Introduction to

PROBABILITY AND STATISTICS

THIRD EDITION

ABDULRAHMAN ABOUAMMOH

KHALAF SULTAN

MOHAMED KAYID

MANSOUR SHRAHILI

KING SAUD UNIVERSITY

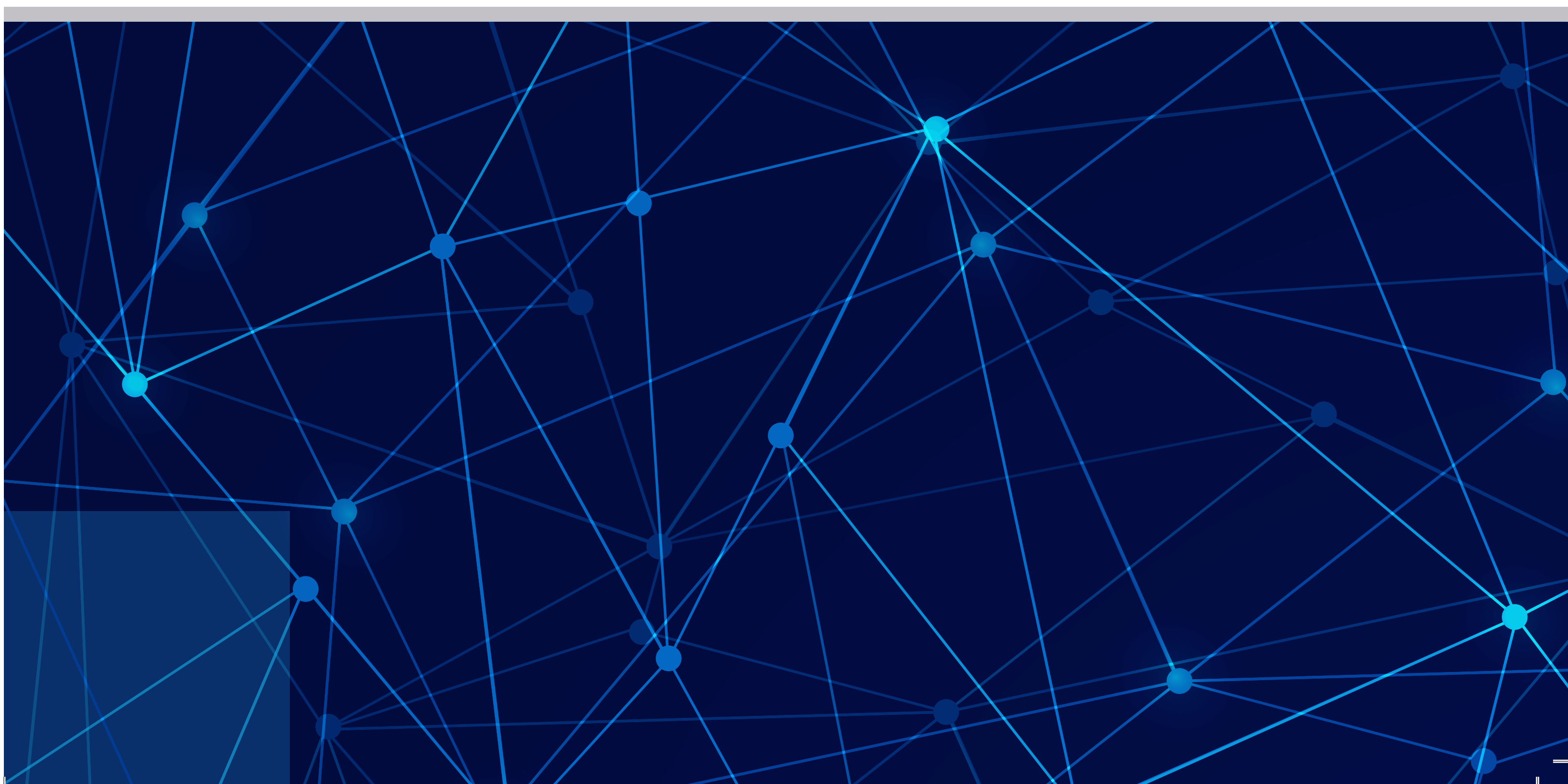


TABLE OF CONTENTS

Table of Contents iii

Introduction vi

Acknowledgements ix

CHAPTER 1 • • Descriptive Statistics 1

1.0 Introduction 2

1.1 Basic Concepts and Definitions 4

1.2 Organization of Data 11

1.3 Graphical Representations 19

1.4 Measures of Central Tendency 27

1.5 Measures of Variability 47

EXERCISES 56

CHAPTER 2 • • Probability 65

2.1 Mathematical Concepts 66

2.2 Definitions And Concepts In Probabilty Calculus 70

2.3 Concept Of Probability Function 81

2.4 Conditional Probabilty And Independence Of Event 89

EXERCISES 102

CHAPTER 3 • • Random Variables and Probability Distributions 107

3.1 Concept Of Random Variables And Their Distributions 108

3.2 Discrete Random Variables and their Distributions 112

3.3 Continuous Random Variables and their Distributions 133

EXERCISES 148

CHAPTER 4 • • Introduction to Statistical Inference 157

4.1 Definitions and Concepts 158

4.2 Estimation of the Population Mean 165

4.3 Estimation of The Population Proportion 171

4.4 Introduction to Hypotheses Testing 179

4.5 Hypothesis Testing for the Population Mean 183

4.6 Hypothesis Testing for the Population Proportion 191

EXERCISES 196

TABLE OF CONTENTS

CHAPTER 5 • • Correlation and Regression 201

5.0 Introduction 202

5.1 Simple Linear Correlation 203

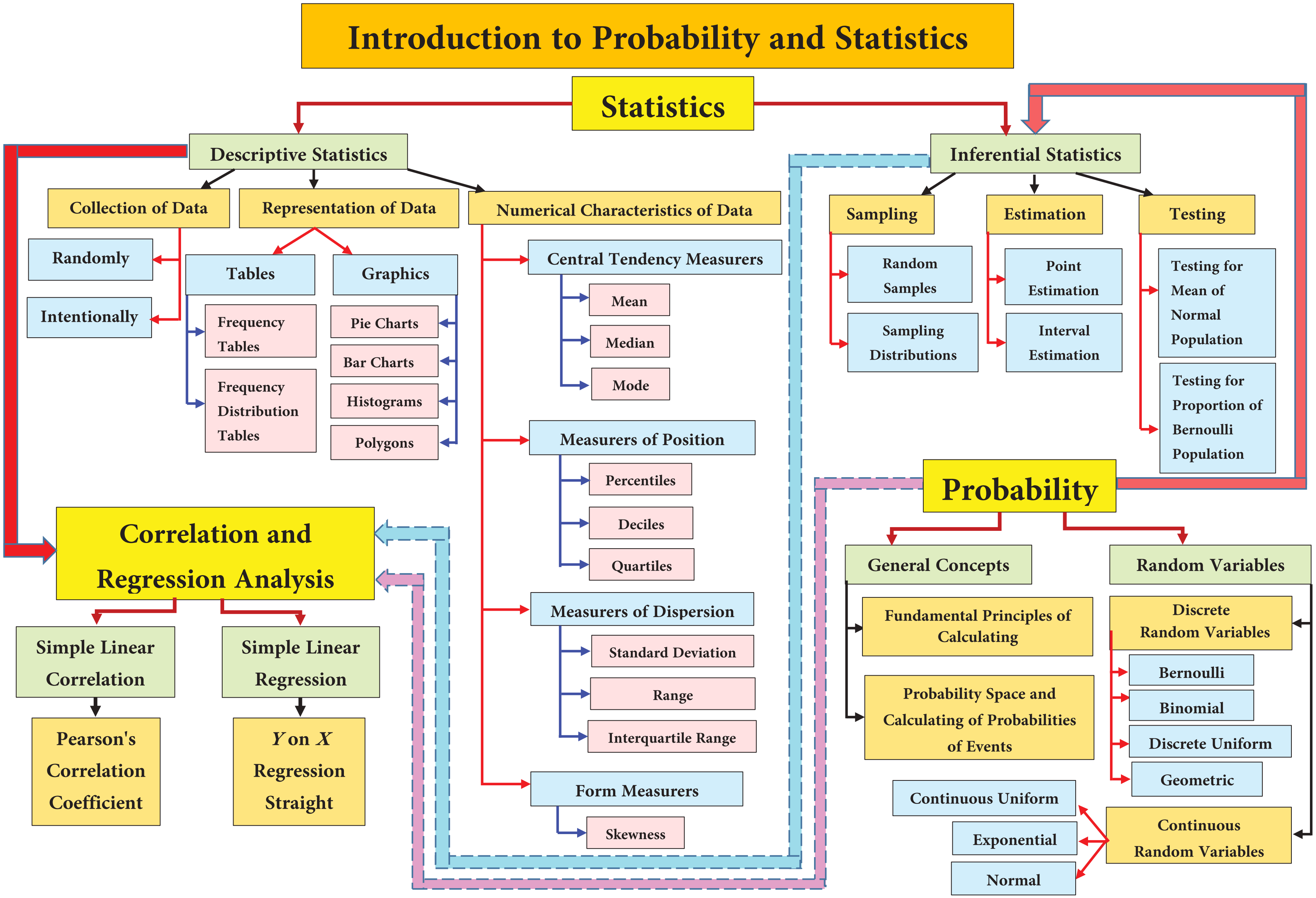
5.2 Simple Linear Regression 209

EXERCISES 221

References *227*

Normal Distribution Table *228*

Subjects Index *230*



Introduction to Probability and Statistics
Statistics
Descriptive Statistics
Collection of Data
Randomly
Intentionally
Representation of Data
Tables
Frequency Tables
Frequency Distribution Tables
Graphics
Pie Charts
Bar Charts
Histograms
Polygons
Numerical Characteristics of Data
Central Tendency Measurers
Mean
Median
Mode
Measurers of Position
Percentiles
Deciles
Quartiles
Measurers of Dispersion
Standard Deviation
Range
Interquartile Range
Form Measurers
Skewness
Inferential Statistics
Sampling
Random Samples
Sampling Distributions
Estimation
Point Estimation
Interval Estimation
Testing
Testing for Mean of Normal Population
Testing for Proportion of Bernoulli Population
Correlation and Regression Analysis
Simple Linear Correlation
Pearson's Correlation Coefficient
Simple Linear Regression
Y on X Regression Straight
Probability
General Concepts
Fundamental Principles of Calculating
Probability Space and Calculating of Probabilities of Events
Random Variables
Discrete Random Variables
Bernoulli
Binomial
Discrete Uniform
Geometric
Continuous Random Variables
Continuous Uniform
Exponential
Normal

INTRODUCTION

Introduction to the First Edition

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

Introduction to the Second Edition

The authors have stated in the first edition of this book that they would welcome any corrections, changes or suggestions by colleagues in the Department of Basic Sciences at the Deanship of the Joint First Year Program - at King Saud University and students.

