2, \dots, X_n)$  a simple random sample, and suppose we want to utilize this sample to obtain a  $100(1 - \alpha)\%$  confidence interval estimator for the population mean  $\mu$ . To obtain such an interval, we start with the sample mean  $\bar{x}$ , which is the point estimator of the population mean  $\mu$ . We now make use of the sampling distribution of the statistic  $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$  and the probability statement as shown in below:

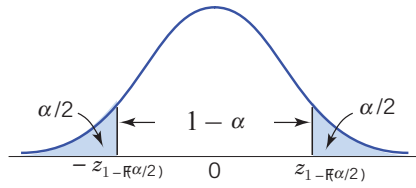


Figure 4.2.1

$$P\left(-z_{1-(\alpha/2)} < Z < z_{1-(\alpha/2)}\right) = P\left(-z_{1-(\alpha/2)} < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < z_{1-(\alpha/2)}\right) = 1 - \alpha$$

Then a  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  in this case is given by:

$$-z_{1-(\alpha/2)} < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < z_{1-(\alpha/2)}$$

$$\bar{x} - z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}}$$

Calculate the value of the statistic  $\bar{x}$  from the sample, so we become a  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  by the interval in which its bounds:

$$\bar{x} \pm z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}}$$

In many applications, the standard deviation is unknown and we do not know whether the underlying population is normal or not. In such cases, we utilize the central limit theorem to obtain the confidence interval for the population mean.

The following table summarizes different cases.

Table 4.2.1

Case	Population	Sample size	Standard deviation	Confidence interval
1	normal	any sample size	known	$\bar{x} \pm z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}}$
2	any population	large ( $n \geq 30$ )	known	$\bar{x} \pm z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}}$
3	any population	large ( $n \geq 30$ )	unknown but can be replaced by the sample standard deviation $s$	$\bar{x} \pm z_{1-(\alpha/2)} \frac{s}{\sqrt{n}}$

**DEFINITION 4.2.4 (Margin Error)**

The margin error is the maximum error (increase or decrease) resulting from using the value of the statistic instead of the value of the parameter.

In a confidence interval of the population mean, the range of values above and below the sample mean is called the **margin of error** (we denote it by  $\delta_\mu$ ) that is:

$$\delta_\mu = z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \delta_\mu = z_{1-(\alpha/2)} \frac{s}{\sqrt{n}}$$

This means that if we use the value  $\bar{x}$  instead of  $\mu$ , we will make a maximum error of  $\delta_\mu$  with probability equal to  $1 - \alpha$ .

► **EXAMPLE 4.2.2** The following measurements were recorded for the drying time, in hours, of a certain brand of paint:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Assuming that the measurements represents a random sample from a normal population with known standard deviation as  $\sigma = 0.96$  hours. Then we will determine:

- 99% confidence interval of the mean  $\mu$  drying time of this brand of paint and calculate the interval length.
- The marginal error  $\delta_\mu$ .

**Solution:**

**For a)** We have the confidence level 99%, therefore  $\alpha = 1 - 0.99 = 0.01$ , and then  $\alpha / 2 = 0.005$ . So is  $z_{1-(\alpha/2)} = z_{0.995}$ . From  $Z$ -table, we find  $z_{0.9949} = 2.57$  and  $z_{0.9951} = 2.58$ , therefore, we take the average of the previous two values for value  $z_{0.995}$ , so we:

$$z_{0.995} = \frac{2.57 + 2.58}{2} = 2.575$$

Now, from the given sample, we have  $n = 15$  and  $\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = 3.8$  hours, and

because our population is normal with known standard deviation  $\sigma = 0.96$ , we find that 99% confidence interval for  $\mu$  is:

$$\bar{x} \pm z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} = 3.8 \pm (2.575) \frac{0.96}{\sqrt{15}} = 3.8 \pm 0.638$$

then, 99% confidence interval is  $(3.8-0.638, 3.8+0.638) = (3.162, 4.438)$ . Hence, one may be 99% confident that the true average drying time of this brand of paint is between 3.162 and 4.438 hours.

**For b)** The marginal error  $\delta_\mu$  is given by  $\delta_\mu = \pm z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} = \pm (2.575) \frac{0.96}{\sqrt{15}} = \pm 0.638$  hours.

► **EXAMPLE 4.2.3** A random sample of 120 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university.

**Solution:** For confidence level 90%,  $\alpha = 1 - 0.90 = 0.10$ ,  $\alpha / 2 = 0.05$ , so  $z_{1-(\alpha/2)} = z_{0.95}$ .

From Z-Table, we have  $z_{0.95} = \frac{1.64 + 1.65}{2} = 1.645$ .

Since our population is large ( $n = 120$ ) and the standard deviation is known to be  $\sigma = 0.51$ , we use the formula:

$$\bar{x} \pm z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} = 2.71 \pm (1.645) \frac{0.51}{\sqrt{120}} = 2.71 \pm 0.0766$$

Then, one may be 90% confident that the true average GPA of all students at the university is contained in the interval  $(2.71 - 0.0766, 2.71 + 0.0766) = (2.63, 2.79)$ .

► **EXAMPLE 4.2.4** Thirty-six cars of the same model are driven the same distance and conditions. The gas mileage for each is recorded. The results give  $\bar{x} = 18$  miles per gallon with standard deviation of  $s = 3$  miles per gallon. Give a 95% confidence interval for the mean mileage  $\mu$  for all cars of this model.

**Solution:** For confidence level 95%,  $\alpha = 1 - 0.95 = 0.05$ ,  $\alpha / 2 = 0.025$ , so  $z_{1-(\alpha/2)} = z_{0.975}$ . From Z-Table, we have  $z_{0.975} = 1.96$ . Since our population is large ( $n = 36$ ) and the standard deviation for the population is unknown, but, the standard deviation for the sample is given to be  $s = 3$  we use the formula

$$\bar{x} \pm z_{1-(\alpha/2)} \frac{s}{\sqrt{n}} = 18 \pm (1.96) \frac{3}{\sqrt{36}} = 18 \pm 0.98$$

Therefore, we are 95% confident that the mean gas mileage of this model of cars is between 17.02 and 18.98 miles per gallon.

**DETERMINING OF THE SAMPLE SIZE**

Questions about sample size are important in research. Too small sample will yield scant information; but ethics, economics, time and other constraints require that a sample size not be too large. “How many subjects do I need?” Neither 7 nor 30 nor any number is an all-purpose answer. A sample size of 30 is a “large sample” in some textbook discussions of “normal approximation”; yet 30,000 observations still may be too few to assess a rare, but serious teratogenic effect. The best first response to “how many?” may be not a number, but a sequence of further questions. A study’s size and structure should depend on the research context, including the researcher’s objectives and proposed analyses. One cautionary note: “How many subjects do I need?” is linked to the companion question, “How many subjects can I afford to get?” The final decision on sample size must pay attention to the various constraints on recruitment.

In this part, we calculate the sample size required in order to get specified margin error. So, we use the relationship between the sample size and the margin error  $\delta_\mu$  to get:

$$n = \left( \frac{\sigma z_{1-(\alpha/2)}}{\delta_\mu} \right)^2$$

where  $\delta_\mu$  is the desired margin error,  $\sigma$  is the population standard deviation and  $z_{\alpha/2}$  is a standard normal percentile ( $z$ -value).

► **EXAMPLE 4.2.5:** In Example 4.2.2, how many records of draying time required for estimating the paint drying mean to be ensured that the marginal error will not exceed 0.5 hour with 95% confidence level.

**Solution:** The marginal error is given by  $\delta_\mu = 0.5$  and the required sample size is given by

$$n \geq \left( \frac{\sigma z_{1-(\alpha/2)}}{\delta} \right)^2 = \left( \frac{(0.96)z_{0.975}}{0.5} \right)^2 = \left( \frac{(0.96)(1.96)}{0.5} \right)^2 = 14.162 \approx 15 \text{ records.}$$



## Section 4.3

# ESTIMATION OF THE POPULATION PROPORTION

In many applications, one may be interested to investigate a certain phenomenon (or a characteristic) in a population. Some objects of the population are following the underlying phenomenon or satisfying the characteristic of interest, while the other population objects are not. Here an important question that is what is the proportion of those objects following the phenomenon or satisfying the characteristic of interest. For example, suppose that a medical researcher wants to know, what is the proportion of the diabetic adult Saudi. This proportion cannot be measured exactly unless the researcher investigates the overall population, which is a highly cost experiment. Instead, he could collect a random sample represents the population and measure this proportion from the collected sample. The measure of proportion from the sample is called a point estimation for  $p$ . Suppose that the researcher has collected a random sample of size 1000 adults Saudi and found 320 diabetics among the sample. Then the population point estimation in this case is  $\frac{320}{1000} = 32\%$ .

► **EXAMPLES 4.3.1** In Example 4.2.1, what is the overall proportion of men have the hemoglobin level less than 16 g/dl the whole city.

**Solution:** From the data in Example 4.2.1, we can count 23 men have the hemoglobin level less than 16 g/dl, then  $\hat{p} = \frac{k}{n} = \frac{23}{50} = 0.46 = 46\%$ . This means that 46% from the men in the whole city have the hemoglobin level less than 16 g/dl.

### DETERMINE THE INTERVAL ESTIMATION

The point estimation of the population proportion is depending on the random sample, this estimate may be being rough estimate in most cases. The true value of  $p$  could be around the point estimate with some margin error. So, constructing a confidence interval for  $p$  would be more appropriate. For example, to estimate the proportion of the adult people in Saudi Arabia whom suffer from the diabetic disease. We can view the investigation of 1000 Saudi adults as a binomial distribution, such that:

- Trial = randomly selected Saudi adult,
- Success =  $S$  = the selected adult is diabetic,

## SECTION 4.3 ESTIMATION OF THE POPULATION PROPORTION

- Failure =  $F$  = the selected adult is non-diabetic,
- $p_S = p$  = proportion of the diabetic,
- $p_F = q = 1 - p$  = proportion of the non-diabetic,
- $n$  = the number of the independent Bernoulli trails, that is the number of persons in the sample,
- $x$  = the number of success, that is the number of diabetic persons in the sample.

One can prove that when  $n$  is sufficiently large; the variable  $X$  can be regarded as approximately normal distribution. We said that  $n$  is large enough if  $np$  and  $n(1 - p)$  are at least 5. If  $p$  is unknown, and suppose that the success in the sample equal to  $k$ , then we may replace  $np$  and  $n(1 - p)$  by  $n\hat{p} = k$  and  $n(1 - \hat{p}) = n - k$  respectively. Then the conditions for the large sample are now both of  $k$  and  $n - k$  are at least 5. Also, from Chapter 3, we have the mean of  $X$  is  $np$  and the standard deviation is  $\sqrt{np(1 - p)}$ . Then  $\hat{p}$  is also approximately normal with mean  $p$  and standard deviation  $\sqrt{\frac{p(1 - p)}{n}}$ , and the following statistic is approximately normal distribution:

$$Z = \frac{\hat{p} - p}{\sqrt{p(1 - p) / n}}$$

When  $n$  is sufficiently large, a  $100(1 - \alpha)\%$  confidence interval for the population proportion  $p$  in this case can be constructed by using the probability statements:

$$P(-z_{1-(\alpha/2)} < Z < z_{1-(\alpha/2)}) = 1 - \alpha$$

or it can be written as

$$\left( \hat{p} - z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) = (\hat{p} - \delta_p, \hat{p} + \delta_p)$$

The quantity  $\delta_p = z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$  is called the margin error of the population proportion estimation with  $100(1 - \alpha)\%$  confidence level.

### DETERMINING OF MARGIN OF ERROR

In a confidence interval of the population proportion, the range of values above and below the sample proportion is called the **margin of error** (or the maximum error) that is

$$z_{1-(\alpha/2)} \sqrt{\hat{p}(1 - \hat{p})}$$

For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

► **EXAMPLE 4.3.2** Assume the medical researcher collected a random sample of size 1000 from Saudi adults (in the previous discussion) and found 320 of them are diabetics.

- Find the point estimation of the proportion of the diabetic Saudi adults.
- Find a 90% confidence interval for  $p$ .
- If the point estimation is used to estimate  $p$ , find the margin error of the estimate with 90% confidence.
- Comment on the how these results should be interpreted.

**Solution:** For confidence level 90% we have  $\alpha = 1 - 0.90 = 0.10$ , so  $z_{1-(\alpha/2)} = z_{0.95} = 1.645$ .

So we have:

**For a)** The point estimation of  $p$  is  $\hat{p} = \frac{k}{n} = \frac{320}{1000} = 0.32$ .

**For b)** To find 90% confidence interval for  $p$ , we see the confidence level 90%, therefore we get  $\alpha = 1 - 0.90 = 0.10$ , then  $\frac{\alpha}{2} = 0.05$ , so  $z_{1-\alpha/2} = z_{0.95} = 1.645$ . Well we have  $n = 1000$ ,  $k = 320$ , and then we get  $\hat{p} = 0.32$ , therefore,  $1 - \hat{p} = 1 - 0.32 = 0.68$ .

Then 90% confidence interval is given by:

$$\hat{p} \pm z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.32 \pm 1.645 \sqrt{\frac{0.32(0.68)}{1000}} = 0.32 \pm 0.02427$$

Therefore, we have 90% confidence interval is (0.344, 0.296).

**For c)** The margin error when we use the point estimation for the population proportion in this case is  $\delta_p = 0.0243396 \approx 0.024$ .

**For d)** From part (b), the researcher can be 90% confidence that the percentage of the diabetic's Saudi adults is between 30% and 34.3%. From part (c) if we consider the point estimation of such percentage (32%), we can sure with 90% that this estimate will be in error by less than 2.4%.

## SECTION 4.3 ESTIMATION OF THE POPULATION PROPORTION

► **EXAMPLE 4.3.3** A random sample of 125 individuals working in a large city indicated that 42 are dissatisfied with their working conditions. Construct a 90% lower confidence bound on the percentage of all workers in that city who are dissatisfied with their working conditions.

**Solution:** Since  $z_{1-(\alpha/2)} = z_{0.975} = 1.96$  and  $\hat{p} = 42 / 125 = 0.336$ , then the 90% lower bound is given by:

$$\hat{p} - z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.336 - 1.96 \sqrt{\frac{0.336(1-0.336)}{125}} = 0.2665$$

That is, we can be 90% certain that over 26.7% of all workers are dissatisfied with their working conditions. ◀

► **EXAMPLE 4.3.4** Out of a random sample of 100 students at a university, 82 stated that they were nonsmokers. Based on this sample, construct a 99% confidence interval estimate of the population proportion of all the students at the university who are nonsmokers.

**Solution:** Since  $100(1-\alpha)\% = 0.99$  when  $\alpha = 0.01$ , we need the value of  $z_{1-(\alpha/2)} = z_{0.995}$  which can be obtained from  $Z$ -table as we have  $z_{0.995} = 2.575$ . Also we have:

$$\hat{p} = \frac{82}{100} = 0.82 \Rightarrow 1 - \hat{p} = 1 - 0.82 = 0.18$$

Then 99% confidence interval estimate of the population proportion  $p$  is given by:

$$\hat{p} \pm z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.82 \pm 2.575 \sqrt{\frac{0.82(0.18)}{100}} = 0.82 \pm 0.0989$$

That is, we can assert with 99% percent confidence that the true percentage of nonsmokers is between 72.1% and 91.9%.

► **EXAMPLE 4.3.5** On December 24, 1991, The New York Times reported that a poll indicated that 46% of the population was in favor of the way that President Bush was handling the economy, with a margin of error of  $\pm 3$  percent. What does this mean?

Can we infer how many people were questioned?

**Solution:** It has become common practice for the news media to present 95% confidence intervals. That is, unless it is specifically mentioned otherwise, it is almost always the case that the interval quoted represents a 95% confidence interval. Since  $z_{1-(\alpha/2)} = z_{0.975} = 1.96$ , a 95% confidence interval for the population proportion  $p$  in this case is given by:

$$\hat{p} \pm z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where  $n$  is the sample size.

Since  $\hat{p}$ , the proportion of those in the random sample who are in favor of the President's handling of the economy, is equal to 0.46, it follows that the 95% confidence interval estimate of  $p$ , the proportion of the population in favor, is:

$$0.46 \pm 1.96 \sqrt{\frac{(0.46)(1 - 0.46)}{n}}$$

Since the margin of error is  $\pm 3$  percent, it follows that:

$$1.96 \sqrt{\frac{(0.46)(1 - 0.46)}{n}} = 0.03$$

Squaring both sides of this equation shows that:

$$(1.96)^2 \frac{(0.46)(1 - 0.46)}{n} = (0.03)^2$$

Hence

$$n = (1.96)^2 \frac{(0.46)(0.54)}{(0.03)^2} = 1060.3$$

That is, approximately 1060 people were sampled, and 46 percent were in favor of President Bush's handling of the economy

### SAMPLE SIZE DETERMINATION WHEN ESTIMATING $P$

From the previous example, we can determine the sample size for estimating the population proportion by the meaning of the maximum error, where  $\hat{p} \pm z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \hat{p} \pm \delta_p$  and maximum error in this case is denoted by  $\delta_p$ . Then

$$\delta_p \leq z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Rewire this form, we get the sample size in terms of some specific value of the maximum error  $\delta_p$  as:

$$n \geq \left( \frac{z_{1-(\alpha/2)}}{\delta_p} \right)^2 \hat{p}(1 - \hat{p})$$

The value of  $\hat{p}(1 - \hat{p})$  is not known, we use  $\hat{p}(1 - \hat{p}) = 1/4$ , this is because if we take the function  $f(\hat{p}) = \hat{p}(1 - \hat{p})$ , then for  $f'(\hat{p}) = 0 \Rightarrow \hat{p} = 1/2$ . Therefore, we get:

$$n \geq 0.25 \left( \frac{z_{1-(\alpha/2)}}{\delta_p} \right)^2$$

## SECTION 4.3 ESTIMATION OF THE POPULATION PROPORTION

The sample size required to estimate the population proportion can be obtained from the preceding form when there no information available about the value of  $\hat{p}(1 - \hat{p})$ .

► **EXAMPLE 4.3.6** How large a sample is needed to ensure that the maximum error of the 95% confidence interval estimate of  $p$  is less than 0.01?

**Solution:** Since the maximum (margin) error required to estimate the population proportion  $p$  is less than 0.01, we need to choose  $n$  so that

$$n \geq 0.25 \left( \frac{z_{1-(\alpha/2)}}{\delta_p} \right)^2 = 0.25 \left( \frac{1.96}{0.01} \right)^2 = 9604$$

That is, the sample size needs to be at least 9604 to ensure with 90% confidence that the maximum error will not exceed 0.01. As we can see as the error decreases, the sample size increases while as the value of  $\alpha$  increases, the level the sample size increases. Some common sample sizes in practice are summarized below.

**Table 4.3.1**

$\delta_p$	$\alpha$	$n$	$\delta_p$	$\alpha$	$n$
0.05	0.05	384.2	0.05	0.01	665.6
0.04	0.05	600.3	0.04	0.01	1040.1
0.03	0.05	1067.1	0.03	0.01	1849.0
0.01	0.05	9604.0	0.01	0.01	16641.0

Sample size should be an integer value, otherwise we approximate to the bigger integer. ◀

Other method for the sample size determination can be based on the length of the confidence interval of the population proportion, since the length of a  $100(1 - \alpha)\%$  confidence interval is

$$\left( \hat{p} + z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) - \left( \hat{p} - z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) = 2 z_{1-(\alpha/2)} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

It can be shown that the product  $\hat{p}(1 - \hat{p})$  is always less than or equal to  $\frac{1}{4}$ , it follows from the preceding expression that an upper bound on the length of the confidence interval is given by  $2 z_{1-(\alpha/2)} \sqrt{1 / (4n)}$  which is equivalent to the statement:

$$\text{Length of } 100(1-\alpha)\% \text{ confidence interval} \leq \frac{z_{1-(\alpha/2)}}{\sqrt{n}}$$

The preceding bound can be used to determine the appropriate sample size needed to obtain a confidence interval whose length is less than a specified value. For instance, suppose that we want to determine a sufficient sample size so that the length of the resulting  $100(1 - \alpha)\%$  confidence interval is less than some fixed value  $L$ . In this case, upon using the preceding inequality, we can conclude that any sample size  $n$  for which  $\frac{z_{1-(\alpha/2)}}{\sqrt{n}} \leq L$  will be sufficient.

That is  $n$  must be chosen  $n$  so that  $\sqrt{n} \geq \frac{z_{1-(\alpha/2)}}{L}$ . Upon squaring both sides, we see that  $n$  must be such that

$$n \geq \left( \frac{z_{1-(\alpha/2)}}{L} \right)^2$$

► **EXAMPLE 4.3.7** A market research firm is interested in determining the proportion of households that are watching a particular sporting event. To accomplish this task, it plans on using a telephone poll of randomly chosen households. How large a sample is needed if the company wants to be 90% certain that its estimate is correct within an interval of maximum length of 3%.

**Solution:** For the 90% certain, we have  $z_{1-(\alpha/2)} = z_{0.95} = 1.645$  and  $L = 0.03$ , then the sample size needed to be 90% certain that its estimate is correct within an interval of maximum length of 3% is given by

$$n \geq \left( \frac{z_{1-(\alpha/2)}}{L} \right)^2 = \left( \frac{1.645}{0.03} \right)^2 = 3006.69$$

This show that the company need at least a sample of size 3007 to conduct this study. ◀

► **EXAMPLE 4.3.8** A geographer is asked to determine the sample size necessary to estimate the proportion of residents of a city who are in favor of declaring the city a nuclear-free zone. The estimate must not differ from the true proportion by more than 0.05 with a 95% confidence level. How large a sample should be taken? at 99%?

**Solution:** We take a conservative approach and assume  $p = 0.5$  in order to produce the widest possible interval.  $\delta_p$  is given as 0.05. For the 95% confidence interval, we see  $z_{1-(\alpha/2)} = z_{0.975} = 1.96$  and thus:

$$n = (1.96 \sqrt{0.5(1 - 0.5)} / 0.05)^2 = 384.16$$

## SECTION 4.3 ESTIMATION OF THE POPULATION PROPORTION

For the 99% confidence interval we use  $z_{1-(\alpha/2)} = z_{0.95} = 2.575$  and get:

$$n = (2.575\sqrt{0.5(1 - 0.5) / 0.05})^2 = 665.06$$





## Section 4.4

# INTRODUCTION TO HYPOTHESES TESTING

One of a statistician's most important jobs is to draw inference about the populations based on the samples taken from the populations. Most of the statistical inference centers around the population parameters of a population, for example, the mean or proportion. One of the two approach of the statistical inference is to make a decision concerning the value of the parameter. Decisions concerning the value of a parameter are obtained by hypotheses testing, the topic we shall study in this chapter.

Students often ask which method should be used on a particular problem that is, should the parameter be estimated, or should we test a hypothesis involving the parameter? The answer lies in the practical nature of the problem and the questions posed about it.

### SETUP HYPOTHESES

The first step in the hypotheses testing approach is to establish a working hypothesis about the underlying parameter. This hypothesis is called the null hypothesis, denoted by the symbol  $H_0$ . The value of the null hypothesis is often a historical value, a claim, a standard value, contract specification, medical fact, Key Performance Indicator (KPI) or a production specification. For example, if the average height of a professional male basketball player was 6.5 feet 10 years ago, we might use a null hypothesis  $H_0 : \mu = 6.5$  feet for a study involving the average height of this year's professional male basketball players. If television networks claim that the average length of time devoted to commercials in a 60-minute program is 12 minutes, we would use  $H_0 : \mu = 12$  minutes as our null hypothesis in a study regarding the average length of time devoted to commercials. Finally, if a car change oil satiation claims that it should take an average of 25 minutes to change the oil of a car, we would use  $H_0 : \mu = 25$  minutes as the null hypothesis for a study of how long the service time of the station is conforming to specify average times for oil changing. Any hypothesis that differs from the null hypothesis is called an alternate hypothesis. An alternate hypothesis is constructed in such a way that it is the one to be accepted when the null hypothesis must be rejected. The alternate hypothesis is denoted by the symbol  $H_1$ . For example, if we believe the average height of professional male basketball players is taller than it was 10 years ago, we would use an alternate hypothesis  $H_1 : \mu > 6.5$  feet with the null hypothesis  $H_0 : \mu = 6.5$  feet.

## SECTION 4.4 INTRODUCTION TO HYPOTHESES TESTING

### DEFINITION 4.4.1 (Statistical Hypothesis)

Statistical hypothesis is an argument about a specific statistical question, and this argument can be true or wrong.

In another words, one can say: A hypothesis is a theory, based on insufficient evidence that lends itself to further testing and experimentation.

### DEFINITION 4.4.2 (Null Hypothesis)

The null hypothesis is a statement under investigation or testing.

Usually the null hypothesis represents a statement of “no effect,” “no difference,” or, put another way, “things haven’t changed.”

### DEFINITION 4.4.3 (Alternate Hypothesis)

The **alternate hypothesis** is a statement we will adopt in the situation in which the evidence (data) is so strong that you reject the null hypothesis.

This test is a statistical test is designed to assess the strength of the evidence (data) against the null hypothesis.

► **EXAMPLE 4.4.1** A car manufacturer advertises that its new models save more gas and get 47 miles per gallon (mpg). Let  $\mu$  be the mean of the mileage distribution for these cars. You assume that the manufacturer will not underrate the car, but you suspect that the mileage might be overrated.

- a. What shall we use for  $H_0$ ?
- b. What shall we use for  $H_1$ ?

**Solution:**

**For a)** We want to see if the manufacturer’s claim that  $\mu = 47$  mpg can be rejected. Therefore, our null hypothesis is simply that  $\mu = 47$  mpg. We denote the null hypothesis as  $H_0 : \mu = 47$  mpg.

**For b)** From experience with this manufacturer, we have every reason to believe that the advertised mileage is too high. If  $\mu$  is not 47 mpg, we are sure it is less than 47 mpg. Therefore, the alternate hypothesis is  $H_1 : \mu < 47$  mpg. ◀

### DEFINITION 4.4.4 (Test Statistic)

A test statistic is a quantity calculated from the given sample and can be used to make a decision in a test of hypotheses.

Some example of test statistics used in this section are summarized below:

**Table 4.4.1**

Different test statistics for population mean hypotheses testing

Case	Population	Sample size	Standard deviation	Statistic and distribution
1	normal	any sample size	known	$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$
2	any population	large ( $n \geq 30$ )	known	$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$
3	any population	large ( $n \geq 30$ )	unknown, use the sample standard deviation instead of	$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim N(0,1)$

**ERRORS AND LEVEL OF SIGNIFICANCE**

When we reject the true null hypothesis, in this case Type I error is occurred with probability  $\alpha$ , while accepting the false null hypothesis causes Type II error with probability  $\beta$ .

**Table 4.4.2**

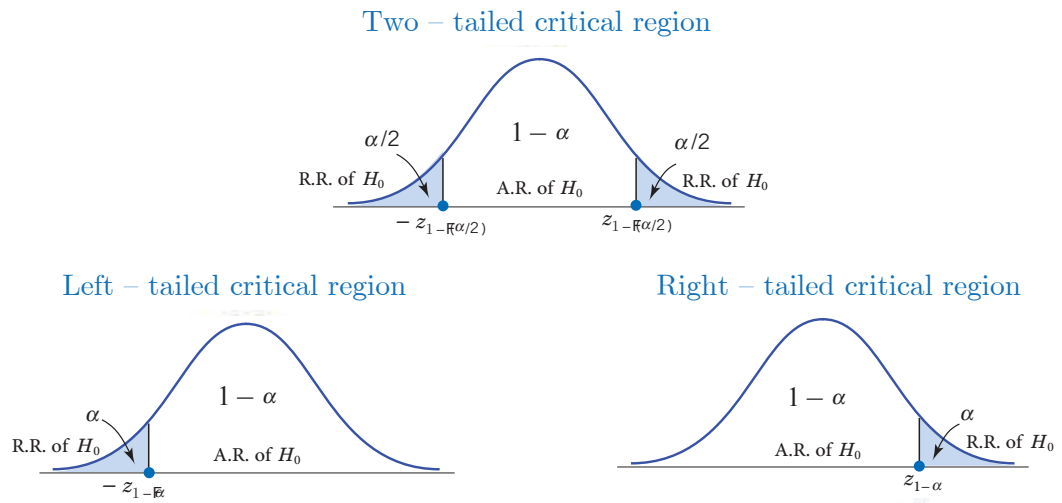
		Null hypothesis	
		Accepted	Rejected
Null hypothesis	True	Correct decision	$P(\text{Type I error}) = \text{Level of Significance} = \alpha$
	False	$P(\text{Type II error}) = \beta$	Correct decision

**DEFINITION 4.4.5 (The Critical Regions)**

The critical region is a region that produced by the value(s) that corresponds to the rejection of the null hypothesis at some chosen level of significance.