The authors have felt the need to include few changes for the book as soon as they have had a critical review of the printed version of the book.

The book has been used, as text for a course in statistics, for the whole year 2017 - 2018 and staffs of the Joint First Year Program and other colleagues have proposed interesting ideas for improving the presentation and readability of the book.

The authors are grateful, in particular, to **Prof. H. Al-Oklah**, for his constructive remarks and the valuable additions, Dean of the Common First Year and Charmin of the Department of Basic Sciences, at King Saud University for their support and inspiration.

Due to theirs and other expedient suggestions, the authors are able to make this edition is almost free of typos. Further standardization of notations, additional examples and more relevant problems have been included in this version. Lists of excursuses have had additions and omissions to make them more suitable and become more gradual in addressing various concepts of the theory and examples for each chapter.

The authors hope, once again, to receive from our teaching colleagues, other readers of the book and students to provide us with any further, remarks and suggestion that might be of help to improve future edition of the book.

Finally, **Mr. Fadi Hasan** deserves our thanks for his skillful efforts in typesetting and production of the printed version of the book, and also the presidency of the Department of Basic Sciences for their trust. As Mr. Fadi Hasan returns thanks to typesetting the book and take it out to print properly.

The Authors
April 2018

Introduction to the Third Edition

The third edition of this book provides an extensive update on the second edition. The book has been used, as text for a course in statistics, for students and staffs of the Joint First Year Program and other colleagues have proposed interesting ideas for improving the presentation and readability of the book.

The authors would like to thank the colleagues in the Department of Basic Sciences at the Deanship of the Joint First Year Program - at King Saud University and students for their thoughtful review of the second edition. They raise important issues and their inputs are very helpful for improving this edition. Due to theirs and other expedient suggestions, the authors are able to make this edition is almost free of typos. Further standardization of notations, additional examples and more relevant problems have been included in this version. Lists of excursuses have had additions and omissions to make them more suitable and become more gradual in addressing various concepts of the theory and examples for each chapter.

However, the authors hope, once again, to receive from our teaching colleagues, other readers of the book and students to provide us with any further, remarks and suggestions that might be of help to improve future edition of the book.

The authors are grateful, in particular, to **Prof. H. Al-Oklah**, for his constructive remarks and the valuable additions, Dean of the Common First Year and Charmin of the Department of Basic Sciences, at King Saud University for their support and inspiration.

Finally, **Mr. Fadi Hasan** deserves our thanks for his skillful efforts in typesetting and production of the printed version of the book, and also the presidency of the Department of Basic Sciences for their trust. As Mr. Fadi Hasan returns thanks of typesetting the book and take it out to print properly.

The Authors
March 2019

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. 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 OF DATA
- SECTION 1.3 GRAPHICAL REPRESENTATIONS
- SECTION 1.4 MEASURES OF CENTRAL TENDENCY
- SECTION 1.5 MEASURES OF DISPERSION

EXERCISES

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 and 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.

DEFINITION 1.0.1 (Data)

Data is a collection of information collected by means of experiments, observations or real life events and stored in a proper format (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 example:

a. Suppose, in a study on flower colors of different plants found in a garden, we obtained the following information:

White, Red, Yellow, Yellow, White, Purple, Blossom, Purple, Red, Red and Red.

These traits, which represent information about the flower colors of those plants studied, are data.

b. Suppose that when measuring the weight of a group of sheep in a fattening cattle farm we obtained the following values:

45, 42, 50, 44, 42, 52, 52, 43, 42, 45, 44, 48, 48, 50

These values, which represent information about the weight of those sheep that have been weighed, are data.

c. Students' performance (includes number of students, marks obtained, age), are data.

d. Country's growth (include literacy rate, crime rates, health conditions), are data.

e. Company's performance (include sales, revenue generated, profit, share price), are data.

Thus, our contemporary life is full of information. That is why 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 with highly 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 Tadawul 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. For example, 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.

DEFINITION 1.0.2 (Statistics)

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

Some other important concepts in descriptive statistics will be present in the following section.

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. One of them is provided by the following definition:

DEFINITION 1.1.1 (Descriptive Statistics)

Descriptive statistics consist of methods and techniques which are used for presenting and summarizing data in tables or graph forms and provide some numerical measures for it.

In effect, the descriptive statistics includes summarizing the data into tables and construction of charts (Pies chart (a circular disk), Bar chart), Histograms, Polygons and other representation graphs, as well calculation of averages, percentiles, dispersions and other descriptive measures.

▶ EXAMPLE 1.1.1:

a. The study of some human characteristics includes average height, average weight and …, based on data taken from a group of people is a descriptive study. It is therefore under the concept of descriptive statistics.

b. The study of some physical characteristics for a kind of petroleum includes boiling point, viscosity and ..., based on data taken from a quantity of oil from a particular well is a descriptive study. It is therefore under the concept of descriptive statistics.

◀

So we note that the descriptive statistics do not allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made. It is a simply way to organize and describe our data.

Now, before presenting the definition of the second sector of statistics, we will provide the following concepts:

DEFINITION 1.1.2 (Population)
Population is a set of all things (which have at least one common characteristic (or feature)) that will be subjected to a study to obtain inferences for a specific problem. The elements of population are called individuals.

EXAMPLES 1.1.2: The following sets represent population:

- **a.** The set of all students in a country.
- **b.** The set of all palm trees in KSA.
- **c.** The set of income of all citizens of a country.
- **d.** The set of all cars in a city.

REMARKS 1.1.1

- **a.** The number of elements of a population is called the size of the population, and if this number is finite then it symbolizes by N or M or ….
- **b.** Population as a whole is often doesn't be used to collect information and make inferences due to its size. Usually a subset of the population which reflects all the major characteristics of the population is used for collecting the information.

DEFINITION 1.1.3 (Sample)
A sample is a subset of population, which is used to collect information and to make inferences about the entire population.

REMARK 1.1.2
The number of elements of a sample is called the size of the sample, and denoted by n or m or ….

EXAMPLES 1.1.3: The following subsets represent samples:

- **a.** The set, which we get by selecting Qassim's palm trees from all palm trees in KSA.
- **b.** The set, which we get by selecting (or withdrawing) citizens from a country in order to identify the income level of the country's citizens.
- **c.** The set, which we get by selecting (or withdrawing) cars in the centrum of Riyadh in order to identify the technical conditions of cars in Riyadh city.

DEFINITION 1.1.4 (Inferential Statistics)
Inferential statistics is some methods and techniques that can be used for drawing conclusions about the entire population using the observations from the samples taken from that population.

In other words, inferential statistics is the branch of statistics that involves drawing conclusions about a population based on information contained in a random sample taken from that population. It includes point estimation, interval estimation, hypothesis testing, statistical modeling, clustering and many more methods based on probability and distribution theories.

▶ **EXAMPLE 1.1.4:** If we want to measure the income of all the citizens of 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 of the KSA population 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.

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.

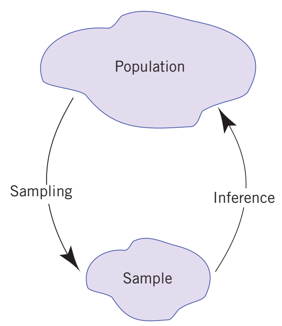


Figure 1.1.1 (Relationship between population and sample)

DEFINITION 1.1.5 (Parameter)
Parameter is a certain quantity or quality for describing a characteristic or phenomenon in a given population that summarizes the data for the entire population.

In a statistical study the measure of a parameter is typically unknown.

DEFINITION 1.1.6 (Statistic)
Statistic is a certain quantity or quality for describing a characteristic or phenomenon of a sample that summarizes the data for the entire sample.

Since parameters are usually unknown, sample statistics are used to estimate them. In other words, one uses a statistic(s) to make inference about a parameter(s).

▶ EXAMPLES 1.1.5:

a. 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. Let's withdraw a sample of size n from the population. Then we calculate the ratio $\hat{p}$ to those people who have aged 20-30 years old having height less than 5 feet.

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.

b. There is millions of passenger personal cars in Riyadh. One can ask, what is the average price of these cars?

It is obviously impractical to attempt to solve this problem directly by assessing the value of every single car in the city, adding up all those numbers, and then dividing by how many they are. Instead, the best we can do would be to estimate the average. One natural way to do so would be to randomly select some of the personal cars, say 50 of them, ascertain the price of each of those cars, and find the average of those 50 numbers.

The set of all those millions of vehicles is the population of interest, and the number attached to each one, its value, is a measurement. The average value is a parameter (a number (monetary worth) that describes a characteristic of the population).

The set of 50 personal cars selected from the population is a sample, and the 50 numbers that represent the monetary values of the personal cars we selected, are the sample data. The average of this data is a statistic (a number calculated from the sample data).

Now, if the average value of the personal cars in our sample was 40000 S.R., then it seems reasonable to conclude that the average value of all cars is about 40000 S.R. In reasoning this way, we have drawn an inference about the population based on information obtained from the sample.

In general, statistics is a study of data (describing properties of the data, which is called descriptive statistics), and drawing conclusions about a population of interest from information extracted from a sample, which is called inferential statistics.

Computing the single number 40000 S.R to summarize the data was an operation of descriptive statistics; using it to make a statement about the population was an operation of inferential statistics.

DEFINITION 1.1.7 (Variables)

A variable is a map (or a function) X defined on the population (or sample) and takes values in an arbitrary set M. That means:

$$X : Population\ (or\ Sample) \longrightarrow \mathrm{M}$$

This variable measures: A characteristic, feature or factor (that varies from one individual to another) in the population.

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

EXAMPLE 1.1.6: For countries, the function that measures: gross domestic product, sex ratio, birth and death rates, number of area, literacy rates are variables. Different countries will have different values for each of these variables. Some variables for humans are represented by measuring the 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

One can classify variables in two directions:

a. According to the type of values that the variable takes,

b. According to the number of values that the variable takes.

DEFINITION 1.1.8 (Qualitative or Categorical Variable)
A qualitative variable is a variable that takes non-numeric values or numeric values which indicate an attribute or property.

EXAMPLES 1.1.7:

a. Variables that determine: marital status, eye-color, gender, hair-color, country, city of a human, are qualitative (or categorical) variables. The elements which take such variables (numeric or non-numeric) no mathematical operations (addition, subtraction, multiplication or division) can be performed on them.

b. The variable that determines the identification or identity (ID) of students in King Saud University is a qualitative variable.

DEFINITION 1.1.9 (Quantitative Variables)
These are the variables which take numerical values, and these numerical values can be undergoing mathematical operations (or calculation operations).

We note that the values of such variables can be ordered in increasing or decreasing.

EXAMPLES 1.1.8: Variables that determine height, weight, temperature, revenue for people are quantitative variables.

DEFINITION 1.1.10 (Discrete Variables)
The variables, which take finite or infinite countable number of values, are called discrete Variables.