The shaded area under the standard normal distribution curve is equal to the level of significance. The critical values are tabulated and thus obtained from the standard normal table. If the absolute value of the statistic is larger than the tabulated value, then statistic value is in the critical region.

The statistical tests use one tailed or two tailed critical regions depending on the nature of the null hypothesis and the alternative hypothesis. The possible critical regions are the parts of the real axis under the shaded areas (see Figure 4.4.1 below):



**Figure 4.4.1** (The one and two tailed critical regions)

R.R: Rejection region and A.R.: Acceptance region.

## Section 4.5

# HYPOTHESES TESTING FOR THE POPULATION MEAN

Below we will take  $\Omega$  as a sample space of a population with mean  $\mu$  and standard deviation  $\sigma$ , and  $\mathcal{X} = (X_1, X_2, \dots, X_n)$  is a simple random sample of  $\Omega$ .

In this section, we consider the hypotheses testing for the population mean, we consider some different cases; they are

- i. Normal population with known variance
- ii. Non-normal population with known variance and large sample size
- iii. Normal or non-normal populations with unknown variance and large sample size

For the cases (i) and (ii), we use  $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ , which is  $N(0,1)$  distributed, while for the case (iii), we use  $Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ , which is approximately  $N(0,1)$  distributed (if the size of sample  $n$  greater than 30) by the central limit theorem. In view of the previous discussion and definition, we may list the test procedure steps that will frequent be used for testing the population mean in the mentioned cases.

### STEPS IN HYPOTHESES TESTING OF POPULATION MEAN

The following steps summarize the hypotheses testing of the population mean:

1. **Identify the null hypothesis  $H_0$  and the alternative (alternate) hypothesis  $H_1$ :**

These will often be conjectures (or suspicions or beliefs) concerning the population mean  $\mu$ . As a rule, the null hypothesis will usually contain an equal sign. The alternate hypothesis will usually contain the symbol  $>$ ,  $<$ , or  $\neq$ , it depends on the situation to be tested. This can be written as

$$H_0 : \mu = \text{spicified value}(\mu_0) \quad \text{versus} \quad H_1 : \begin{cases} \mu > \text{spicified value}(\mu_0), \\ \mu < \text{spicified value}(\mu_0), \\ \mu \neq \text{spicified value}(\mu_0). \end{cases}$$

In reality, we use only one case of the different cases of the alternate hypotheses  $H_1$  depending on the situation to be tested.

**2. Select the test statistic and determine its value under the null hypothesis  $H_0$ :**

In this step, select the test statistic depending on the patent population of the study. The statistic can be calculated using the available random sample and the specified value of the population mean under the null hypothesis  $H_0$ . The calculated value of the test statistic is called the observed value and denoted by  $z_0$  and given by one of the following cases:

$$z_0 = \begin{cases} \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} & ; \text{ when } \sigma \text{ known} \\ \frac{\bar{x} - \mu_0}{s / \sqrt{n}} & ; \text{ when } \sigma \text{ unknown and } n \geq 30 \end{cases}$$

**3. Determine the critical region:**

The critical region consists of those values of the test statistic that strongly favor the alternate hypothesis  $H_1$ . The actual size of the critical region depends on the level of significance  $\alpha$ . This is because the critical region is chosen in such way that the probability will be  $\alpha$  that the test statistic will fall in the critical region (if  $H_0$  were true). The following graphs show three different type of critical region as shaded according the type of the test based on the nature of  $H_1$ :

- i. two-tailed test ( $H_1 : \mu \neq \mu_0$ ): this case can be used when you believe that the population mean  $\mu$  is different from the stated value  $\mu_0$  (pre-specified value of the mean).

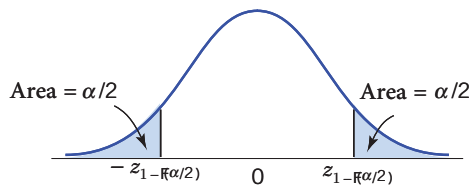


Figure 4.5.1-a

- ii. right-tailed test ( $H_1 : \mu > \mu_0$ )

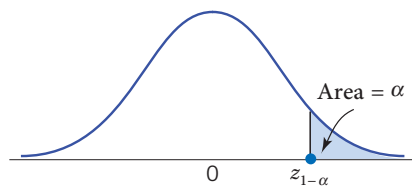


Figure 4.5.1-b

- iii. left-tailed test ( $H_1 : \mu < \mu_0$ )

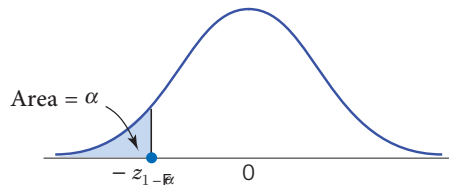


Figure 4.5.1.c

**4. Decision:**

If the test statistic falls in the critical region, we reject  $H_0$ . When this occurs, some statisticians say the results are statistically significant, also giving the  $\alpha$  level. If the test statistic does not fall in the critical region, we fail to reject  $H_0$ . That is, we conclude that there is not enough evidence to reject  $H_0$ . You should interpret your decision in ordinary, nontechnical language.

**P -VALE APPROACH FOR HYPOTHESES TESTING OF POPULATION MEAN**

**DEFINITION 4.5.1 (p-value)**  
 Assuming the null hypothesis  $H_0$  is true, the probability that the test statistic will take on values as extreme as or more extreme than the observed test statistic (computed from sample data) is called the  $p$ -value of the test.

**REMARK 4.5.1**

The smaller  $p$ -value computed from sample data, the stronger evidence against the null hypothesis.

The  $p$ -value approach can be used as an alternative way to conduct the hypotheses testing. This approach is summarized as the following steps.

1. **Identify the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$  :**

As before we can write the hypotheses as

$$H_0 : \mu = \text{spicified value}(\mu_0) \quad \text{versus} \quad H_1 : \begin{cases} \mu > \text{spicified value}(\mu_0), \\ \mu < \text{spicified value}(\mu_0), \\ \mu \neq \text{spicified value}(\mu_0). \end{cases}$$

2. **Select the test statistic and determine its value under the null hypothesis  $H_0$  :**

The calculated value of the test statistic is called the observed value and denoted by  $z_0$  and given by one of the following cases

$$z_0 = \begin{cases} \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} & ; \text{ when } \sigma \text{ known} \\ \frac{\bar{x} - \mu_0}{s/\sqrt{n}} & ; \text{ when } \sigma \text{ unknown and } n \geq 30 \end{cases}$$

**3. Determine  $p$ -value:**

The  $p$ -value can be calculated for the different cases of the tests (right-tailed, left-tailed and two-tailed) as:

- a. For two-tailed test ( $H_1 : \mu \neq \mu_0$ ) we have:  $p\text{-value} = 2P(Z > |z_0|)$

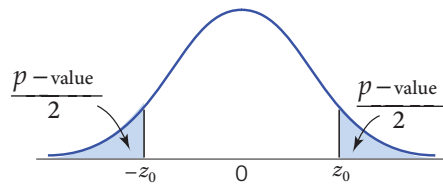


Figure 4.5.2-a

- b. For right-tailed test ( $H_1 : \mu > \mu_0$ ) we have:  $p\text{-value} = P(Z > z_0)$

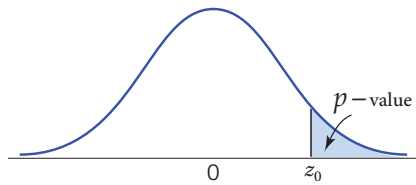


Figure 4.5.2-b

- c. For left-tailed test ( $H_1 : \mu < \mu_0$ ) we have:  $p\text{-value} = P(Z < -|z_0|)$

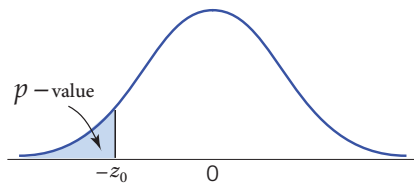


Figure 4.5.2-c

**4. Decision:**

The decision rule is to reject  $H_0$  when  $p\text{-value} \leq \alpha$ , otherwise, do not reject  $H_0$ . Then give a simple explanation of your conclusion in the context of the applications.

► **EXAMPLE 4.5.1** Suppose we would like to determine if the typical amount spent per customer for dinner at a new restaurant in a town is more than \$20.00. A sample of 49 customers over a three-week period was randomly selected and the average amount spent was



\$22.60. Assume that the standard deviation is known to be \$2.50. Using a 0.02 level of significance, would we conclude the typical amount spent per customer is more than \$20.00?

**Solution:** In this example, there is no need to state that the population is normal because the sample size is  $n = 49$  (more than 30) with average  $\bar{x} = 22.6$  dollars. The standard derivation in this case is known to be  $\sigma = 2.5$  dollars. Then, we follow the usual test steps.

**1. Identifying the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$ :**

$$H_0 : \mu = 20 \quad \text{versus} \quad H_1 : \mu > 20$$

**2. Selecting the test statistic and determine its value under the null hypothesis  $H_0$ :**

In this case we use the test statistic as

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \Rightarrow z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{22.6 - 20}{2.5 / \sqrt{49}} = 7.28$$

**3. Determine the critical region:**

The test in this case is right-tailed ( $H_1 : \mu > 20$ ), then critical region is the part of the real axis under the shaded area, which determined by  $z_{1-\alpha} = z_{0.98} = 2.05$  (see below):

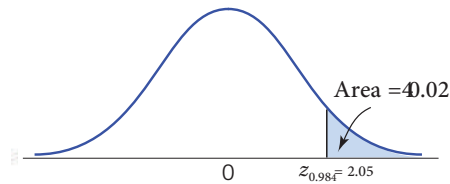


Figure 4.5.3

**4. Decision:**

The test statistic falls in the critical region, so we reject  $H_0$  at the significance level of  $\alpha$ . Then, we conclude the typical amount spent per customer is more than \$20.

One may use the  $p$ -value approach, since the test is right-tailed ( $H_1 : \mu > 20$ ), then:

$$p\text{-value} = P(Z > z_0) = 1 - P(Z \leq 7.28) \approx 1 - 1 = 0 < \alpha = 0.02$$

So we reject  $H_0$ . There is sufficient evidence to conclude the typical amount spent per customer is more than \$20 at  $\alpha = 0.02$ .

► **EXAMPLE 4.5.2** Suppose an editor of a publishing company claims that the mean time to write a textbook is at most 15 months. A sample of 16 textbook authors is randomly selected and it is found that the mean time taken by them to write a textbook was 12.5

## SECTION 4.5 HYPOTHESES TESTING FOR THE POPULATION MEAN

months. Assume also that the standard deviation is known to be 3.6 months and the time to write a textbook is normally distributed and using a 0.025 level of significance. Would you conclude the editor's claim is true?

**Solution:** In this example, there population in is normal with known standard deviation  $\sigma = 3.5$ . The sample has a size of 16 with average  $\bar{x} = 13.5$  months. Then, we follow the usual test steps.

1. **Identifying the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$ :**

$$H_0 : \mu = 15 \quad \text{versus} \quad H_1 : \mu < 15$$

2. **Selecting the test statistic and determine its value under the null hypothesis  $H_0$ :**

In this case we use the test statistic as

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{13.5 - 15}{3.6 / \sqrt{16}} = -1.67$$

3. **Determine the critical region:**

The test in this case is left-tailed ( $H_1 : \mu < 15$ ), then critical region is the part of the real axis under the shaded area, which determined by  $-z_{1-\alpha} = -z_{0.975} = -1.96$  (see below):

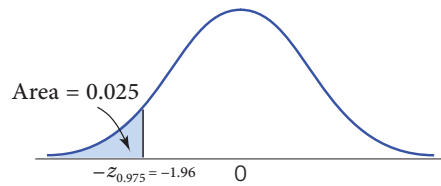


Figure 4.5.4

4. **Decision:**

The test statistic falls in the acceptance region, so we don't reject  $H_0$  at the significance level of  $\alpha$  (this means that we do not have information from the sample, that drives us to reject it). Then, there is sufficient evidence to conclude the editor's claim is true at  $\alpha = 0.025$ .

By using the  $p$ -value approach, we have a right-tailed test ( $H_1 : \mu < 15$ ), then:

$$p\text{-value} = P(Z < - | z_0 |) = P(Z < -1.67) = 0.0475 > \alpha = 0.025$$

So we accept  $H_0$ .

► **EXAMPLE 4.5.3** Test the claim that on average there are three TV sets in each U.S. home. Assume you know that the population standard deviation is 1 set. You have collected a

random sample of 100 households and found the average to be 3.2 sets. Can you conclude that this claim is false at  $\alpha = 0.05$ ?

**Solution:** In this example, there population need not to be normal since the sample size in 100. The population standard deviation  $\sigma = 1$  set. The sample average  $\bar{x} = 3.2$  sets. Then, we follow the usual test steps.

**1. Identifying the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$ :**

$$H_0 : \mu = 3 \quad \text{versus} \quad H_1 : \mu \neq 3.$$

**2. Selecting the test statistic and determine its value under the null hypothesis  $H_0$ :**

In this case we use the test statistic as

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{3.2 - 3}{1 / \sqrt{100}} = 2.0$$

**3. Determine the critical region:**

The test in this case is two-tailed ( $H_1 : \mu \neq 3$ ), then the critical regions are the parts of the real axis under the shaded areas, which determined by  $\pm z_{1-(\alpha/2)} = \pm z_{0.975} = \pm 1.96$  (see below):

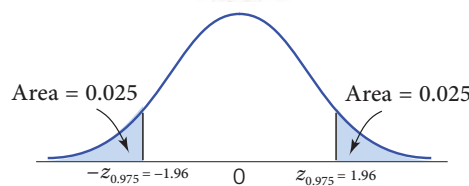


Figure 4.5.5

**4. Decision:**

The test statistic falls in the critical region, so we reject  $H_0$  at the significance level of  $\alpha$ . Then, we do have enough evidence to conclude that the average number of TV sets in U.S. homes differs from three,  $\alpha = 0.025$ .

By using the  $p$ -value approach, we have a two-tailed test ( $H_1 : \mu \neq 3$ ), then:

$$p\text{-value} = 2 P(Z > | z_0 |) = 2[1 - P(Z \leq 2.0)] \approx 2(1 - 0.9772) = 0.0456 < \alpha = 0.05$$

So we reject  $H_0$ . We do have enough evidence to conclude that the average number of TV sets in U.S. homes differs from three sets at  $\alpha = 0.05$ .

► **EXAMPLE 4.5.4** A nutritionist believes that a 12-ounce box of breakfast cereal should contain an average of 1.2 ounces of bran. The nutritionist measures a random sample of 60 boxes of popular cereal for bran content. Suppose the data yield  $\bar{x} = 1.170$ , and  $s = 0.111$ ,

## SECTION 4.5 HYPOTHESES TESTING FOR THE POPULATION MEAN

do the data indicate that the mean bran content of all boxes of this brand of cereal differs from 1.2 ounces? Use 5% level of significance.

### Solution

The population in this example need not to be normal since the sample size is 60. The sample mean and standard deviation are  $\bar{x} = 1.17$  and  $s = 0.111$ , respectively. Then, we follow the usual test steps.

1. **Identifying the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$ :**

$$H_0 : \mu = 1.2 \quad \text{versus} \quad H_1 : \mu \neq 1.2$$

2. **Selecting the test statistic and determine its value under the null hypothesis  $H_0$ :**

In this case we use the test statistic as

$$z_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{1.17 - 1.2}{0.111 / \sqrt{60}} = -2.094$$

3. **Determine the critical region:**

The test in this case is two-tailed ( $H_1 : \mu \neq 1.2$ ), then the critical regions are the parts of the real axis under the shaded areas, which are determined by  $\pm z_{1-(\alpha/2)} = \pm z_{0.975} = \pm 1.96$  (see below):

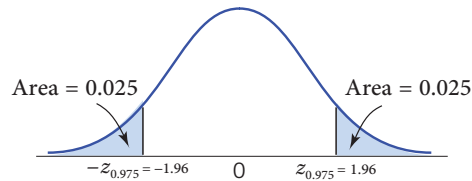


Figure 4.5.6

4. **Decision:**

The test statistic falls in the critical region, so we reject  $H_0$  at the significance level of  $\alpha$ . Then, we do not have enough evidence to conclude that a 12-ounce box of breakfast cereal should contain an average of 1.2 ounces of bran.

One may use the p-value approach, since the test is two-tailed ( $H_1 : \mu \neq 3$ ), then:

$$p\text{-value} = 2P(Z > |z_0|) = 2[1 - P(Z \leq 2.094)] \approx 2(1 - 0.9817) = 0.0366 < \alpha = 0.05$$

So we reject  $H_0$ . So, we do not have enough evidence to conclude that a 12-ounce box of breakfast cereal should contain an average of 1.2 ounces of bran at 5% level of significance. ◀

## Section 4.6

# HYPOTHESES TESTING FOR THE POPULATION PROPORTION

In this section, we consider the tests concerning the proportion of members of a population that possess a certain characteristic. We suppose that the population is very large (in theory, of infinite size), and we let  $p$  denote the unknown proportion of the population with the characteristic. We will be interested in testing the null hypothesis, that is the population proportion  $p$  is equal some specified value  $p_0$ . Proceed as the population mean, we follow four steps as:

**1. Identify the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$**

Similar as the population mean we can formulate the null hypothesis and the alternate hypothesis about the population proportion as:

$$H_0 : p = \text{spicified value}(p_0) \quad \text{versus} \quad H_1 : \begin{cases} p > \text{spicified value}(p_0), \\ p < \text{spicified value}(p_0), \\ p \neq \text{spicified value}(p_0). \end{cases}$$

**2. Select the test statistic and determine its value under the null hypothesis  $H_0$**

The statistic can be calculated using the available random sample and the specified value of the population proportion under the null hypothesis  $H_0$ . The calculated value of the test statistic is called the observed value and denoted by  $z_0$  similar to the population mean and given by:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

As, we have seen before, this statistic follows  $N(0,1)$ .

**3. Determine the critical region:**

Similar to the critical regions in the population mean, we have three cases.

- i. two-tailed test ( $H_1 : p \neq p_0$ ): this case can be used when you believe that the population proportion  $p$  is different from the stated value  $p_0$ .

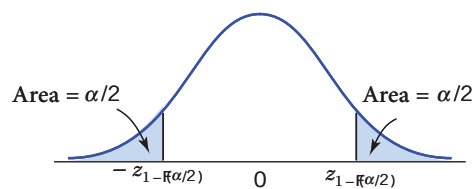


Figure 4.6.1-a

- ii. right-tailed test ( $H_1 : p > p_0$ )

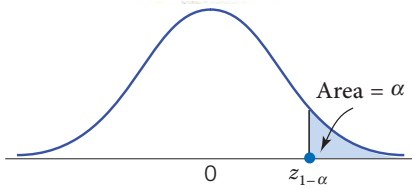


Figure 4.6.1-b

- iii. left-tailed test ( $H_1 : p < p_0$ )

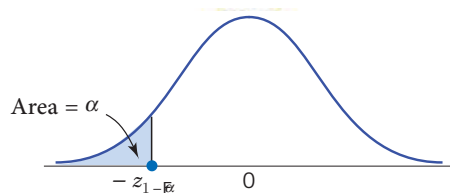


Figure 4.6.1-c

#### 4. Decision:

Again as the population mean hypotheses testing, if the test statistic falls in the critical region, reject  $H_0$ , otherwise, we do not have enough evidence to reject  $H_0$ . Finally, you should interpret your decision in ordinary, nontechnical language.

### P-VALUE APPROACH FOR HYPOTHESES TESTING OF POPULATION MEAN

The  $p$ -value approach can be used as an alternative way to conduct the hypotheses testing. The approach works similar to the population mean case, except replace the test statistic corresponding to the population proportion.

► **EXAMPLE 4.6.1** A computer chips manufacturer claims that at most 2 percent of the produced chips are defective. An electronics company, impressed by that claim, has purchased a large quantity of chips. To determine if the manufacturer's claim is plausible, the company has decided to test a sample of 400 of these chips. If there are 13 defective chips (3.25 percent) among these 400, does this disprove (at the 5 percent level of significance) the manufacturer's claim?

**Solution:** In this example we have:

$$p_0 = 0.02, \quad n = 400, \quad \hat{p} = \frac{13}{400} = 0.0325, \quad \alpha = 0.05 \quad \text{and} \quad z_{1-(\alpha/2)} = z_{0.95} = 1.645$$

Following the test steps described before, we have

1. Identifying the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$  :

$$H_0 : p = 0.02 \quad \text{versus} \quad H_1 : p > 0.02$$

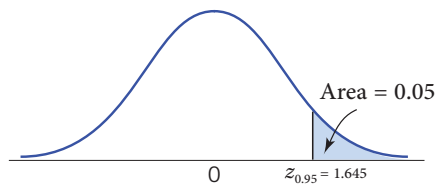
**2. Selecting the test statistic and determine its value under the null hypothesis  $H_0$ :**

The statistic can be calculated using the available random sample and the specified value of the population proportion under the null hypothesis  $H_0$  as:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.0325 - 0.02}{\sqrt{\frac{0.02(0.98)}{400}}} = 1.79$$

**3. The critical region:**

The test is right-tailed ( $H_1 : p > 0.02$ ), then the critical region is the part of the real axis under the shaded area (see below):



**Figure 4.6.2**

**4. Decision:**

The test statistic falls in the critical region, reject  $H_0$ , then the test does not support the manufacturer's claim at 5% level of significant.

On the other hand:

$$p\text{-value} = P(Z > 1.79) = 1 - P(Z \leq 1.79) = 1 - 0.9633 = 0.0367 < \alpha = 0.05$$

Then reject  $H_0$  which is also, does not support the manufacturer's claim at 5% level of significant.

► **EXAMPLE 4.6.2** Historical data indicate that 4 percent of the components produced at a certain manufacturing facility are defective. A particularly acrimonious labor dispute has recently been concluded, and management is curious about whether it will result in any change in this figure of 4 percent. If a random sample of 500 items indicated 16 defectives (3.2 percent), is this significant evidence, at the 5 percent level of significance, to conclude that a change has occurred?

**Solution**

In this example  $p_0 = 0.04$ ,  $n = 500$ ,  $\hat{p} = \frac{16}{500} = 0.032$ ,  $\alpha = 0.05$ , then

$$z_{1-(\alpha/2)} = z_{0.975} = 1.96,$$

following the test steps described before, we have:

## SECTION 4.6 HYPOTHESES TESTING FOR THE POPULATION PROPORTION

1. Identifying the null hypothesis  $H_0$  and the alternate hypothesis  $H_1$  :

$$H_0 : p = 0.04 \quad \text{versus} \quad H_1 : p \neq 0.04$$

2. Selecting the test statistic and determine its value under the null hypothesis  $H_0$  :

The statistic can be calculated using the available random sample and the specified value of the population proportion under the null hypothesis  $H_0$  as

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.032 - 0.04}{\sqrt{\frac{0.04(0.96)}{500}}} = -0.913$$

3. **The critical region:** the test in this case is two-tailed ( $H_1 : p \neq 0.04$ ), then the critical regions are the parts of the real axis under the shaded areas, which determined by  $\pm z_{1-(\alpha/2)} = \pm z_{0.975} = \pm 1.96$  (see below):

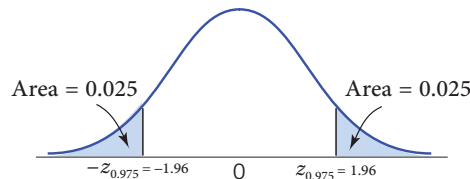


Figure 4.6.3

4. **Decision:**

The test statistic does not fall in the critical region, so, we cannot reject  $H_0$ , this is not significant evidence, at the 5% level of significance, to conclude that a change has occurred.

On the other hand:

$$p - \text{value} = 2P(Z > | -0.913 |) = 2[1 - P(Z \leq 0.913)] = 2(1 - 0.8186) = 0.3628 > \alpha = 0.05$$

This is not significant evidence, at the 5% level of significance, to conclude that a change has occurred.







## EXERCISES



1. The circumference of a certain type of palm trees in Saudi Arabia farms follows normal distribution with mean 100 cm and standard deviation 5. Find the proportion of trees with circumference less than 95 cm.
2. It was found that the cholesterol level in a population of 14-years school boys is approximately normal distribution with mean 160 and standard deviation 25.
  - a. What is the form of the pdf for the r.v  $X$ ?
  - b. What proportion of 14-year-old boys has cholesterol level between 120 and 200?
  - c. What proportion of 14-year-old boys has cholesterol level greater than 200?
  - d. What cholesterol level for the boys who have level in the highest 75%?
3. Lifetimes of a certain brand of cars tires is approximately normally distributed with mean 42500 miles and standard deviation 3200 miles.
  - a. What is the probability that the life time is greater than 50000 miles?
  - b. What ranges do the middle 90% of tire lifetimes?
  - c. What is the probability that  $P(X < 3200)$ .
  - d. For a population 2000 tires what is the number of tires whose lifetimes between 30000 miles and 45000 miles.
4. A new brand of milk is being market tested. It is estimated that 70% of consumers like the new milk. A sample of 108 taste-tested the new milk. Now find the following
  - a. The standard error of the proportion.
  - b. The probability that equal to or more than 75% of consumers will like the milk.
  - c. The probability that equal to or more than 40% of consumers will not like the milk.
5. A random sample is drawn from a population of known standard deviation 11.3, construct a 90% confidence interval for the population mean based on the information given
  - a.  $n = 36$ ,  $\bar{x} = 105.2$  and  $s = 11.2$
  - b.  $n = 100$ ,  $\bar{x} = 105.2$  and  $s = 11.2$
  - c. Compare between the results in (a) and (b).
6. A random sample is drawn from a population of known standard deviation 22.1. Construct a confidence interval for the population mean based on  $n = 36$ ,  $\bar{x} = 182.4$  and  $\sigma = 9$ , when

- 
- a.  $1 - \alpha = 0.9$
    - b.  $1 - \alpha = 0.95$
    - c. Compare between the results in (a) and (b).
  7. A random sample of 85 group leaders, supervisors, and similar personnel revealed that a person spent an average 6.5 years on the job before being promoted. The population standard deviation was 1.7 years. Using the 0.95 confidence level, what is the confidence interval for the population mean?
  8. The yields in metric tons per hectare of potatoes in a randomly selected sample of 10 farms in a small region are 32.1, 34.4, 34.9, 30.6, 38.4, 29.4, 28.9, 32.6, 32.9, and 44.9. Assuming that these yields are normally distributed with standard deviation of 4.77. Determine a 99% confidence interval on the population mean yield.
  9. Past experience shows that the standard deviation of the distances traveled by consumers to patronize a “big-box” retail store is 4 km. Adopting an error probability of 0.05, how large a sample is needed to estimate the population mean distance traveled to within 0.5 km? 1 km? 5 km?
  10. An engineering firm manufactures a space rocket component that will function for a length of time that is normally distributed with a standard deviation of 3.4 hours. If a random sample of nine such components has an average life of 10.8 hours, find a 95% and 99% confidence intervals estimate of the mean length of time that these components function.
  11. To estimate  $p$ , the proportion of all newborn babies who are male, the gender of 10,000 newborn babies was noted. If 5106 were male, determine a 90% and 99% confidence interval estimate of  $p$ .
  12. A researcher wishes to estimate the proportion of all adults who own a cell phone. He takes a random sample of 1572 adults; 1298 of them own a cell phone.
    - a. What is the point estimation of population of interest?
    - b. Construct 90% confidence interval of the population proportion.
  13. The mean number of travel days per year for salespeople employed by hardware distributors needs to be estimated with 90% level of confidence. For a small pilot study, the mean was 150 days and the standard deviation was 14 days. If the population mean is estimated within margin error of two days, how many salespeople should be sampled?