DEFINITION 1.1.11 (Continuous Variables)
The variables, which take uncountable number of values, are called continuous variables.

EXAMPLES 1.1.9:

a. Variables, which record 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 and so on, are discrete variables.

b. Variables, which record the blood groups of people, the type of cars, the fruit varieties and so on, are discrete variables.

c. Variables, which measure weight of person, distance between the cities, temperature, and income and so on, are continuous variables.

d. The variable, which records all spectrum colors resulting from white light analysis is a continuous variable.

REMARK 1.1.3

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

The following Diagram summarizes the classification of variables

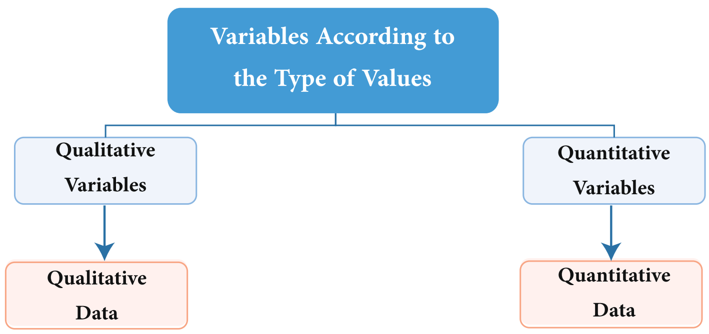


Figure 1.1.2

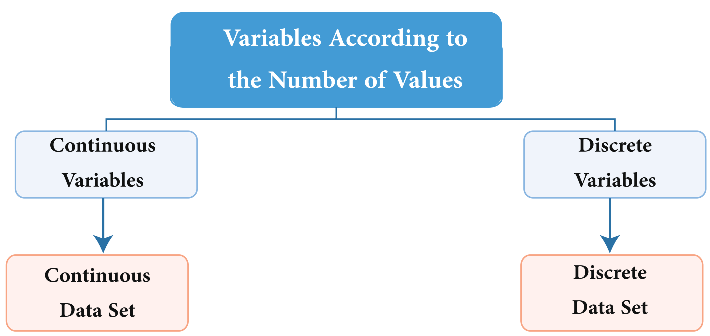


Figure 1.1.3

Section 1.2

ORGANIZATION OF DATA

In many problems, statisticians are provided with data in a format which is not well organized, and usually these data are the result of measuring variables directly and without performing any treatments on them. Such set of data is known as raw data.

▶ **EXAMPLE 1.2.1** **(Qualitative data):** Consider the month of the birth of 25 members of a community, where we find the following raw data:

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

We note that all data are qualitative data.

◀

▶ **EXAMPLE 1.2.2** **(Discrete quantitative data):** Consider the number of children in 40 families of a society, where we find the following raw data:

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

We note that all data are quantitative data.

◀

▶ **EXAMPLE 1.2.3** **(Continuous quantitative data):** Consider the height of 25 adult men (in centimeters), where we find the following raw data:

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

We note that all data are quantitative data also, but its continuous data because the generator of data is continuous variable.

◀

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 the different ways of representing data one should understand the concepts of frequency, relative frequency, cumulative frequency and percent frequency.

The frequency of a particular data value is the number of times the data value occurs. Accordingly, 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:

$$\text{The relative frequency of a class} = \frac{\text{The frequency of class}}{\text{The sum of frequencies}}$$

$$\text{The percent frequency of a class} = (\text{The relative frequency}) \times 100\%$$

FREQUENCY TABLE (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 month of 25 persons)

Month	Frequency	Relative Frequency	Percent Frequency
January	2	$2/25 = \mathbf{0.08}$	$0.08\times100 = \mathbf{8}$ %
February	3	$3/25 = \mathbf{0.12}$	$0.12\times100 = \mathbf{12}$ %
March	2	$2/25 = \mathbf{0.08}$	$0.08\times100 = \mathbf{8}$ %
April	3	$3/25 = \mathbf{0.12}$	$0.12\times100 = \mathbf{12}$ %
May	2	$2/25 = \mathbf{0.08}$	$0.08\times100 = \mathbf{8}$ %
June	4	$4/25 = \mathbf{0.16}$	$0.16\times100 = \mathbf{16}$ %
July	3	$3/25 = \mathbf{0.12}$	$0.12\times100 = \mathbf{12}$ %
August	2	$2/25 = \mathbf{0.08}$	$0.08\times100 = \mathbf{8}$ %
September	1	$1/25 = \mathbf{0.04}$	$0.04\times100 = \mathbf{4}$ %
October	0	$0/25 = \mathbf{0}$	$0\times100 = \mathbf{0}$ %
November	1	$1/25 = \mathbf{0.04}$	$0.04\times100 = \mathbf{4}$ %
December	2	$2/25 = \mathbf{0.08}$	$0.08\times100 = \mathbf{8}$ %
Total	$n = \sum_i f_i = 25$	1	100%

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

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

▶ **EXAMPLE 1.2.5:** 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 %
Total	40	1	100%

FREQUENCY TABLE (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 1.2.6:** Construct 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)

Number 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 %
Total	40	1	100%

▶ **EXAMPLE 1.2.7:** We want to see in how many subjects each student failed in the 5th standard. Therefore, we consider a sample of 40 students. So we find the following data.

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 construct 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 %
Total	40	1	100%

◀

FREQUENCY DISTRIBUTION TABLE (CONTINUOUS QUANTITATIVE DATA)

A continuous quantitative variable takes infinite (and uncountable also) number of values and hence it is impossible to create a frequency distribution table which gives frequency of each possible value of the variable, therefore, we define sets of classes of that variable to calculate the frequency.

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

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

$$\mathbf{R} = x_\ell - x_s$$

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

$$\text{Length} = \frac{\text{Range+one measuring unite}}{\text{No. of classes}} \quad \Leftrightarrow \quad \mathbf{C} = \frac{\mathbf{R}\text{+ one measuring unite}}{k}$$

c. If the length of a class is calculated, using the above formula comes out to be a fraction value t. Then (for easier viewing of classes by data dump) 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 is calculated by adding the number $(\mathbf{C}-1)$ to the lower limit of the current class. The lower limit of the next class limit is 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 $(\mathbf{C}-1)$ 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 $\mathbf{C}$, which is previously calculated is for the class boundaries.

f. It is common practice to represent the classes in a frequency distribution 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:

$$\text{Class midpoint} = \frac{\text{Upper limit of the class} + \text{Lower limit of the class}}{2}$$

g. Determine the frequency value of a class i. This value (we denote it by f_i) equal to the number of data in the interval $\left[b_{i-1}, b_i\right)$, where b_{i-1} is the lower limit of the class boundary i, and b_i is the upper limit of the class boundary i.

h. Determine the value of the ascending cumulative frequency (ACF) of a class i. This value (we denote it by F_i) gives us the value of how many items are less than the upper limit of the class boundary i.

REMARKS 1.2.1

1. When raw data are dumped in a frequency distribution table, they are emptied in the class boundaries.

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:

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

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

We note here that it is unwanted to dump the data in a frequency distribution table if the number of data is less than 32.

▶ **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.

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:

$$\mathbf{R} = 25 - 1 = 24$$

Therefore, the length of class boundary is given by:

$$\mathbf{C} = (24+1)/5 = 5$$

Therefore, the length of class limit equals to $\mathbf{C} - 1 = 5 - 1 = 4$.

Table 1.2.5-a

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Less than	Ascending Cumulative Frequency (ACF) F_i
1 - 5	$0.5 \to 5.5$	3	7	0.14	5.5	**7**
6 - 10	$5.5 \to 10.5$	8	9	0.18	10.5	7+9 = **16**
11 - 15	$10.5 \to 15.5$	13	14	0.28	15.5	7+9 +14 = **30**
16 - 20	$15.5 \to 20.5$	18	12	0.24	20.5	7+9 +14+12 = **42**
21 - 25	$20.5 \to 25.5$	23	8	0.16	25.5	7+9 +14+12 8 = **50**
Total	----------------	--------	50	1	-------	----------------------------

Table 1.2.5-b

Class Boundaries	Frequency	greater than	Descending Cumulative Frequency (DCF) Φ_i
$0.5 \to 5.5$	7	0.5	**50**
$5.5 \to 10.5$	9	5.5	50-7 = **43**
$10.5 \to 15.5$	14	10.5	50-9-7 = **34**
$15.5 \to 20.5$	12	15.5	50-14-9-7 = **20**
$20.5 \to 25.5$	8	20.5	50-12-14-9-7 = **8**
---------------	50	-------	---------------------------

EXAMPLE 1.2.9: Consider the mileage of 40 cars per liter of fuel in a particular city, so we get the following results:

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 distribution table for this data in the following manner:

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

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

$$k = \lfloor 3.322 \log n \rfloor = \lfloor 3.322 \log 40 \rfloor = \lfloor 5.322 \rfloor = 5$$

So the class boundary length equals to $\mathbf{C} = (8+1)/5 = 1.8$

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

Table 1.2.6

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency F_i	Descending Cumulative Frequency Φ_i
12-13	$11.5 \to 13.5$	12.5	8	0.20	8	40
14-15	$13.5 \to 15.5$	14.5	10	0.25	18	32
16-17	$15.5 \to 17.5$	16.5	12	0.30	30	22
18-19	$17.5 \to 19.5$	18.5	6	0.15	36	10
20-21	$19.5 \to 21.5$	20.5	4	0.10	40	4
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 F_i	Ascending Cumulative Relative Frequencies	Ascending Cumulative Percentages Frequencies
11.5→13.5	8	8	$8/40 = \mathbf{0.20}$	$0.20\times100 = \mathbf{20}$
13.5→15.5	10	18	$18/40 = \mathbf{0.45}$=0.20+0.25	$0.45\times100 = \mathbf{45}$
15.5→17.5	12	30	$30/40 = \mathbf{0.75}$=0.45+0.30	$0.75\times100 = \mathbf{75}$
17.5→19.5	6	36	$36/40 = \mathbf{0.90}$=0.75+0.15	$0.90\times100 = \mathbf{90}$
19.5→21.5	4	40	$40/40 = \mathbf{1}.00$=0.90+0.10	$1.00\times100 = \mathbf{100}$
Total	40	----------	---------	---------

◀

Section 1.3

GRAPHICAL REPRESENTATIONS

DEFINITION 1.3.1 (Pie chart)

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.