14. A random sample of 50 households is taken, and 38 automobile commuters are found. Determine the 95% confidence interval of the proportion of commuters by automobile in the neighborhood.
15. A random sample of medical files is used to estimate the proportion  $p$  of all people who have blood type  $B$ . If you have no preliminary estimate for  $p$ , how many medical files should you include in a random sample in order to be 85% sure that the point estimate will be within a distance of 0.05 from  $p$ ?
16. Suppose a production line operates with a mean filling weight of 16 ounces per container. Since over- or under-filling can be dangerous, a quality control inspector samples 30 items to determine whether or not the filling weight has to be adjusted. The sample revealed a mean of 16.32 ounces. From past data, the standard deviation is known to be .8 ounces. Using a 0.10 level of significance, can it be concluded that the process is out of control (not equal to 16 ounces)?
17. In the past the average waist size of adult males in a town has been 36 inches with a standard deviation of 3 inches. You wish to determine if the average waist size of males in a town is now greater than 36. You collect a random sample of 36 men and determine that the average waist size is 37.5 inches. Use  $\alpha = 0.01$
18. A tire manufacturing plant produces 15.2 tires per hour. This yield has an established variance of 2.5 ( $\sigma = 1.58$  tires/hour). New machines are recommended, but will be expensive to install. Before deciding to implement the change, 12 new machines are tested. They produce 16.8 tire per hour. Is it worth buying the new machines? Use 2% level of significance.
19. A bus company advertised a mean time of 150 minutes for a trip between two cities. A consumer group had reason to believe that the mean time was more than 150 minutes. A sample of 40 trips showed a mean  $\bar{x} = 153$  minutes and a standard deviation  $s = 7.5$  minutes. At the .05 level of significance, test the consumer group's belief.
20. An environmentalist collects a liter of water from 45 different locations along the banks of a stream. He measures the amount of dissolved oxygen in each specimen. The mean oxygen level is 4.62 mg, with the overall standard deviation of 0.92. A water purifying company claims that the mean level of oxygen in the water is 5 mg. Conduct a hypothesis test to determine whether the mean oxygen level is less than 5 mg. Use 5% level of significance.

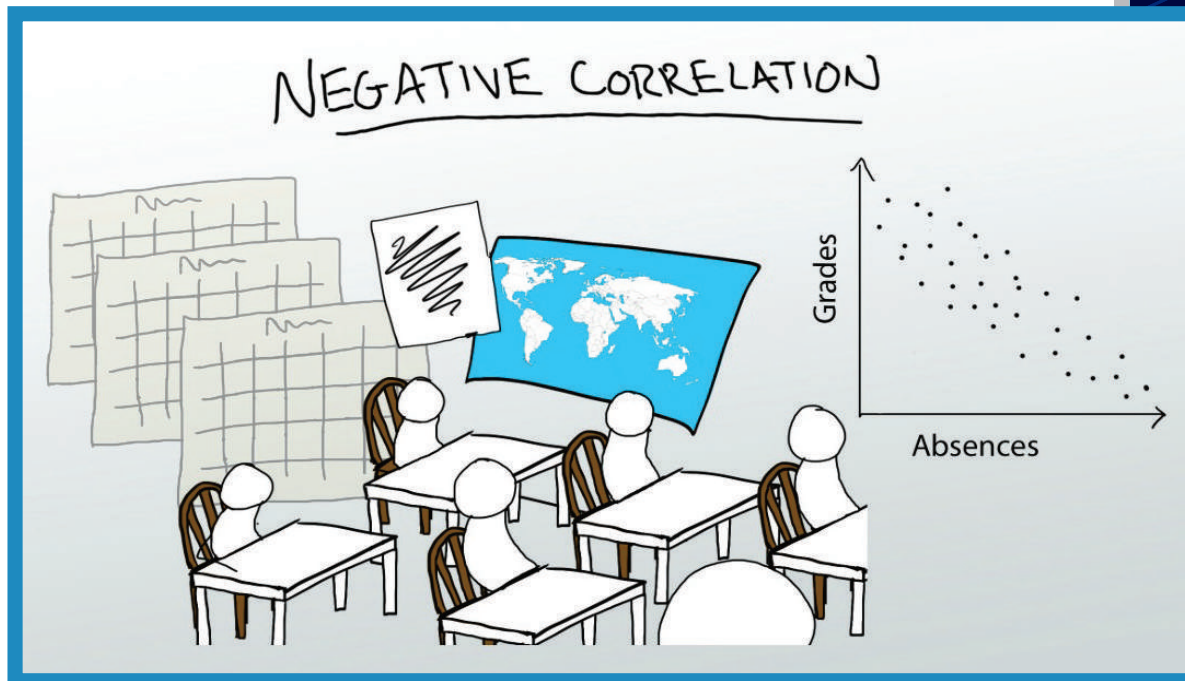
- 
21. Traffic authorities claim that traffic lights are red for a time that is normal with mean 30 seconds and standard deviation 1.4 seconds. To test this claim, a sample of 40 traffic lights was checked. If the average time of the 40 red lights observed was 32.2 seconds, can we conclude, at the 5% level of significance, that the authorities are incorrect? What about at the 1% level of significance?
22. Consider the following hypothesis test:  $H_0 : p = 0.8$  vs  $H_1 : p > 0.8$ . A sample of 400 provided a sample proportion of 0.853.
- Using  $\alpha = 0.05$ , what is the conclusion based on classical hypothesis test?
  - Using  $\alpha = 0.01$ , what is the conclusion based on  $p$ -value?
23. A marketing company claims that it receives 4% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test the company claims at .05 level of significance.
24. A researcher was interested to know about the proportion of females in the population of all patients visiting a certain clinic. The researcher claims that 70% of all patients in this population females. Would you agree with this claim if a random survey s that 24 out of 45 people are females? Use a 0.10 level of significance.
25. In a study on the fear of dental care in a certain city, a survey showed that 60 out of 200 adults said that they would hesitate to take a dental appointment due to fear. Test whether the proportion of adults in this city who hesitate to take dental appointment is less than 0.25. Use a level of significance of .025.
26. An economist thinks that at least 60 percent of recently arrived immigrants who have been working in the health profession in the United States for more than 1 year feel that they are underemployed with respect to their training. Suppose a random sample of size 450 indicated that 294 individuals (65.3 percent) felt they were underemployed. Is this strong enough evidence, at the 5% level of significance, to prove that the economist is correct? What about at the 1% level of significance?
27. Suppose a consumer group suspects that the proportion of households that have three cell phones is not known to be 30%. A cell phone company has reason to believe that the proportion is 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

- 28.** A random survey of 75 death row inmates revealed that the average length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population average time on death row could likely be 15 years.
- 29.** A bank wishes to estimate the mean balances owed by customers holding MasterCard. The population standard deviation is estimated to be \$300. If a 98 percent confidence interval is used and an interval of \$75 is desired, how many cardholders should be sampled?
- 30.** A group of statistics students decided to conduct a survey at their university to find the average (mean) amount of time students spent studying per week.
- Assuming a standard deviation of 6 hours, what is the required sample size if the error should be less than  $\frac{1}{2}$  hour with a 95% level of confidence?
  - Assuming a standard deviation of 3 hours, what is the required sample size if the error should be less than  $\frac{1}{2}$  hour with a 99% level of confidence
- 31.** A survey of 144 retail stores revealed that the average price of a microwave was \$375 with a standard error of \$20.
- What is the 95% confidence interval to estimate the true cost of the microwave?
  - What is the 99% confidence interval to estimate the true cost of the microwave?
  - If 90% and 95% confidence intervals were developed to estimate the true cost of the microwave, what similarities would they have?
  - If 95% and 98% confidence intervals were developed to estimate the true cost of the microwave, what differences would they have?
- 32.** Your company sells exercise clothing and equipment on the Internet. To design the clothing, you collect data on the physical characteristics of your different types of customers. We take a sample of 24 male runners and find their mean weight to be 61.79 kilograms. Assume that the population standard deviation is  $\sigma = 4.5$ .
- Calculate a 95% confidence interval for the mean weight of all such runners.
  - Based on this confidence interval, does a test of
$$H_0 : \mu = 61.3 \text{ kg}$$
$$H_A : \mu \neq 61.3 \text{ kg}$$
Reject  $H_0$  at the 5% significance level?



# CHAPTER 5

## CORRELATION AND REGRESSION



### LEARNING OBJECTIVES

After completing this chapter, you should be able to:

1. Recognize the terms: Correlation, Positive Correlation, Negative Correlation, Simple linear Regression Line, Response Variable, Independent Variable, Predictor Variable, Residuals and the Coefficient of Determination.
2. Regression Line Coefficients: Intercept and Slope.
3. Model parameter estimation: least square method
4. Model interpretation: the meaning of the model parameters.
5. The coefficient of determination and the correlation coefficient.
6. Applications of the simple linear regression model.

- SECTION 5.0 INTRODUCTION
- SECTION 5.1 LINEAR CORRELATION COEFFICIENT
- SECTION 5.2 SIMPLE LINEAR REGRESSION

## Section 5.0

# INTRODUCTION

Statistics is often used to investigate the relationship between two (or more) variables of interest. The following are some examples of relations are often studied:

- Is there a relationship between high school grade and the first year college grade point average (GPA)? If so, what is the relationship?
- What is the relationship between the expenditure and income of a Saudi family?
- What is the relationship between the age and blood pressure?
- The relationship between body mass index and systolic blood pressure, or between hours of exercise per week and percent body fat.

In the above examples, we see that there are two basic questions of interest when investigating a pair of variables:

1. Is there a relationship between the two variables?
2. What is the relationship (if any) between the two variables?

In this chapter, we study these two questions. We study the correlation analysis, which is concerned with the question of whether there is a relationship between variables. We also study regression analysis, where our objective to find a relationship between the variables. This relationship will take the form of an equation relating the two variables. Then a given value of one variable, we can solve for the value of the other variable.

In this part, we list some basic definitions and concepts, they are:

- **Correlation**  
A method used to determine if a relationship between variables exists.
- **Correlation Coefficient**  
A statistic or parameter which measures the strength and direction of a relationship between two variables.
- **Dependent Variable**  
A variable in correlation or regression that cannot be controlled, that is, it depends on the independent variable.



- **Independent Variable**

A variable in correlation or regression which can be controlled, that is, it is independent of the other variable.

- **Pearson Correlation Coefficient**

A measure of the strength and direction of the linear relationship between two variables

- **Regression**

A method used to describe the relationship between two variables.

- **Scatter Plot**

A plot of the data values on a coordinate system. The independent variable is graphed along the  $x$ -axis and the dependent variable along the  $y$ -axis.

- **Coefficient of Determination**

The percent of the variation that can be explained by the regression equation.

## Section 5.1

# LINEAR CORRELATION COEFFICIENT

In this section, we consider the problem of measuring the linear relationship (linear association) between two variables  $X$  and  $Y$ . In the case of studying the correlation (linear association) between two variables  $X$  and  $Y$ , the data may be represented by pairs of observations  $(x_1, y_1)$ ,  $(x_2, y_2)$ , ... and  $(x_n, y_n)$  where  $x_i$  is the value of  $X$  for the  $i^{\text{th}}$  observation,  $y_i$  is the value of  $Y$  for the  $i^{\text{th}}$  observation, and  $n$  is the number of observations.

### SCATTER PLOT

The study of the correlation between  $X$  and  $Y$  variables usually begins with the so-called **scatter plot**, where the scatter plot is an important medium in correlation studies. The purpose of this painting is to take a first impression of the behavior of the mutual influence between the explanatory variable  $X$  and the variable of response  $Y$ .

#### DEFINITION 5.1.1 (Scatter Plot)

Scatter plot is a graph of data, that given in the form of binaries  $(x_1, y_1)$ ,  $(x_2, y_2)$ , ... and  $(x_n, y_n)$ , so that each binary is represented by a point in the coordinate plane  $XoY$  (i.e., we represent the data by points). It is usually we take the orthogonal coordinates in this representation (see the following Figure 5.1.1).

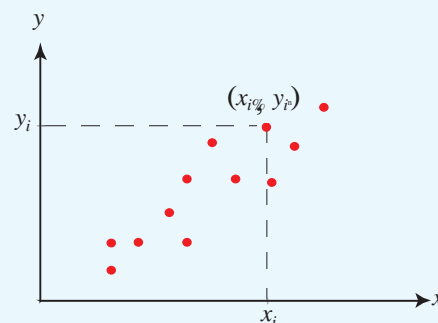


Figure 5.1.1 (Scatter diagram)

### CORRELATION COEFFICIENT

In fact, there are many measures to show the correlation between two phenomena  $X$  and  $Y$ . Below we will present one of these metrics known as "Pearson coefficient of linear correlation".

If you look in different statistics textbooks, you are likely to find different-looking (but equivalent) formulas for computing a correlation coefficient. In this section, we present several formulas that you may encounter. The most common formula for computing a product-moment correlation coefficient ( $r$ ) is given by the following definition.

**DEFINITION 5.1.2 (Pearson's Correlation Coefficient)**

Let  $(x_1, y_1), (x_2, y_2), \dots$  and  $(x_n, y_n)$  be binaries given data. Then the **Pearson's Correlation Coefficient** (or Pearson coefficient of linear correlation) is given by the following relation:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Or using the following relation:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \cdot \left( \sum_{i=1}^n y_i \right)}{\sqrt{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2}}$$

*How to interpret the correlation coefficient?*

The sign and the absolute value of a correlation coefficient ( $r$ ) describe the direction and the magnitude of the relationship between two variables or two phenomena.

- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of the correlation coefficient, the greater the correlation between the two variables (or phenomena).
- The strong linear relationship is indicated by a correlation coefficient, that is close to  $\pm 1$  or equal to  $\pm 1$ , and when the correlation coefficient ( $r$ ) is equal to  $\pm 1$ , then one says that the relationship between two variables is complete linear.
- The weak linear relationship is indicated by a correlation coefficient, that is close to zero or equal to zero, and when the correlation coefficient ( $r$ ) equal to zero, then one says not, that doesn't relationship between two variables, because it is possible that the relationship between the two variables is not linear (see upcoming drawings for models of the correlation).

## SECTION 5.1 LINEAR CORRELATION COEFFICIENT

- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger, i.e. the relationship between the two phenomena is positive monotone.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller, i.e. the relationship between the two phenomena is negative monotone.

The scatterplots below show different patterns of data produce different degrees of correlation.

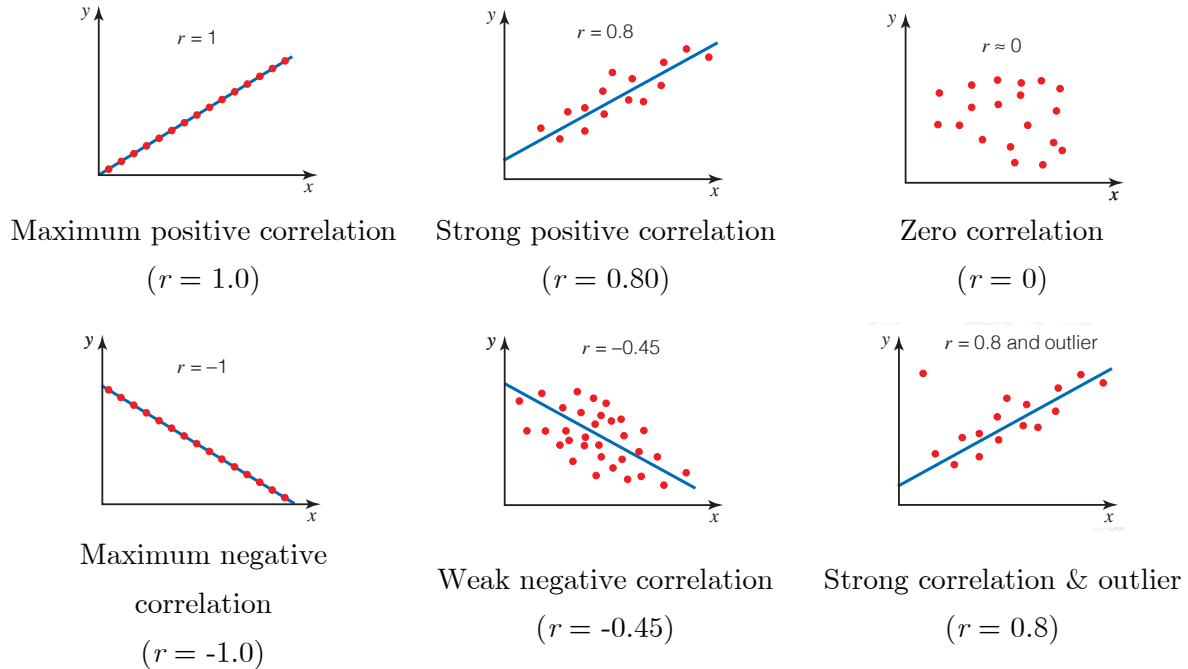


Figure 5.1.2: Different shape of liner correlations

Several points are evident from the scatterplots.

- When the slope of the line in the plot is negative, the correlation is negative; and vice versa.
- The strongest correlations ( $r = 1.0$  and  $r = -1.0$ ) occur when data points fall *exactly* on a straight line.
- The correlation becomes weaker as the data points become more scattered.
- If the data points fall in a random pattern, the correlation is equal to zero.
- Correlation is affected by outliers. Compare the first scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 1.00 to 0.71).

- There are many statistical tests to determine the strength and the significance of the linear relationship between  $X$  and  $Y$ . In general, we might use the following rule to determine the strength of the linear relationship.

**ASSESSMENT OF CORRELATION STRENGTH**

The Relationship between the two variables (or phenomena)	The Range of $r$
Very weak or no linear	$0 \leq  r  \leq 0.30$
Weak (an acceptable degree of linearity)	$0.30 <  r  \leq 0.50$
Moderately strong linear	$0.50 <  r  \leq 0.70$
Strong (the linearity very clear)	$0.70 <  r  \leq 0.86$
Very Strong (high degree of linearity)	$0.86 <  r  < 1$
Complete (all points are located on one straight)	$ r  = 1$

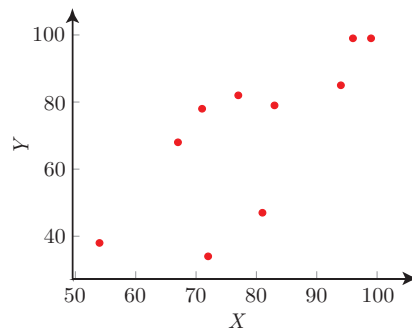
► **EXAMPLE 5.1.1** The results of a class of 10 students on midterm exam marks ( $X$ ) and on the final examination marks ( $Y$ ) are as follows:

The values of $X$	77	54	71	72	81	94	96	99	83	67
The values of $Y$	82	38	78	34	47	85	99	99	79	68

- Construct the scatter diagram.
- Is there a linear relationship (linear association) between  $X$  and  $Y$ ? Is it positive or negative?
- Calculate the sample coefficient of correlation ( $r$ ).

**Solution:** We have:

**For a)** The scatter diagram for the given data is:



**Figure 5.1.3**

**For b)** The scatter diagram suggests that there is a positive linear association between  $X$  and  $Y$  since there is a linear trend for which the value of  $Y$  linearly increases when the value of  $X$  increases.

## SECTION 5.1 LINEAR CORRELATION COEFFICIENT

**For c)** To calculating the coefficient of correlation ( $r$ ) we will create the following table:

**Table 5.1.1**

$i$	$x_i$	$y_i$	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	77	82	-2.4	11.1	5.76	123.21	-26.64
2	54	38	-25.4	-32.9	645.16	1082.41	835.66
3	71	78	-8.4	7.1	70.56	50.41	-59.64
4	72	34	-7.4	-36.9	54.76	1361.61	273.06
5	81	47	1.6	-23.9	2.56	571.21	-38.24
6	94	85	14.6	14.1	213.16	198.81	205.86
7	96	99	16.6	28.1	275.56	789.61	466.46
8	99	99	19.6	28.1	384.16	789.61	550.76
9	83	79	3.6	8.1	12.96	65.61	29.16
10	67	68	-12.4	-2.9	153.76	8.41	35.96
<b>Total</b>	<b>794</b>	<b>709</b>	<b>0</b>	<b>0</b>	<b>1818.4</b>	<b>5040.9</b>	<b>2272.4</b>

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{794}{10} = 79.4, \quad \bar{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{709}{10} = 70.9$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1818.4, \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 5040.9 \quad \text{and} \quad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 2272.4$$

Then the sample coefficient of correlation is:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{2272.4}{\sqrt{1818.4} \sqrt{5040.9}} = 0.75056 \approx 0.75$$

Based on our rule, there is a strong positive linear relationship between  $X$  and  $Y$ . (The values of  $Y$  increase when the values of  $X$  increase).

## Section 5.2

# SIMPLE LINEAR REGRESSION

The statistical use of the word regression dates back to Francis Galton, who studied heredity in the late 1800's. One of Galton's interests was whether or not a man's height as an adult could be predicted by his parents' heights. He discovered that it could, but the relationship was such that very tall parents tended to have children who were shorter than they were, and very short parents tended to have children taller than themselves. He initially described this phenomenon by saying that there was a "reversion to mediocrity" but later changed to the terminology "regression to mediocrity."

The idea behind regression in the social sciences is that the researcher would like to find the relationship between two or more variables. Regression is a statistical technique that allows the scientist to examine the existence and extent of this relationship. Regression shows that given a population, if the researcher can either examine the entire population or perform a random sample of sufficient size, it is possible to mathematically recover the parameters that describe the relationships between variables. Once the researcher has established such a relationship, he/she can then use these parameters to predict values of a new dependent variable given a new independent variable. Regression does not make any specifications about the way that the independent variables are distributed or measured (discrete, continuous, binary, etc.), but in order for regression to be the appropriate technique, some advanced assumptions must be fulfilled.

In its simplest (bivariate) form, regression shows the relationship between one independent variable ( $X$ ) and a dependent variable ( $Y$ ). The magnitude and direction of that relation are given by a parameter ( $b$ ), and an intercept term ( $a$ ) captures the status of the dependent variable when the independent variable is absent. A final error term ( $\varepsilon$ ) captures the amount of variation that is not predicted by the slope and intercept terms. The regression coefficient ( $r$ ) shows how well the values fit the data. More sophisticated forms of regression allow for more independent variables, interactions between the independent variables, and other complexities in the way that one variable affects another.

Regression thus shows us how variation in one variable co-occurs with variation in another. What regression cannot show is causation; causation is only demonstrated analytically, through substantive theory. For example, a regression with shoe size as an independent variable and

## SECTION 5.2 SIMPLE LINEAR REGRESSION

foot size as a dependent variable would show a very high regression coefficient and highly significant parameter estimates, but we should not conclude that higher shoe size causes higher foot size. All that the mathematics can tell us is whether or not they are correlated, and if so, by how much.

### DIFFERENCE BETWEEN CORRELATION AND REGRESSION

It is important to recognize that regression analysis is fundamentally different from ascertaining the correlations among different variables.

- Correlation can tell you how the values of your variables co-vary, but regression analysis is aimed at making a stronger claim: demonstrating how one variable, your independent variable, causes another variable, your dependent variable.
- Correlation determines the strength of the relationship between variables, while regression attempts to describe the relationship between these variables.

Of course, it is apparent that regression may lead to what is called “spurious correlation,” where the co-variation of two variables implies a causal relationship that does not exist. For example, we might find that there is a significant relationship between being a basketball player and being tall. Of course, being a basketball player does not cause one to become taller; the relationship is almost certainly the opposite. It is important to recognize that regression analysis cannot itself establish causation, only describe correlation. Causation is established through theory.

We’ve seen how to explore the relationship between two quantitative variables graphically with a scatterplot. When the relationship has a straight-line pattern, the Pearson correlation coefficient describes it numerically. We can analyze the data further by finding an equation for the straight line that best describes the pattern. This equation predicts the value of the response( $Y$ ) variable from the value of the explanatory variable.

Much of mathematics is devoted to studying variables that are deterministically related. Saying that  $X$  and  $Y$  are related in this manner means that once we are told the value of  $X$ , the value of  $Y$  is completely specified. For example, suppose the cost for a small pizza at a restaurant is SR10 plus SR2 per topping. If we let  $Y =$  toppings and  $Y =$  price of pizza, then  $Y = 10 + 2X$ . If we order a 3-topping pizza, then  $Y = 10 + 2(3) = 16$  SR.

There are many variables  $X$  and  $Y$  that would appear to be related to one another, but not in a deterministic fashion. Suppose we examine the relationship between  $X$  (high school GPA) and  $Y$  (college GPA). The value of  $Y$  cannot be determined just from knowledge of  $X$ , and two



different students could have the same  $X$  value but have very different  $Y$  values. Yet there is a tendency for those students who have high (low) high school GPAs also to have high (low) college GPAs. Knowledge of a student's high school GPA should be quite helpful in enabling us to predict how that person will do in college.

Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a nondeterministic fashion, it is probabilistic fashion. Regression analysis is used to study the relationship between variables. Some variables are called independent variables and other variables are called dependent variables. In the case of simple linear regression, we study the linear relationship between a single independent variable ( $X$ ) and a single dependent variable ( $Y$ ). The independent variable  $X$  is called an explanatory (or predictor) variable, while the dependent variable  $Y$  is called response variable.

The simple linear regression line of a population describing the linear relationship between explanatory variable  $X$  and the response variable  $Y$  is given by the following relation:

$$Y = a + bX + \varepsilon$$

Where:

- $\varepsilon$  is a normal random variable with zero expectation  $E(\varepsilon) = 0$ . This term ( $\varepsilon$ ) in the form of simple regression line makes the regression analysis as a probabilistic approach?
- $a$  and  $b$  are the parameters of the simple regression line, where  $a$  is a constant term (intercept) and  $b$  is the coefficient of the variable  $X$  (slope).

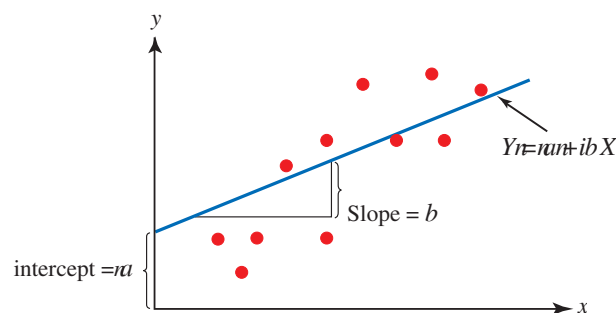


Figure 5.2.1

### THE METHOD OF LEAST SQUARES FOR ESTIMATING $a$ and $b$

Now, when we have a sample  $(x_1, y_1), (x_2, y_2), \dots$  and  $(x_n, y_n)$ , where  $x_1, x_2, \dots$  and  $x_n$  are values of  $X$  with mean  $\bar{x}$  and standard deviation  $S_X$ , and  $y_1, y_2, \dots$  and  $y_n$  are values of  $Y$  with mean  $\bar{y}$  and standard deviation  $S_Y$ . Then the least squares method is used to find the

## SECTION 5.2 SIMPLE LINEAR REGRESSION

estimation of parameters  $a$  and  $b$ . The estimated line makes the sum of the squares of the vertical distances of the data points from the line as small as possible, computationally (the sum of the error equal zero), this can be seen as the expected value of the random term  $E(\varepsilon) = 0$ . So, the estimated regression line for the given sample can be obtained (without proof) to be:

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

where the coefficients  $\hat{a}$  and  $\hat{b}$  can be estimated as:

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{Or by the relation} \quad \hat{b} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

and

$$\hat{a} = \bar{y} - \hat{b} \bar{x} \quad \text{Or by the relation} \quad \hat{a} = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

The constant  $b$  denotes the slope. The slope in the equation equals the amount that  $\hat{Y}$  changes when  $X$  increases by one unit. The constant  $a$  denotes the  $Y$ -intercept. The  $Y$ -intercept is the predicted value of  $Y$  when  $X = 0$ .