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

$$\textbf{(The relative frequency of the class } (i)) \times (360) = \ldots \textbf{ degree}$$

EXAMPLE 1.3.1: To draw the pie chart for example 1.2.6 we have the central angle for each category as follows:

For category (AB) the measure angle is $0.05 \times 360 = 18$ **deg**

For category (B) the measure angle is $0.10 \times 360 = 36$ **deg**

For category (A) the measure angle is $0.45 \times 360 = 162$ **deg**

For category (O) the measure angle is $0.40 \times 360 = 144$ **deg**

Therefore, the pie chart for example 1.2.6 is given as follows:

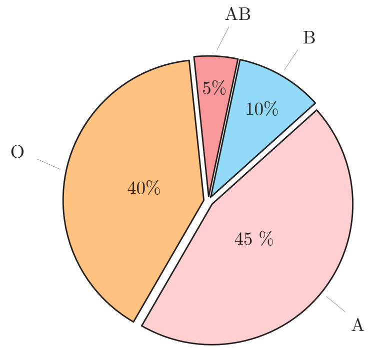


Figure 1.3.1 (Pie Chart for example 1.2.6)

The following graph method is a simple representation of classes and their frequencies with provide a simple way to compare the frequencies of different classes.

DEFINITION 1.3.2 (Bar chart)

A bar chart is a representation of data of discrete variable with finite values (qualitative or quantitative). This is done through vertical or horizontal bars; so that it draws over each statement a bar with height (or length) equals to the frequency of that statement.

REMARK 1.3.1

1. Note that if the values of the variable are large, it is useless for these graphs.
2. The width of the bar doesn't matter, but the width of bars must be uniform. Moreover, the bars must be separated from each other.

Below is the bar chart representation for the data in Example 1.2.6:

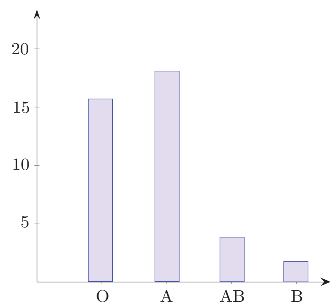


Figure 1.3.2-a (Vertical Bar Chart)

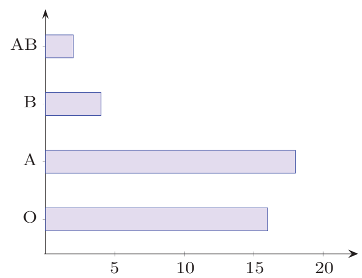


Figure 1.3.2-b (Horizontal Bar Chart)

DEFINITION 1.3.3 (Two directional bar chart)

Two directional bar chart is a bar chart, where we can represent both positive and negative values of different classes.

▶ **EXAMPLE 1.3.2:** In the following table, we consider the 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 %

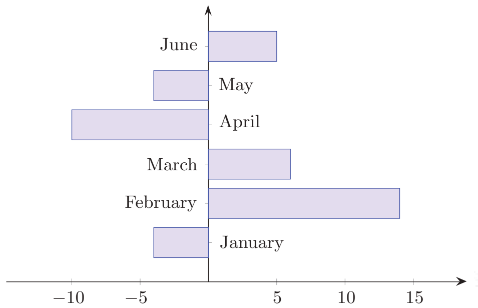


Figure 1.3.3 (Two Directional Bar Chart)

DEFINITION 1.3.4 (Multiple bar chart)
A multiple bar chart is a bar chart, where we can use it 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:

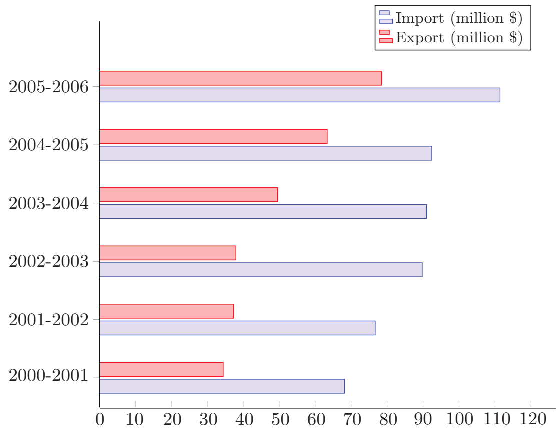


Figure 1.3.4 (Horizontal Bar Chart)

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

DEFINITION 1.3.5 (Component, or Stacked bar chart)
Component bar chart is a bar chart, where we can represent each component by a section in the bar, whose size is proportional to its contribution in the class.

Such bar charts when each class or category has components, which make up 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:

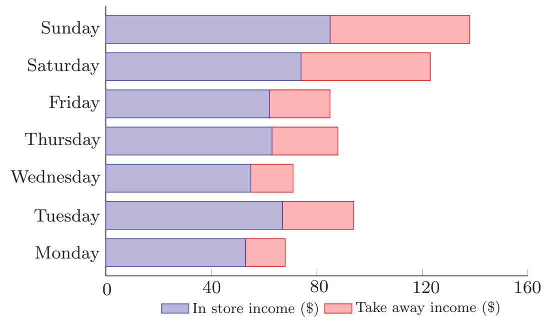


Figure 1.3.5 (Horizontal Bar Chart)

◀

DEFINITION 1.3.6 (Histogram)
A histogram is a graphical display used for data generated by continuous variables. It is a graph in which class boundaries are marked on the horizontal axis and the frequencies are marked on a vertical axis, and is constructed by drawing a rectangular column above each actual category so that its height equals the frequency of that category.

From this definition, we note that the histogram is similar to the bar chart but the classes of the variable are adjacent to each other and the rectangular bars must touch each other.

EXAMPLE 1.3.5:

Consider the data from Example 1.2.9, where we have:

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

Then the histogram for the given data is as follow:

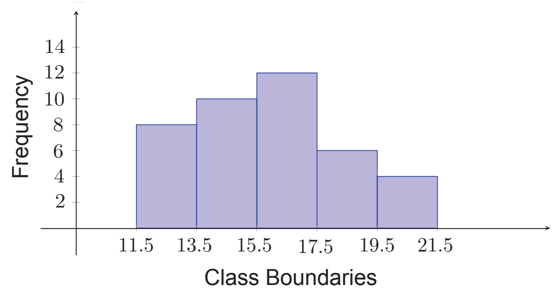


Figure 1.3.6

REMARKS 1.3.2

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

I- Symmetric Histogram:

The symmetrical Histogram is characterized by being divided into two identical sections with respect to a column on the middle of its base.

In such types of histograms, the bins in the middle have higher frequency and the frequency values keep 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 sides 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.

Uniform: Histogram with no peaks (all classes have the same frequency).

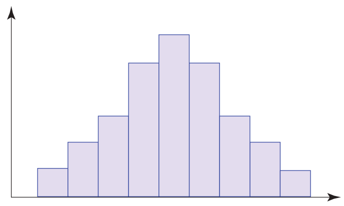


Figure 1.3.7-a (Unimodal Histogram)

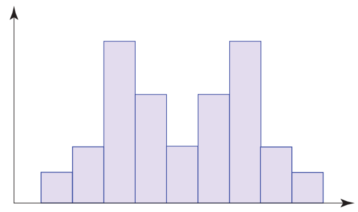


Figure 1.3.7-b (Bimodal Histogram)

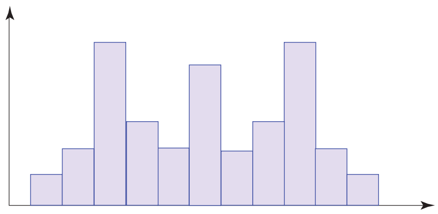


Figure 1.3.7-c (Multimodal Histogram)

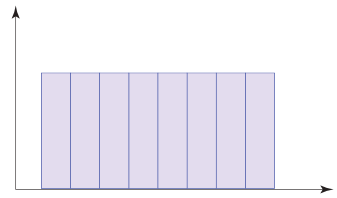


Figure 1.3.7-d (Uniform Histogram)

II- Skewed Histogram:

Before talking about twisted distributions, we will provide 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 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 a left skewed histogram is one, which has longer tail on the left side.

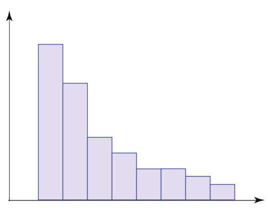


Figure 1.3.8-a (Right Skewed Histogram)

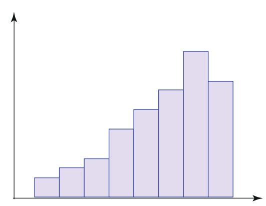


Figure 1.3.8-b (Left Skewed Histogram)

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 are the midpoint and the frequency of class boundary i respectively, and it closes from the left to the middle of a default class boundary located before the first class boundary, and from the right to the middle of a default class boundary located after the last class boundary.

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.

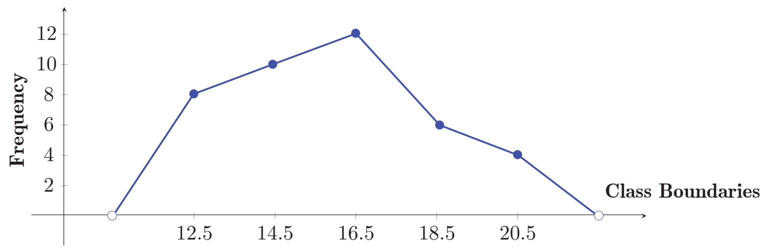


Figure 1.3.9 (Polygon frequency chart for mileage data)

DEFINITION 1.3.9 (Ascending Cumulative Frequency Polygon (ACFP))

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

The ascending cumulative frequency polygon for the data from Example 1.2.9 is presented in the following graph.

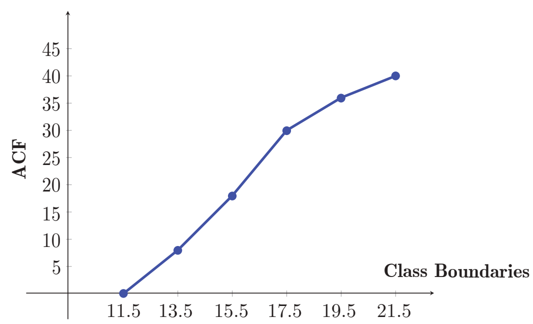


Figure 1.3.10-a (ACFP for mileage data)

DEFINITION 1.3.10 (Descending Cumulative Frequency Polygon (DCFP)):

The descending cumulative frequency polygon (DCFP) is a polygon which connects with a straight line the points (b_i, Φ_i), whereas b_i and Φ_i are the lower bound and the descending cumulative frequency of class boundary i respectively, and closes from the right to the end of the last class boundary.

The descending cumulative frequency polygon for the data from Example 1.2.9 is presented in the following graph.

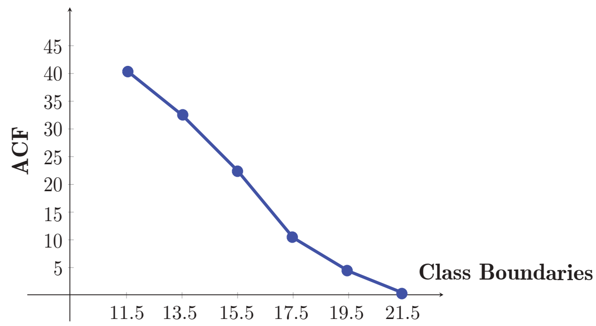


Figure 1.3.10-b (DCFP for mileage data)

REMARK 1.3.3

The ascending cumulative frequency polygon is known as the less than ogive, and the descending cumulative frequency polygon is known as the greater than ogive.