### COEFFICIENT OF DETERMINATION $r^2$

The coefficient of determination can also be obtained by squaring the Pearson correlation coefficient. This method works only for the linear regression model

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

The method does not work in general. The coefficient of determination  $\mathbf{R}^2$ , represents the proportion of the total sample variation in  $Y$  (measured by the sum of squares of deviations of the sample  $y_1, y_2, \dots$  and  $y_n$  values about their mean  $\bar{y}$ ) that is explained by (or attributed to) the linear relationship between  $X$  and  $Y$ . Some other way to calculate the coefficient of determination as:

$$r^2 = \frac{SSR}{SS_{tot}} = 1 - \frac{SSE}{SS_{tot}}$$

where

$$\text{The total sum of squared error} = SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\text{The sum of squared regression error} = SSR = \sum_{i=1}^n (\bar{y} - \hat{y}_i)^2$$

$$\text{The sum of squared error (or residuals)} = SSE = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

and

$$SS_{tot} = SSR + SSE$$

The coefficient of determination is a number between 0 and 1, inclusive. That is,  $0 \leq r^2 \leq 1$ . If  $r^2 = 0$ , the least squares regression line has no explanatory value. If  $r^2 = 1$ , the regression line explains 100% of the variation in the response variable  $Y$ .

► **EXAMPLE 5.2.1** The example data given below

$X$	$Y$
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

are plotted in Figure 5.2.2.

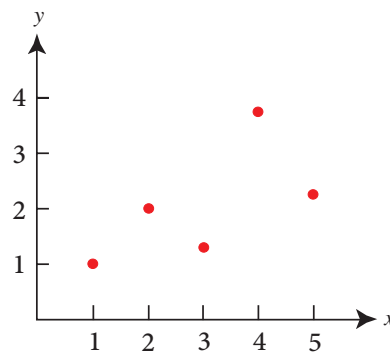


Figure 5.2.2

You can see that there is a positive relationship between  $X$  and  $Y$ . If you were going to predict  $Y$  from  $X$ , the higher the value of  $X$ , the higher your prediction of  $Y$ .

- Calculate the correlation coefficient between  $X$  and  $Y$ .
- Estimate the simple linear regression line  $\hat{Y} = \hat{a} + \hat{b}X$

SECTION 5.2 SIMPLE LINEAR REGRESSION

c. Calculate the sum of the square residuals,

**Solution**

a. From the given data, we have:

**Table 5.2.1**

$i$	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i \cdot y_i$
1	1	1	1	1	1
2	2	2	4	4	4
3	3	1.3	9	1.69	3.9
4	4	3.75	16	14.0625	15
5	5	2.25	25	5.0625	11.25
---	$\sum x_i = 15$	$\sum y_i = 10.3$	$\sum x_i^2 = 55$	$\sum y_i^2 = 28.815$	$\sum x_i \cdot y_i = 35.15$

Then the linear correlation coefficient is given by:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \cdot \left( \sum_{i=1}^n y_i \right)}{\sqrt{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2}}$$

$$= \frac{5(35.15) - (15)(10.3)}{\sqrt{[5(55) - (15)^2][5(28.815 - (10.3)^2)]}} = 0.6268327 \approx 0.63 = 63\%$$

b. Linear regression consists of finding the best-fitting straight line through the points.

The best-fitting line is the simple regression line given by:

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

The coefficients  $\hat{b}$  and  $\hat{a}$  can be estimated by using the forms:

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

And

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

In order to use these form to calculate the estimates of  $\hat{b}$  and  $\hat{a}$  we need to calculate the following table

**Table 5.2.2**

$i$	$x_i$	$y_i$	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	1	-2	-1.06	4	2.12
2	2	2	-1	-0.06	1	0.06
3	3	1.3	0	-0.76	0	0
4	4	3.75	1	1.69	1	1.69
5	5	2.25	2	0.19	4	0.38
<b>Total</b>	<b>15</b>	<b>10.3</b>	<b>0</b>	<b>0</b>	<b>10</b>	<b>4.25</b>

From the table, we have:

$$\bar{x} = \frac{15}{5} = 3 \text{ and } \bar{y} = \frac{10.3}{5} = 2.06$$

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{4.25}{10} = 0.425$$

And

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 2.06 - (0.425)(3) = 0.785$$

Hence, the estimated simple linear regression model is

$$\hat{Y} = 0.785 + 0.425X$$

- c. The sum of the square residuals can be calculated by using  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ . For this we can use the estimated equation  $\hat{Y} = 0.785 + 0.425X$ , we can calculate the residuals as:

**Table 5.2.3**

$i$	$x_i$	$y_i$	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	$\varepsilon_i^2 = (y_i - \hat{y}_i)^2$
1	1	1	1.21	-0.21	0.0441
2	2	2	1.635	0.365	0.133225
3	3	1.3	2.06	-0.76	0.5776
4	4	3.75	2.485	1.265	1.600225
5	5	2.25	2.91	-0.66	0.4356
<b>Total</b>	<b>15</b>	<b>10.3</b>	<b>10.3</b>	<b>0</b>	<b>2.79075</b>

The blue diagonal line in Figure 5.6 is the regression line and consists of the predicted score on  $Y$  for each possible value of  $X$ . The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the 3.75 point is much higher than the regression line and therefore its error of prediction is large.

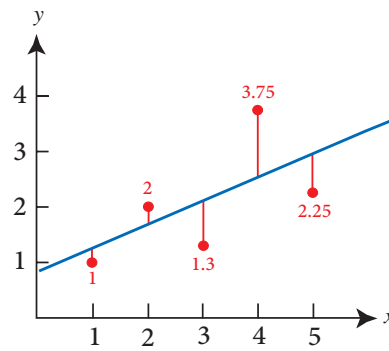


Figure 5.2.3

The blue line consists of the predictions, the points are the actual data, and the vertical lines between the points and the blue line represent errors of prediction. The error of prediction for a point is the value of the point minus the predicted value (the value on the line). The above table shows the predicted values  $\hat{y}_i$  and the errors of prediction  $\varepsilon_i = y_i - \hat{y}_i$ . For example, the first point has a  $y_1$  of 1.00 and a predicted  $y_1$  (called  $\hat{y}_1$ ) of 1.21. Therefore, its error of prediction is -0.21.

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 5.2.3. The last column in above shows the squared errors of prediction. The sum of the squared errors of prediction shown in the table is lower than it would be for any other regression line.

► **EXAMPLE 5.2.2** A certain spare part is manufactured by Westwood Company once a month in lots which vary in size as demand fluctuates. Let  $X$  represent the lot size and  $Y$  the number of Man-hours labor for 10 recent production runs. The data is given in the table below

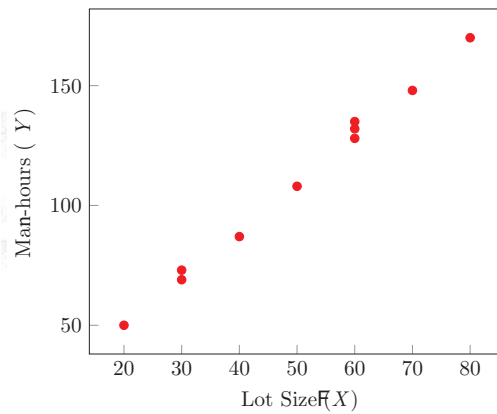
<b>The values of <math>X</math></b>	30	20	60	80	40	50	60	30	70	60
<b>The values of <math>Y</math></b>	73	50	128	170	87	108	135	69	148	132

- Construct the scatter diagram.
- Is the linear relationship appropriate to describe the relationship between  $X$  and  $Y$ ?  
In other words, do you think that the simple linear regression model
 
$$Y = a + bX + \varepsilon$$
 is appropriate to describe the relationship between  $X$  and  $Y$ ?
- Estimate the parameters of the linear regression line  $Y = a + bX$  and write down the estimated regression line.

- d. Plot the estimated regression line on the scatter diagram.
- e. Estimate (or predict) the man-hours for a lot of size 65 ( $X = 65$ ).
- f. Calculate the coefficient of determination ( $r^2$ ) and hence deduce the simple linear correlation coefficient ( $r$ ) and interpret the results.

**Solution** In this example, we have  $X$  (lot size) is independent variable (regressor/predictor variable) and  $Y$  (Man-hours) is the dependent variable (response variable).

- a. The scatter diagram is



**Figure 5.2.4**

The scatter diagram suggests that there is a strong positive linear association between  $X$  and  $Y$  since there is a linear trend for which the value of  $Y$  linearly increases when the value of  $X$  increases.

- b. The scatter diagram if Figure 5.2.4 shows that there is a linear trend since the value of  $Y$  linearly increases when the value of  $X$  increases. Hence, the linear relationship is appropriate for describing the relationship between  $X$  and  $Y$ , i.e., the regression model  $Y = a + bX + \varepsilon$  is appropriate to describe the relationship between  $X$  and  $Y$ .
- c. Estimating the parameters of the regression  $a$  and  $b$ .

**Table 5.2.4-a**

$i$	$x_i$	$x_i^2$	$y_i$	$y_i^2$	$x_i \cdot y_i$
1	30	900	73	5329	2190
2	20	400	50	2500	1000
3	60	3600	128	16384	7680
4	80	6400	170	28900	13600
5	40	1600	87	7569	3480
6	50	2500	108	11664	5400

## SECTION 5.2 SIMPLE LINEAR REGRESSION

$i$	$x_i$	$x_i^2$	$y_i$	$y_i^2$	$x_i \cdot y_i$
7	60	3600	135	18225	8100
8	30	900	69	4761	2070
9	70	4900	148	21904	10360
10	60	3600	132	17424	7920
<b>Total</b>	<b>500</b>	<b>28400</b>	<b>1100</b>	<b>134660</b>	<b>61800</b>

From the table, we have:

$$\sum_{i=1}^{10} x_i = 500, \sum_{i=1}^{10} y_i = 1100, \sum_{i=1}^{10} x_i^2 = 28400, \sum_{i=1}^{10} y_i^2 = 134660 \text{ and } \sum_{i=1}^{10} x_i \cdot y_i = 61800$$

Then the estimation of the parameters are:

$$\hat{b} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{10(61800) - (500 \times 1100)}{10(28400) - (500)^2} = \frac{68000}{34000} = 2$$

$$\hat{a} = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{(1100 \times 28400) - (500 \times 61800)}{10(28400) - (500)^2} = 10$$

The estimated simple linear regression equation is  $\hat{Y} = 10 + 2X$ . From this equation, we see that when the lot size increases by one unit, the Man-hours increases by 2 hours, while there are 10 hours do not depend on the lot size.

- d. The estimated regression line on the scatter diagram as shown below

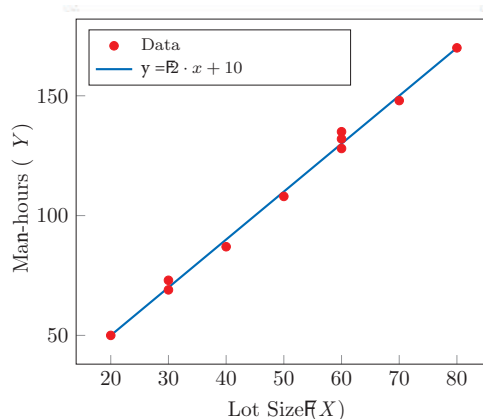


Figure 5.2.5

- e. Estimate (or predict) the man-hours for a lot of size 65 ( $X = 65$ ) is

$$\hat{Y} = 10 + 2X = 10 + 2(65) = 140 \text{ hours.}$$



- f. To calculate the coefficient of determination ( $R^2$ ) we need to get the sum of squared errors as:

Table 5.2.4-b

$i$	$y_i$	$\hat{y}_i$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	73	70	-37	1369	3	9	-40	1600
2	50	50	-60	3600	0	0	-60	3600
3	128	130	18	324	-2	4	20	400
4	170	170	60	3600	0	0	60	3600
5	87	90	-23	529	-3	9	-20	400
6	108	110	-2	4	-2	4	0	0
7	135	130	25	625	5	25	20	400
8	69	70	-41	1681	-1	1	-40	1600
9	148	150	38	1444	-2	4	40	1600
10	132	130	22	484	2	4	20	400
<b>Total</b>	<b>1100</b>	<b>1100</b>	<b>0</b>	<b>13660</b>	<b>0</b>	<b>60</b>	<b>0</b>	<b>13600</b>

From the table 5.2.5-a and table 5.2.5-b, we have:

- The total sum of squared error =  $SS_{tot} = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 13660$
- The sum of squared regression error =  $SSR = \sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2 = 13600$
- The sum of squared error =  $SSE = \sum_{i=1}^{10} (y_i - \hat{y}_i)^2 = 60$

It is clear that  $SS_{tot} = SSR + SSE$ .

The coefficient of determination is:

$$r^2 = \frac{SSR}{SS_{tot}} = \frac{13600}{13660} = 0.9956076 \approx 99.6\%$$

Or

$$r^2 = 1 - \frac{SSE}{SS_{tot}} = 1 - \frac{60}{13660} = 1 - 0.00439 = 0.99561 \approx 99.6\%$$

This shows that 99.6% of the total variation of the Man-hours is explained by the lot size and hence we can conclude that the lot size is the most important variable to predict the Man-hours.

**REMARK 5.2.1**

The simple linear correlation coefficient can be calculate from the coefficient of determination as  $r = \pm\sqrt{r^2}$  where the sign is chosen as the sign of the slope estimation  $\hat{b}$ .