Section 1.4

MEASURES OF CENTRAL TENDENCY AND POSITION

In the previously methods, we became familiar with the concepts that help us to organize raw data in tables and graphs. In fact, 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 percentiles, deciles and quartiles of the data. Some of these are only defined for quantitative data but some are defined for both.

We will start these central tendency measures with the following measure, which is defined only for quantitative data.

DEFINITION 1.4.1.a (Mean for Raw Data)

Let $x_1, x_2, \ldots, x_n$ be values of quantitative variable. Then the mean of these values (one denote it by $\overline{x}$) is defined by the following relation:

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

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: Calculate the mean of the following data:

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

According to the definition of the mean, we have:

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

Hence, the mean of the above data is 11.

DEFINITION 1.4.1.b (Mean for Organized Data in a Frequency Table)

We suppose that the data (of a discrete variable) are given by the following Frequency table:

Table 1.4.1

i	Values	Frequency
1	x_1	f_1
2	x_2	f_2
$\vdots$	$\vdots$	$\vdots$
m	x_m	f_m
Total	--------	$\sum f_i = n$

Then, the mean for these data (we denote it by $\overline{x}$) is given by the following relation:

$$\overline{x} = \frac{1}{\sum f_i} \sum_{i=1}^{m} f_i\, x_i = \frac{1}{n} \sum_{i=1}^{m} f_i\, x_i$$

EXAMPLE 1.4.2: Let us consider the following data (No. of subjects in which student failed):

Table 1.4.2

i	No. of subjects in which student failed	Frequency
1	0	8
2	1	18
3	2	12
$m = 4$	3	2
Total	----------	40

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

$$\overline{x} = \frac{1}{\sum f_i} \sum_{i=1}^{m} f_i\, x_i = \frac{(0\times 8)+(1\times 18)+(2\times 12)+(3\times 2)}{8+18+12+2} = \frac{0+18+24+6}{40} = \frac{48}{40} = 1.2$$

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.

DEFINITION 1.4.1.c (Mean for Organized Data in a Frequency Distribution Table)

We suppose that the data (of a continuous variable) are given by the following frequency distribution table:

Table 1.4.3

i	Class Boundaries	Class Midpoint x_i	Frequency f_i	Ascending Cumulative Frequency (ACF)
1	$b_0 \to b_1$	x_1	f_1	$F_1 = f_1$
2	$b_1 \to b_2$	x_2	f_2	$F_2 = f_1 + f_2$
$\vdots$	$\vdots \quad \vdots \quad \vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots \quad \vdots \quad \vdots$
k-1	$b_{k-2} \to b_{k-1}$	x_{k-1}	f_{k-1}	$F_{k-1} = f_1 + f_2 + \dots + f_{k-1}$
k	$b_{k-1} \to b_k$	x_k	f_k	$F_k = f_1 + f_2 + \dots + f_k$
Total	-----------	--------	$\sum f_i$	-----------------

Then, the mean for these data (we denote it by $\overline{x}$) is given by the following relation:

$$\overline{x} = \frac{1}{\sum f_i} \sum_{i=1}^{k} f_i \, x_i$$

The example below will help the reader in better understanding.

▶ **EXAMPLE 1.4.3 (For Continuous quantitative data):** Let us consider the following data (the mileage of 40 cars per liter of fuel in a particular city):

Table 1.4.4

i	Class Boundaries	Class Midpoint	Frequency
1	11.5→13.5	12.5	8
2	13.5→15.5	14.5	10
3	15.5→17.5	16.5	12
4	17.5→19.5	18.5	6
$k = 5$	19.5→21.5	20.5	4
Total	------------	----------	**40**

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

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

◀

DEFINITION 1.4.2 (Weighted Mean)

Let $x_1, x_2, \dots, x_\ell$ be observed values and their weights are $w_1, w_2, \dots, w_\ell$ respectively. Then the weighted mean (it is a mean, and we denote it by $\overline{x}$) of this data is given by the following relation:

$$\overline{x} = \frac{1}{\sum w_i} \sum_{i=1}^{\ell} w_i x_i$$

One notes that the weighted mean is equal to the ordinary mean when all the weights are equal.

In a weighted mean, some of the observations are given more weight than the others. This is used when some values are more significant than others.

▶ **EXAMPLE 1.4.4:** Consider a student in King Saud University received the following grades in the first semester:

Courses	Grades	Credit hours
Mathematics	B	4
Statistics	A	3
English	C	3
Physics	C	4

Where grades $A = 5$, $B = 4$, $C = 3$ and $D = 2$ points.

Now to calculate the Grade Point Average (GPA) for this student in this semester we have

Courses	Grades	Points (x_i)	Credit Hours (w_i)	w_i x_i
Mathematics	B	4	4	16
Statistics	A	5	3	15
English	C	3	3	9
Physics	C	3	4	12
Total	------------	----------	14	52

Then, the Grade Point Average is calculated as:

$$\text{GPA} = \bar{x} = \frac{1}{\sum w_i} \sum_{i=1}^{\ell} w_i x_i = \frac{52}{14} = 3.71$$

◀

▶ **EXAMPLE 1.4.5:** 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.

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.

From the previous example we note that the weighted mean can be used to compare between the objects in hand or to make decisions using this measure.

Advantages of The Mean

- It is quick and easy to compute.
- All values are considered by calculating the mean.
- 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 1.4.3 (Median)
Median (we denote it by $\tilde{x}$) is that value which divides the data in two halves after ordering them, in ascending or descending order.

One notes that a 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.

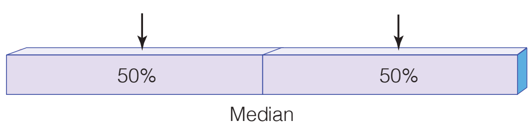


Figure 1.4.1: The concept of median

CALCULATING THE MEDIAN

One calculates the median as follow:

For Raw Data: To calculate the median for given data we must arrange the data in the increasing order, and suppose it $x_1, x_2, ..., x_n$. Then the median is given by the following relation:

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

EXAMPLE 1.4.6: 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 1.4.7: 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:

$$\begin{array}{cccccccc} 21, & 22, & 23, & 24, & 26, & 28, & 29, & 35 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \end{array}$$

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 less than 25 and 4 observations greater than 25.

For Organized Data in a Frequency Table (*data of a discrete variable*)**:** We suppose that the data are given by the Frequency table 1.4.1

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 $\sum f_i = n$ is odd, then we consider the value x_i with the smallest cumulative frequency greater than or equal to $\frac{n+1}{2}$ as median of data.
- If $\sum f_i = n$ is even, then:

a. If a value x_i has cumulative frequency equal to $\frac{n}{2}$ exactly, we will take the mean of values x_i and x_{i+1} as median of data.

b. If a value x_i has cumulative frequency greater than $\frac{n}{2}$, and the values x_{i-1} has cumulative frequency less than $\frac{n}{2}$, we will take the value x_i itself as median of data

EXAMPLE 1.4.8: Consider the following data:

Table 1.4.5-a

Subjects	No. of subjects in which student failed	Frequency (Number of students whose failed in the subject)
English	2	12
Mathematics	1	18
Statistics	0	8
Chemistry	3	2

We arrange the representative values of the data in the given table, so we get the following frequency table:

Table 1.4.5-b

i	Subjects	No. of subjects in which student failed	Frequency	ACF
1	Statistics	0	8	**8**
2	Mathematics	1	18	8+18 = **26**
3	English	2	12	26+12 = **38**
$m=4$	Chemistry	3	2	38 +2 = **40**
Total		---------	40	------------

For this example, we have $n = 40$ (even number). Therefore, the smallest cumulative frequency greater than or equal to $\frac{n}{2} = 20$ is 26. So we get that the value of the data corresponding to that cumulative frequency is 1. Hence, the median is 1.

For Organized Data in a Frequency Distribution Table (*data of a continuous variable*)**:**

We conceder data in a frequency distribution table as in Table 1.4.3. 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 is greater than or equal to $\frac{1}{2}\sum f_i$. Such class is called median class.

- We use the following formula to calculate the median for continuous quantitative data:

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

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

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

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

C is the class length of the median class boundary.

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

▶ **EXAMPLE 1.4.9:** We will calculate the median for the data in example 1.4.3.

Table 1.4.6

i	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency
1	11.5→13.5	12.5	8	0.20	8
2	13.5→15.5	14.5	10	0.25	18
3	15.5→17.5	16.5	12	0.30	30
4	17.5→19.5	18.5	6	0.15	36
5	19.5→21.5	20.5	4	0.10	40
Total		-------	40	1	-----------

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

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

Where we have, $\tilde{F} = 30$, $\tilde{f} = 12$ and $\mathrm{C} = 2$, then we get:

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

◀

Advantages of The Median

- It is easy to compute and understand.
- It is not affected by outliers or extreme values.
- It can be used even if you loss some data (known argument) that is not in the middle.

Disadvantages of The Median

- It does not take all values into account.
- It is not used in many statistical tests.
- It cannot be identified for qualitative data

DEFINITION 1.4.4 (The Mode)
The mode (we denote it by $\hat{x}$) of data is a value or observation, which has the highest frequency.

CALCULATING THE MODE

One finds or calculates the mode(s) as follow:

For Raw Data: To find the mode for raw data we take the value or observation that have the highest frequency. The following examples illustrate this concept.

▶ **EXAMPLE 1.4.10:** We consider the following data that represents grades of 12 students in an exam:

A, A, C, A, D, A, B, B, C, D, A, B

So we find that the mode of this data is A.

◀

▶ **EXAMPLE 1.4.11:** The following data represent the time spent in 10 Km race for ten bicycle runners:

28 22 26 29 21 23 28 28 25 29

If we put this data in a frequency table, then we have:

Table 1.4.7

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

For this data, we note that the mode is equal to 28.

◀

▶ **EXAMPLE 1.4.12:** Consider the following data representing the age (in years) of 14 students:

12	11	13	14	13	12	11
12	13	12	12	13	14	13

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 1.4.13:** The data below represents the outcome when a die is rolled 40 times.

1	5	2	4	6	3	3	5
3	1	5	2	4	5	5	6
1	6	2	4	2	6	2	3
2	1	5	2	6	1	1	1
2	4	2	3	4	5	6	2

So, we note that the highest frequency is 10 for the observation 2. Therefore, the mode for the given data is the value 2.

◀

▶ **EXAMPLE 1.4.14:** The finishing times of 7 bike racers who had finishing times as:

28 22 26 29 21 23 24

If we put this data in a frequency table, then we have:

Table 1.4.8

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

In this example we note that all data have the same frequency 1. Therefore, the given data have no mode.