In our example, the linear correlation coefficient is

$$r = +\sqrt{r^2} = \sqrt{0.99561} = 9978 \approx 99.8\% ,$$

this value of the linear correlation coefficient can be also obtained by using Pearson correlation coefficient to get the same result.



## EXERCISES

1. What is the chart called when the paired data (the dependent and independent variables) are plotted?
  - a. Scatter diagram
  - b. Bar chart
  - c. Pie chart
  - d. Histogram
2. What is the variable used to predict the value of another called?
  - a. Independent
  - b. Correlation
  - c. Dependent
  - d. Determination
3. Which of the following statements regarding the coefficient of correlation is true?
  - a. It ranges from  $-1.0$  to  $+1.0$  inclusive
  - b. It measures the strength of the relationship between two variables
  - c. A value of  $0.00$  indicates two variables are not related
  - d. All of the above
4. What does a coefficient of correlation of  $0.70$  infer?
  - a. Almost no correlation because  $0.70$  is close to  $1.0$
  - b.  $70\%$  of the variation in one variable is explained by the other
  - c. Coefficient of determination is  $0.49$
  - d. Coefficient of non-determination is  $0.30$
5. What is the range of values for a coefficient of correlation?
  - a.  $0$  to  $+1.0$
  - b.  $-1.0$  to  $+1.0$  inclusive
  - c.  $-3$  to  $+3$  inclusive
  - d. Unlimited range
6. If the correlation between two variables is close to one, the association is
  - a. strong
  - b. weak
  - c. moderate
  - d. none
7. If the correlation coefficient between two variables equals zero, what can be said of the variables  $X$  and  $Y$ ?
  - a. Not related
  - b. Dependent on each other
  - c. Highly related
  - d. All of the above are correct

8. What can we conclude if the coefficient of determination is 0.94?
- a. Strength of relationship is 0.94
  - b. Direction of relationship is positive
  - c. 94% of total variation of one variable is explained by variation in the other variable.
  - d. All of the above are correct
9. If  $r = -1.00$ , what inferences can be made?
- a. The dependent variable can be perfectly predicted by the independent variable
  - b. All of the variation in the dependent variable can be accounted for by the independent variable.
  - c. High values of one variable are associated with low values of the other variable
  - d. Coefficient of determination is 100%.
  - e. All of the above are correct.
10. If  $r = 0.65$ , what does the coefficient of determination equal?
- a. 0.194
  - b. 0.577
  - c. 0.423
  - d. 0.806
11. What does the coefficient of determination equal if  $r = 0.89$ ?
- a. 0.94
  - b. 0.79
  - c. 0.89
  - d. 0.06
12. Which value of  $r$  indicates a stronger correlation than 0.40?
- a.  $-0.30$
  - b.  $+0.38$
  - c.  $-0.50$
  - d. 0
13. What is the range of values for the coefficient of determination?
- a.  $-1$  to  $+1$  inclusive
  - b.  $-100\%$  to  $+100\%$  inclusive
  - c.  $-100\%$  to  $0\%$  inclusive
  - d.  $0\%$  to  $100\%$  inclusive
14. Suppose the least squares regression equation is  $\hat{Y} = 1202 + 1,133X$ . When  $X = 3$ , what does  $\hat{Y}$  equal?
- a. 5,734
  - b. 4,601
  - c. 8,000
  - d. 4,050
15. What is the general form of the regression equation?
- a.  $\hat{Y} = ab$
  - b.  $\hat{Y} = a + bX$
  - c.  $\hat{Y} = a - bX$
  - d.  $\hat{Y} = abX$

16. Based on the regression equation, we can
- Predict the value of the dependent variable given a value of the independent variable.
  - Predict the value of the independent variable given a value of the dependent variable.
  - Measure the association between two variables.
  - all of the above.
17. In the regression line equation,  $\hat{Y} = 10 + 20X$  the value of 20 indicates
- The  $Y$  intercept.
  - For each unit increase in  $X$ ,  $Y$  increases by 20.
  - For each unit increase in  $Y$ ,  $X$  increases by 20.
  - None of the above.
18. In the equation  $\hat{Y} = a + bX$ , what is  $\hat{Y}$ ?
- Slope of the line
  - $Y$  intercept
  - Predicted value of  $Y$ , given a specific  $X$  value
  - Value of  $Y$  when  $X = 0$ .
19. Assume the least squares equation is  $\hat{Y} = 10 + 20X$ . What does the value of 10 in the equation indicate?
- $Y$  intercept
  - For each unit increased in  $Y$ ,  $X$  increases by 10
  - For each unit increased in  $X$ ,  $Y$  increases by 10
  - None of the above
20. Given the following five points:  $(-2,0)$ ,  $(-1,0)$ ,  $(0,1)$ ,  $(1,1)$ , and  $(2,3)$ .
- What is the slope of the line?
  - What is the  $Y$  intercept?
21. A company wants to study the relationship between an employee's length of employment and their number of workdays absent. The company collected the following information on a random sample of seven employees.

Number of workdays absent	2	3	3	5	7	7	8
Length of employment (in yrs)	5	6	9	4	2	2	0

- What is the independent variable ( $X$ )?
- What is the dependent variable ( $Y$ )?

- c. What is the slope of the linear equation?
- d. What is the  $y$  intercept of the linear equation?
- e. What is the regression line equation for the data?
- f. What is the meaning of a negative slope?

**22.** The relationship between interest rates as a percent ( $X$ ) and housing starts ( $Y$ ) is given by the linear equation  $\hat{Y} = 4094 - 269X$ .

- a. What will be the number of housing starts if the interest rate is 8.25%?
- b. What will be the number of housing starts if the interest rate rose to 16%?
- c. For what interest rate will the maximum number of housing starts be achieved?

**23.** A sales manager for an advertising agency believes there is a relationship between the number of contacts and the amount of the sales. To verify this belief, the following data was collected:

Salesperson	Number of Contacts	Sales (in thousands)
1	14	24
2	12	14
3	20	28
4	16	30
5	46	80
6	23	30
7	48	90
8	50	85
9	55	120
10	50	110

- a. What is the dependent variable?
- b. What is the independent variable?
- c. What is the  $Y$ -intercept of the linear equation?
- d. What is the slope of the linear equation?
- e. What is the value of the coefficient of correlation?
- f. What is the value of the coefficient of determination?

**24.** Let  $X$  be the body weight of a child (in kilograms), and let  $Y$  be the metabolic rate of the child (in 100 kcal/24h).

$X$	3	5	9	11	15	17	19	21
$Y$	1.4	2.7	5.0	6.0	7.1	7.8	8.3	8.8

- a. Estimate the regression line and use it to the metabolic rate for a child of body weight 13 kilograms.
- b. Calculate the coefficients of determination and correlation.
- c. Interpret the results in (a) and (b).
- 25.** Suppose that we are interested in estimating the blood glucose levels (mg/100ml) of adult women in a certain population using his weights (in kg). A study was made for this purpose and gave the following results:

$X$ (weights)	63	65	72	80	90
$Y$ (blood glucose levels)	107	109	106	101	100

- 26.** In the following problems, suppose that the simple linear regression model  $Y = a + bX$  is appropriate to describe the relationship between  $X$  and  $Y$ . We have the following information:

$$\sum (x_i - \bar{x})^2 = 498, \quad \sum (y_i - \bar{y})^2 = 61.2, \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = -164$$

- A.** The coefficient of correlation ( $r$ ) equals to:
- a.  $-0.939$                       b.  $0.25$                       c.  $0.68$                       d.  $-0.81$
- B.** Based on the value of the coefficient of correlation ( $r$ ), the linear relationship between  $X$  and  $Y$  is:
- a. Strong                      b. Complete                      c. Weak                      d. Moderate

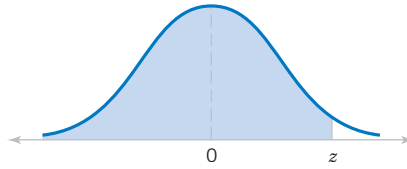




## REFERENCES

- ① Bernstein, Stephen, and Ruth Bernstein. *Schaums Outline of Elements of Statistics I: Descriptive Statistics and Probability*. McGraw Hill Professional, 1999.
- ② Brase C.H. and Brase, C.P. *Understanding Statistics*, Fifth edition, Brooks/Cole, Canada, 2010.
- ③ Chase W. and Bown, F. *General Statistics*, Second edition, John Wiley and Sons, New York, 1992.
- ④ Feller, W. *An introduction to Probability Theory and its Applications*, Vol. II. John Wiley & Sons Inc., New York, 1971.
- ⑤ Gharamani, S. *Fundamentals of Probability and Stochastic Processes*. 2nd edition. Pearson and Prentice Hall, New Jersey, 2005.
- ⑥ Goos, Peter, and David Meintrup. *Statistics with JMP: Graphs, Descriptive Statistics and Probability*. John Wiley & Sons, 2015.
- ⑦ Isotalo, Jarkko. *Basics of Statistics*. Finland: University of Tampere, 2001.
- ⑧ Jaggi, Seema, *Descriptive Statistics and Exploratory Data Analysis*. Indian Agricultural Statistics Research Institute, 2003.
- ⑨ Mann, Prem S. *Introductory Statistics*. Seventh Edition International Student Version. N.p.: John Wiley & Sons, 2010.
- ⑩ Moore, David S. *The Basic Practice of Statistics*. Vol. 2. New York: WH Freeman, 2007.
- ⑪ Nicholas, Jackie. *Introduction to Descriptive Statistics*. Mathematics Learning Centre, University of Sydney, 1990.
- ⑫ Ross, S. M., *A first Course in Probability*. Prentice-Hall, Inc., New Jersey, sixth edition, 1984.
- ⑬ Ross, S. M., *Introduction to Probability and Statistics for Engineers and Scientists*, Fourth edition, 2009.
- ⑭ Ross, S. M., *Introductory Statistics*. Academic Press, 2005.
- ⑮ Samules, M.L., Witmer, J.A and Schaffner, A., *Statistics for the Life Sciences*. Fourth edition, Pearson, New York, 2012.
- ⑯ Soong, T. T. *Fundamental of Probability and Statistics for Engineers*. John Wiley and Sons, 2004.
- ⑰ Ware, William B., John M. Ferron, and Barbara Manning Miller. *Introductory Statistics: A Conceptual Approach Using R*. Routledge, 2013.
- ⑱ Walpole, R.E., Myers, R.H. and Myers, S.L. and Ye, K., *Probability and Statistics for Engineers and Scientists*, Ninth Edition, Prentice, New York, 2012.

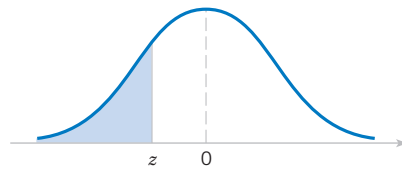
## STANDARD NORMAL DISTRIBUTION TABLE



For positive values of  $z$

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.50 and up	.9999									

# STANDARD NORMAL DISTRIBUTION TABLE



For negative values of  $z$

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.50 and lower	.0001									
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

## SUBJECTS INDEX

- A**
  - Acceptance region
  - Alternative hypotheses
  - Axioms
  
- B**
  - Bar chart
  - Bayes' theorem
  - Binomial distribution
  - Box plot
  
- C**
  - Central limit theorem
  - Chebyshev's theorem
  - Class interval
  - Class lower limit
  - Class midpoint
  - Class upper limit
  - Classical approach
  - Coefficient of determination
  - Coefficient of variation
  - Combination
  - Complement
  - Concepts of probability
  - Conditional probability
  - Conditional probability.
  - Confidence interval
  - Confidence level
  - Continuous
  - Continuous
  - Continuous data
  - Contrast independent
  - Correlation
  - Correlation coefficient
  - Counting techniques
  - Counting techniques
  - Critical region

- Cumulative distribution
- Cumulative frequency
- Cumulative relative frequency
  
- D** Data
- Deciles
- Dependent events
- Dependent variable
- Descriptive statistics
- Discrete
- Discrete data
  
- E** Empirical rule
- Event
- Expectation
- Exponential distribution
  
- F** Five number summary
- Frequency
- Frequency table
  
- G** Geometric distribution
- Graphical representation
- Grouped data
  
- H** Histogram
- Hypotheses testing
  
- I** Independent variable
- Inference
- Inferential statistics
- Intercept
- Interquartile range
- Intersection
- Interval estimation



## SUBJECTS INDEX

- J**    Joint probability
  
- L**    Length  
      Level of significance  
      Linear
  
- M**    Margin error  
      Marginal  
      Marginal probability  
      Mean  
      Measures of central tendency  
      Measures of dispersion  
      Median  
      Mode  
      Multiple bar chart  
      Multiplicative rule  
      Mutually exclusive
  
- N**    Negative correlation  
      Normal distribution  
      Null hypotheses
  
- O**    Observation  
      Ogive
  
- P**    Parameter  
      Pearson correlation coefficient  
      Percent frequency  
      Percentiles  
      Permutation  
      Pie chart  
      Point estimation  
      Poisson distribution

Polygon  
Population  
Positive correlation  
Predictor variable  
Probability  
Probability density function  
Probability distribution  
Probability mass function  
 $p$  -value

**Q** Qualitative data  
Qualitative variable  
Quantitative variable  
Quartiles

**R** Random experiment  
Random variable  
Range  
Raw data  
Regression error  
Rejection region  
Relative frequency  
Relative frequency approach  
Residuals  
Response variable

**S** Sample  
Sample mean  
Sample proportion  
Sample space  
Sampling distribution  
Simple event  
Simple regression line  
Skewed histogram  
Slope



## SUBJECTS INDEX

Stacked bar chart  
Standard deviation  
Standard error  
Standard scores  
Statistic  
Statistics  
Step function  
Strong correlation  
Subjective approach  
Sure event  
Symmetric histogram

**T** Test statistic  
Total error  
Total probability  
Two directional bar chart

**U** Uniform distribution  
Uniform histogram  
Union

**V** Variable  
Variance  
Venn diagrams

**W** Weak correlation  
Weighted mean

**Z** Z-table