◀

For Organized Data in a Frequency Table (*data of a discrete variable*)**:** We suppose that the data are given by the Frequency table 1.4.1 or by the following table for qualitative data:

Table 1.4.9

i	Observations Qualitative data	Frequency
1	α_1	f_1
2	α_2	f_2
$\vdots$	$\vdots$	$\vdots$
m	α_m	f_m

Then we take the values or observations that have the highest frequency as mode(s).

The following examples illustrate this concept.

▶ **EXAMPLE 1.4.15:** The following table represents marks (out of 10) obtained by 20 students:

Table 1.4.10

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

So, we note that the mode of the marks obtained by students is 9.

EXAMPLE 1.4.16: We consider the data in example 1.2.5 (the data about the blood group of 40 persons), which we can display in the following table:

Table 1.4.11

Blood group	O	A	B	AB
Frequency	16	18	4	2

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

For Organized Data in a Frequency Distribution Table (*data of a continuous variable*):

We conceder data in a frequency distribution table as in Table 1.4.3. Then for continuous quantitative data, the mode is calculated by the following steps:

Look carefully at the values of frequencies, then the class whose frequency is greater than the frequency of the previous and subsequent classes directly is a modal class.

For each modal class we use the following relation to calculate the value of its mode.

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

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

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

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

C is the class length of the modal class boundary.

REMARK 1.4.1

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

The following examples illustrate this concept.

EXAMPLE 1.4.17: Consider the time taken by 24 persons in a certain run race.

Table 1.4.12

Seconds	Class Boundaries	Frequency
1	50.5→55.5	2
2	55.5→60.5	7
3	60.5→65.5	8
4	65.5→70.5	4
5	70.5→75.5	3

We note that then the class which has frequency greater than the frequency of the previous and subsequent classes directly is the third class, and this class is unique. Therefore, the modal class for the given table is the class 60.5-65.5. So, we have:

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

EXAMPLE 1.4.18: The minutes spent per week by the teenagers in watching movies are given by the following table:

Table 1.4.13

Number of Minutes per week	Number of Teenagers
$0 \to 90$	26
$90 \to 180$	32
$180 \to 270$	65
$270 \to 360$	75
$360 \to 450$	60
$450 \to 540$	42
Total	300

Then we find the class modal for the above data is 270→360 because it has the highest frequency 75. Therefore, we have:

$$\hat{x} = \hat{L} + \frac{d_1}{d_1 + d_2} C = 270 + \frac{10}{10+15} \times 90 = 306$$

REMARKS 1.4.2

a. One can use the mode for qualitative and quantitative data.

b. In case if the highest frequency is constant for all data, then the data is said to have no mode.

c. 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.

Advantages of The Mode

- It is quick and easy to compute.
- It can be evaluated for both quantitative and qualitative data.
- It is not affected by extreme values.

Disadvantages of The Mode

- There may be more than one mode for a certain data set.
- Sometimes, there is no mode for a given data set.
- It may not reflect the center of the distribution very well.

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 distribution of data is symmetric and asymmetric (or skewed) and learn the relationship between these measures.

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

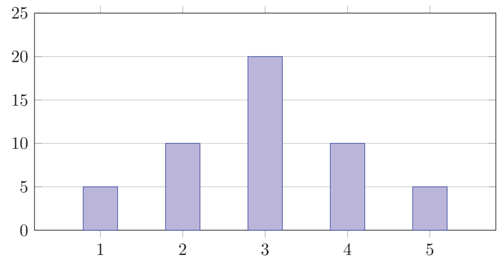


Figure 1.4.2 (A symmetric distribution with mode = mean = median = 3)

For this data we have: Mode = Mean = Median = 3.

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

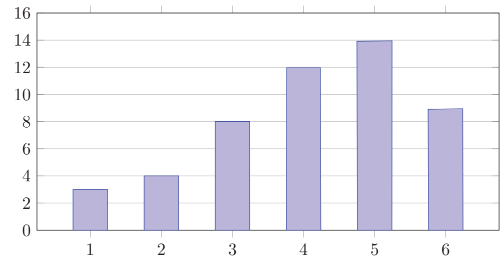


Figure 1.4.3 (A left skewed distribution with mode greater than median greater than mean)

For the above data, the mode is 5, the median is 4 and the mean is 4.14.

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

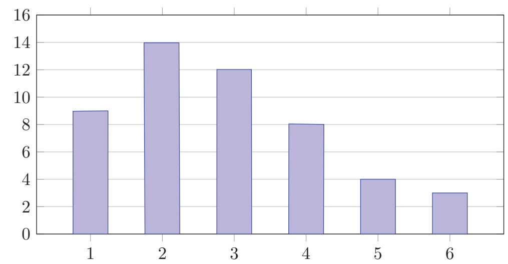


Figure 1.4.4 (A right skewed distribution with mean greater than median greater than mode)

For the above distribution of data, the mean is 2.86, the median is 3 and the mode is 2.

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

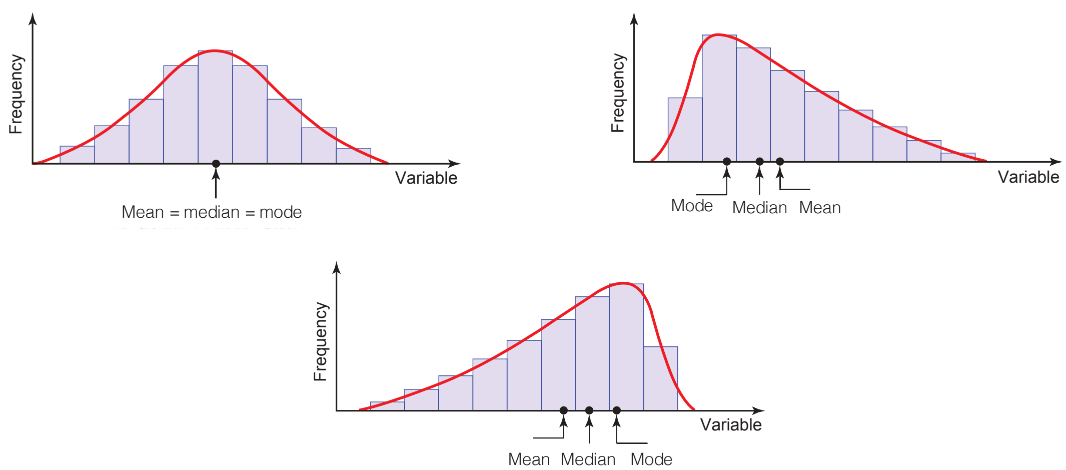


Figure 1.4.5 (The relationship among the three measures of tendency)

DEFINITION 1.4.5 (Percentiles)

The percentiles (denoted by P_1, P_2, ...and P_{99}) of a variable divide the ordered observed values into 100 equal parts. Median is the 50th percentile and it divides ordered data into two equal halves. The percentile P_1 divides the ordered observed data into 1% from bottom and 99% from top. Similarly, any j^{th} percentile, P_j divides the ordered 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.

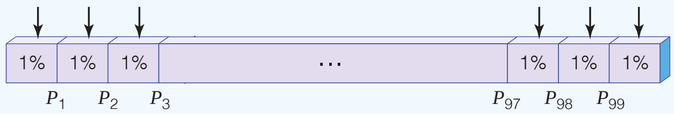


Figure 1.4.6 (The concept of percentiles)

How can we calculate the percentiles?

Let x_1, x_2, ..., x_n be arranged data. Then:

- We calculate the **rank** of r^{th} percentile (we denote it by p_r) by the following relation:

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

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

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

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

▶ **EXAMPLE 1.4.19:** Calculate the 35th 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:

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

◀

Note that the percentiles are more useful when the number of data is large (the number of data $n \geq 99$). Some commonly used percentiles are deciles and quartiles.

DEFINITION 1.4.6 (Deciles)

The deciles (denoted by D_1, D_2, ...and D_9) divide the ordered data into 10 equal parts. D_1 is the 10^{th} percentile; D_2 is 20^{th} percentile and so on till D_9 which is the 90^{th} percentile. The following graph explains the concept of deciles.

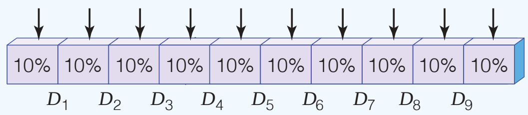


Figure 1.4.7 (The concept of deciles)

How can we calculate the deciles?

Let x_1, x_2, ..., x_n be arranged data. Then:

- We calculate the **rank** of r^{th} decile, whose denoted by d_r, and is calculated by the following relation:

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

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

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

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

For example, refer to the example 1.4.19 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:

$$D_6 = x_k + s\,(x_{k+1} - x_k) = x_{18} + 0.6\,(x_{19} - x_{18}) = 60 + 0.6(70 - 60) = 66$$

DEFINITION 1.4.7 (Quartiles)

The Quartiles (denoted by Q_1, Q_2 and Q_3) divide the ordered data into 4 equal parts. The first quartile Q_1 is 25^{th} percentile, the second quartile Q_2 is 50^{th} percentile which is also the median of the data and the third quartile Q_3 is 75^{th} percentile. The following graph explains the concept of quartiles.

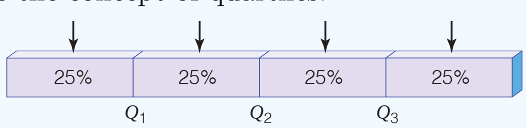


Figure 1.4.8 (The concept of quartiles)

How can we calculate the quartiles?

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

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

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

- The r^{th} **quartile** Q_r can be calculated by the following relation:

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

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 1.4.20: Find the quartiles for the data given below:

28 22 26 29 21 23 24

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 1.4.21:** Find the quartiles of the data given below:

39	40	41	56	7	8	15	36

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:

$$q_1 = 2.25 \quad \Rightarrow \quad Q_1 = 8 + 0.25(15 - 8) = 9.75$$
$$q_2 = 4.5 \quad \Rightarrow \quad Q_2 = 36 + 0.5(39 - 36) = 37.5$$
$$q_3 = 6.75 \quad \Rightarrow \quad Q_3 = 40 + 0.75(41 - 40) = 40.75$$

◀

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

DEFINITION 1.4.8 (Extreme Value)
We say that a value x of given data is said to be extreme if one of the following relations is realizing:

$$x < LF := \mathrm{Q}_1 - 1.5(\mathrm{Q}_3 - \mathrm{Q}_1) \quad [LF \text{ is the abbreviation of “Lower Fence”}]$$

or

$$x > HF := \mathrm{Q}_3 + 1.5(\mathrm{Q}_3 - \mathrm{Q}_1) \quad [HF \text{ is the abbreviation of “High Fence”}]$$

▶ **EXAMPLE 1.4.22:** Refer to the example 1.4.21, we find:

$$\mathrm{Q}_1 - 1.5(\mathrm{Q}_3 - \mathrm{Q}_1) = 9.75 - 1.5(40.75 - 9.75) = -36.75$$
$$\mathrm{Q}_3 + 1.5(\mathrm{Q}_3 - \mathrm{Q}_1) = 40.75 + 1.5(40.75 - 9.75) = 87.25$$

So, the given data haven't an extreme value.

◀

DEFINITION 1.4.9 (Five Numbers)
Five Numbers are a summary of the variable data which includes the below mentioned five characteristics:

Smallest value, Q_1, Q_2, Q_3 and Largets value

The five numbers summary for example 1.4.20 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 three quartiles and join them making the box.
- Calculating the end of right whisker as follow:
  a. If the data have not great extreme value, then the end of the right whisker stops at the greatest value in the data.

  b. If the data have great extreme value, then the end of the right whisker stops at HF.
- Calculating the end of left whisker as follow:
  a. If the data have not small extreme value, then the end of the left whisker stops at the smallest value in the data.

  b. If the data have small extreme value, then the end of the left whisker stops at LF.

REMARK 1.4.3

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

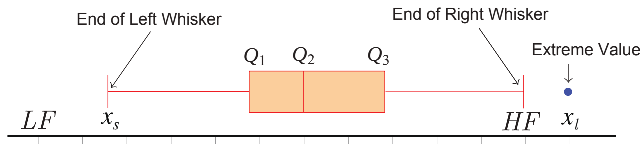


Figure 1.4.9

The box plot for the data given in Example 1.4.21 with five numbers summary are 7, 9.75, 37.5, 40.75 and 56 is given below

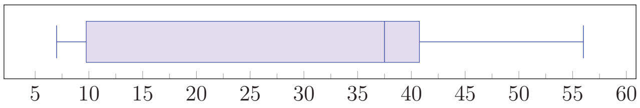


Figure 1.4.10 (Box plot for example 1.4.21)

For continuous data (as in frequency distribution tables) one can use box plot to determine whether the distribution of data is symmetric or skewed.

The following graph represents three different sets of data displaying the symmetric and the skewed distribution of data.

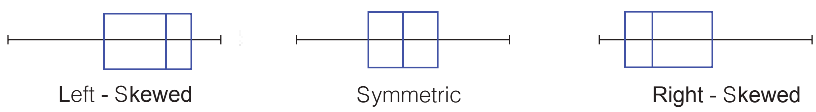


Figure 1.4.11 (Types of data using box plot)

Section 1.5

MEASURES OF VARIABILITY

In statistics, variability measures give us the information about how the data is spread out, and how to compare the dispersion of two sets of data or more. With the dispersion measures we can tell whether the data is stretched or squeezed. In other words, measures of dispersion provide us how the data is distributed about its central tendency measure. With the variation measures we can tell which of two sets of data has a greater dispersion. In fact, measures of central tendency and dispersion together provide a good summary of the data in hand.

We will now provide some measures of dispersion.

DEFINITION 1.5.1 (Variance for raw data)

Let $x_1, x_2, ..., x_n$ be raw data with mean $\overline{x}$ and $n \geq 2$. Then the variance of this data (one denote it by S^2) is given by the following relation:

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

DEFINITION 1.5.2 (Standard Deviation)

The standard deviation (one denotes it by S) is the positive square root of the variance and calculated by the following relation:

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

REMARKS 1.5.1

a. If the number of data n equal to 1, then we put $S^2 = 0$.

b. The variance value is expressed in square units

c. We can also calculate the variance of raw data by using the following formula:

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

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

d. Standard deviation is the best measure of dispersion.

e. The dispersion measure 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).

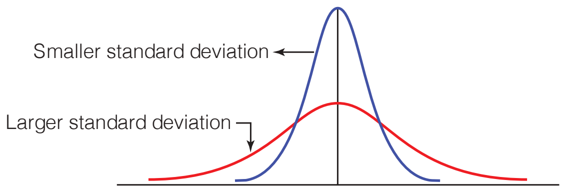


Figure 1.5.1

▶ **EXAMPLE 1.5.1:** Let 2, 3, 6, 8, 10, 13 and 14 be given data. Then the mean of this data is:

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

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

Table 1.5.1

Variable values	Squared Variable values	Deviation from mean $\left(x_i - \overline{x}\right)$	Squared Deviation $\left(x_i - \overline{x}\right)^2$
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	$\sum_{i=1}^{7} (x_i - \overline{x})^2 = 130$

Therefore, we have:

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

We can solve the above example using the formula in 1.5.1-c and calculate the standard deviation and the variance.

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

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

In the latter case, it is possible to obtain a value different from the previous value because the numbers are rounded during the calculations.

DEFINITION 1.5.3 (Variance for grouped data in a frequency table)

We suppose that the data are given by the Frequency table 1.4.1. Then, the variance for these data is given by the following relation:

$$S^2=\frac{1}{\left(\sum_{i=1}^{m}f_i\right)-1}\sum_{i=1}^{m}f_i\left(x_i-\overline{x}\right)^2$$

▶ **EXAMPLE 1.5.2:** Let us consider the following data:

Table 1.5.2

i	Subjects	Frequency
1	2	6
2	5	10
3	7	16
4	12	8
Total	----------	40

The mean of the data for the above table is calculated as:

$$\overline{x}=\frac{1}{\sum f_i}\sum_{i=1}^{m}f_i\,x_i=\frac{(6\times 2)+(10\times 5)+(16\times 7)+(8\times 12)}{40}$$
$$=\frac{12+50+112+96}{40}=\frac{270}{40}=6.75$$

We calculate the variance by the following relation:

$$S^2=\frac{1}{\left(\sum_{i=1}^{m}f_i\right)-1}\sum_{i=1}^{m}f_i\left(x_i-\overline{x}\right)^2$$
$$=\frac{6(2-6.75)^2+10(5-6.75)^2+16(7-6.75)^2+8(12-6.75)^2}{39}=\frac{387.5}{39}=9.94$$

Therefore, the standard deviation becomes that:

$$S = +\sqrt{S^2} = +\sqrt{9.94} = 3.15$$

DEFINITION 1.5.4 (Variance for grouped data in a frequency distribution table)
We suppose that the data are given by the Frequency table 1.4.3. Then, the variance for these data is given by the following relation:

$$S^2 = \frac{1}{\left(\sum_{i=1}^{m} f_i\right) - 1} \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2$$

EXAMPLE 1.5.3: Let us consider the following frequency distribution table:

Table 1.5.3

Class boundary	Midpoint	Frequency
$4 \to 10$	7	8
$10 \to 16$	13	10
$16 \to 22$	19	18
$22 \to 28$	25	12
$28 \to 34$	31	2
Total		50

The mean of the data for the above table is calculated as:

$$\begin{aligned}\overline{x} &= \frac{1}{\sum f_i} \sum_{i=1}^{m} f_i\, x_i = \frac{(8 \times 7) + (10 \times 13) + (18 \times 19) + (12 \times 25) + (2 \times 31)}{50} \\ &= \frac{56 + 130 + 342 + 300 + 62}{50} = \frac{890}{50} = 17.8\end{aligned}$$

We calculate the variance by the following relation.

$$\begin{aligned}S^2 &= \frac{1}{\left(\sum_{i=1}^{m} f_i\right) - 1} \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2 \\ &= \frac{8\,(7 - 17.8)^2 + 10\,(13 - 17.8)^2 + 18\,(19 - 17.8)^2 + 12\,(25 - 17.8)^2 + 2\,(31 - 17.8)^2}{49} \\ &= \frac{2160}{49} = 44.08\end{aligned}$$

Therefore, the standard deviation becomes that:

$$S = +\sqrt{S^2} = +\sqrt{44.08} = 6.64$$

REMARKS 1.5.2

a. We know that variance is the sum of squares of deviations and since squares are always positive or zero, therefore, we have $S^2 \geq 0$.

b. Since the standard deviation is the positive square root of the variance, then $S \geq 0$.

c. 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 (The variable itself is a constant map - or constant function-).

DEFINITION 1.5.5 (Range for raw data)

We have introduced the definition of the range earlier when building a frequency distribution table, were we had $\mathbf{R} = x_\ell - x_s$. Whereas x_l is the greatest value of data, and x_s is the smallest value of data.

DEFINITION 1.5.6 (Range for organized data in a frequency table)

We suppose that the data are given by the Frequency table 1.4.1. Then, the range for these data is given by the following relation:

$$\mathbf{R} = x_m - x_1$$

Where x_m is the value of the last subject, and x_1 is the first subject in the frequency table.

EXAMPLE 1.5.4: Let us consider the following data:

Table 1.5.4

i	Subjects	Frequency
1	2	4
2	5	10
3	7	16
$m = 4$	12	8
Total	----------	38

The range of the number of subjects is calculated as:

$$\mathbf{R} = x_m - x_1 = 12 - 2 = 10$$

DEFINITION 1.5.7 (Range for grouped data)
We conceder grouped data in a frequency distribution table as in Table 1.4.3. 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 1.5.5:

1. We conceder 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:

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

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

2. We consider the data in the table 1.5.3 (frequency distribution table). Then we find the range of data equal to $R = x_k - x_1 = 31 - 7 = 24$.

DEFINITION 1.5.8 (Interquartile Range)
The Interquartile Range (one denotes it by IQR) of given data is defined as the difference between the third quartile and the first quartile.

$$IQR = Q_3 - Q_1$$

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

Example 1.5.6: Find the Interquartile range of the data given in Example 1.4.20.

Solution: In the example 1.4.20 we calculated the quartiles of the given data which were as $Q_1 = 22$, $Q_2 = 24$ and $Q_3 = 28$. Therefore, we get that:

$$IQR = Q_3 - Q_1 = 28 - 22 = 6$$

DEFINITION 1.5.9 (Coefficient of Variation)
Let $x_1, x_2, \ldots, x_n$ be raw data with mean $\overline{x} \neq 0$ and standard deviation S. Then the coefficient of variation (we denote it by CV) is calculated as:

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

The coefficient of variation is a useful measure of variation to compare between sets of data with different units (measures).

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 1.5.10 (*z*-scores)

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

$$z_{x_i} = \frac{x_i - \overline{x}}{S}$$

This numerical value is used to assign a degree of data (or observation) that distinguishes its position relative to the rest of the data (or observations), but its meaning is more apparent when we compare the standard of two sets of data. The following examples illustrate this concept.

▶ **EXAMPLE 1.5.7:** Let 2, 5, 3, 3, 7 be given data. Then to calculate the z - scores for this data we must calculate the mean and the standard deviation of data. We find $\overline{x} = 4$ and $S = 2$. So we get:

$$z_2 = \frac{x_1 - \overline{x}}{S} = \frac{2-4}{2} = \frac{-2}{2} = -1 \quad \& \quad z_5 = \frac{x_2 - \overline{x}}{S} = \frac{5-4}{2} = \frac{1}{2}$$

$$z_3 = \frac{x_3 - \overline{x}}{S} = \frac{3-4}{2} = \frac{-1}{2} \quad \& \quad z_7 = \frac{x_5 - \overline{x}}{S} = \frac{7-4}{2} = \frac{3}{2}$$

Thus we find that the standardized values $-1, \frac{1}{2}, \frac{-1}{2}, \frac{-1}{2}, \frac{3}{2}$ have mean (we denote it by $\overline{z}$):

$$\overline{z} = \frac{-1 + 0.5 - 0.5 - 0.5 + 1.5}{5} = \frac{0}{5} = 0$$

And variance (we denote it by S_z^2):

$$\begin{aligned} S_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \overline{z}\right)^2 \\ &= \frac{(-1-0)^2 + (0.5-0)^2 + (-0.5-0)^2 + (-0.5-0)^2 + (1.5-0)^2}{4} \\ &= \frac{4}{4} = 1 \end{aligned}$$

Therefore, we get:

$$S_z = +\sqrt{S_z^2} = +\sqrt{1} = 1$$

EXAMPLE 1.5.8: Let 2, 5, 7, 7, 8, 9, 9, 9, 7 and 0, 0, 1, 5, 1, 5, 7, 9, 3, 4 be degrees of students in two classes A and B respectively. Now let's see which of the students having 5, are the best in terms of level.

For this, we calculate the z-score for both students. To calculate the z - scores we must calculate the mean and the standard deviation for the two sets of data.

We find $\overline{x}_A = 7$, $S_A = 2.29$, $\overline{x}_B = 3.5$ and $S_B = 3.06$. So we get:

$$\begin{aligned} z_{5,A} &= \frac{x - \overline{x}_A}{S_A} \\ &= \frac{5-7}{2.29} = -0.873 \end{aligned}$$

$$\begin{aligned} z_{5,B} &= \frac{x - \overline{x}_B}{S_B} \\ &= \frac{5-3.5}{3.06} = 0.4902 \end{aligned}$$

Therefore, we find $z_{5,A} < z_{5,B}$. This means that the student who has 5 in class B has a best level than the student who has 5 in class A.

THE EMPIRICAL RULE

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

1. Approximately 68.2% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\overline{x} \pm S$.

2. Approximately 95.4% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 2S$.

3. Approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\overline{x} \pm 3S$.

The following graph illustrates the concept of the empirical rule.

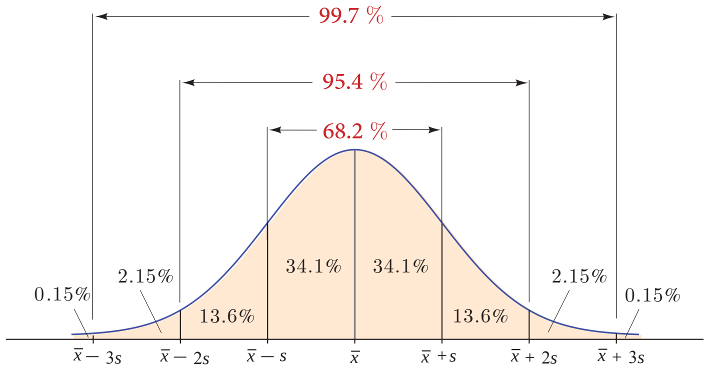


Figure 1.5.2 (Illustration of the concept of the empirical rule)

▶ **EXAMPLE 1.5.9:** Scores of some tests (as IQ test) 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 68.2% of the IQ scores in the population lie between 90 and 110,
2. Approximately 95.4% 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 1.5.10:** We will assume that apartment's prices in a certain Saudi city have a bell-shaped with mean 500000 (in S.R.) and standard deviation 100000. We will determine the price range for which at least 95.4% of the houses will sell.

The empirical rule states that 95.4% of the data values will fall within 2 standard deviations of the mean. Thus,

$$\begin{aligned}\left(\bar{x} - 2S, \bar{x} + 2S\right) &= \left(500000 - 2 \times 100000,\ \ 500000 + 2 \times 1000000\right)\\ &= \left(300000,\ \ 700000\right)\end{aligned}$$

Hence, at least 95.4% of all apartments sold will have a price range from SR 300000 to SR 700000.

◀

EXERCISES

1. The value of π till 50 decimal places is given below:

$$3.14159265358979323846264338327950288419716939937510$$

   a. Make a frequency table of the digits from 0 to 9 after the decimal point.

   b. What are the most and the least frequently occurring digits?

2. Consider the data shown in the following frequency table:

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

   a. Represent them graphically using pie chart and bar chart.

   b. Compute the range of this data.

   c. Compute the mean of this data.

3. The following data give the results of a sample survey. The letters A, B and C represent 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

   a. Prepare a frequency table of this data.

   b. Calculate the relative frequencies and percentages for all symbols.

   c. What percentage of the elements belongs to category B?

   d. Draw a bar chart and the pie chart for the frequency table.

4. 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	2.05	4.1	3.5	4.4	4.45	3.8
3.5	2.3	3.2	3.4	3.8	3.2	4.05	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 boundary intervals of size 0.5 starting from the class boundary $2 \rightarrow 2.5$ (the measure unit is 0.1).

5. 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 5 classes for the data given above?

b. Draw the histogram for the data of above frequency distribution table.

c. Draw the polygon for the data of frequency distribution table.

d. Draw the less than and greater than 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?

6. Forty 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.

b. Calculate mean, median and mode for the given data and the frequency distribution table. What do you note?

7. A sample of 100 children was asked how many times they play computer games for a period of one week. The following frequency table gives their answers:

Times of playing (in hours)	Number of children
Less than 2	23
$2 \rightarrow 5$	40
$5 \rightarrow 10$	28
$10 \rightarrow 18$	6
More than 18	3

a. Prepare the relative frequency and percentage columns.

b. What percentage of these children plays 10 hours or more weekly?

c. Draw the bar chart of the given data.

8. Consider the following frequency distribution table, representing the weights of 50 students of a class:

Weights (in kg)	Number of students	Relative Frequency	Percentage	A.C.F
$40 \rightarrow 50$	6			
$50 \rightarrow 60$		0.20		
$60 \rightarrow 70$			36%	
$70 \rightarrow 80$				44
$80 \rightarrow 90$	6			
Total	**50**			

a. Complete the above given frequency distribution table.

b. Calculate mean, median and mode for the frequency distribution table.

c. Draw the less and greater than ogive (ACFP and DCFP) for the frequency distribution table.

d. How many students have weights less than 70 Kg?

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

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

a. Represent the given information by a histogram.

b. How many lamps have a life time of more than 700 hours?

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

Grade	Section A Frequency	Section B Frequency
F	3	5
D	9	19
C	17	15
B	12	10
A	9	1

Represent the grades of the students of both the sections on the same graph by a multiple bar chart. From the multiple bar chart, compare the performance of the two sections.

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

Class Limit	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Ascending Cumulative Frequency (ACF)
2 - 6			6		
7 - 11				0.24	
12 - 16					22
17 - 21				0.25	
22 - 26			8		
Total			50		

a. Complete the above frequency distribution table.

b. Calculate the variance and standard deviation of the above frequency distribution table.

12. The points scored by a team in series of matches are as follows:

17	2	7	27	15	5	14	8
10	24	48	10	8	7	18	28

a. Calculate the mean and standard deviation of the given data.

b. Calculate the standard score of the value (7) in the given data.

c. Calculate the coefficient of variation for the given data.

d. Calculate Q_1 , Q_2 , Q_3 , LF and HF for the given data.

e. Draw the box plot for the given data.

13. Consider the following frequency distribution table**:**

Class No.	Class Boundaries	Class Midpoint	Frequency	Relative Frequency	Percentage %	A.C.F
1	$2 \rightarrow 6$		4			
2				0.25		
3					20	36
4				0.12		32
5	$18 \rightarrow 22$		8			
Total			40			

a. Complete the above frequency distribution table.

b. Draw the histogram, polygon and ACFP for the above frequency distribution table.

c. Calculate the mode(s) for the above frequency distribution table.

14. Consider the marks obtained (out of 100 marks) by 50 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

a. Calculate P_{10}, P_{50} and P_{93}.

b. Calculate D_3, D_5 and D_8.

c. Calculate Q_1, Q_2 and Q_3.

d. Check if the given data have extreme value or not? Draw the boxplot of them.

15. 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

a. Calculate the average daily sale of sugar.

b. Calculate the variance and the standard deviation of the above data.

c. Determine the coefficient of variation.

16. 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

a. Calculate Q_1, Q_2 and Q_3.

b. Calculate the IQR.

c. Have the given data got extreme values.

d. Construct the box plot for the given data.

17. Consider the following data:

65 70 70 10 20 40 60 65

85 90 90 150 75 75 80 75

a. Calculate Q_1, Q_2 and Q_3.

b. Calculate the IQR.

c. Have the given data got extreme values?

d. Construct the box plot for the given data.

e. Comment the skewness for the given data.

18. Consider the following ordered data:

40 45 55 65 ? ? 75 75 78 183

Then use the suitable measure to calculate the central tendency and dispersion for the given data.

19. Consider the following ordered data:

-15 20 40 50 65 65 70 73 75 137

a. Have the given data got extreme values?

b. Use the suitable measure to calculate the central tendency and dispersion for the given data.

20. 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. What is the age of the sixth person?

21. 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 .

22. Consider the following two data sets (we will assume that they are marks of students in a 60-degree test).

Data set ***X***:	8	12	25	37	40
Data set ***Y***:	23	27	40	52	55

Notice that each value of the second data set is obtained by adding 15 to the corresponding value of the first data set. Then:

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 (40) in data sets X and Y .

d. Calculate the coefficient of variation for each of these two data sets, and then compare them.

23. Consider the following three data sets.

Data set X:	5	10	15	20	25
Data set Y:	10	20	30	40	50
Data set Z:	20	40	60	80	100

a. Calculate the mean for each of these data sets. Comment on the relationship between the three means.

b. Calculate the standard deviation for each of these data sets. Comment on the relationship between the three standard deviations.

c. Calculate the standard score of the value (20) in data set Y.

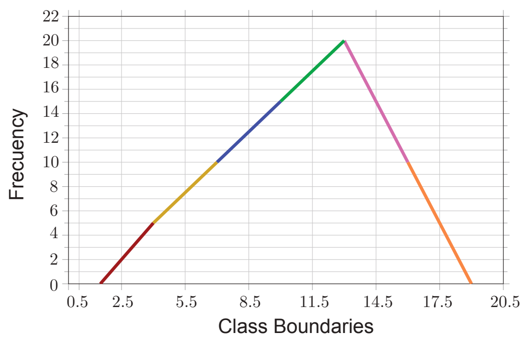
d. Calculate the coefficient of variation for each of these data sets, and then compare them.

24. Consider the 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?

25. Consider the following polygon of grouped data, representing the degrees of an examination of 60 students:

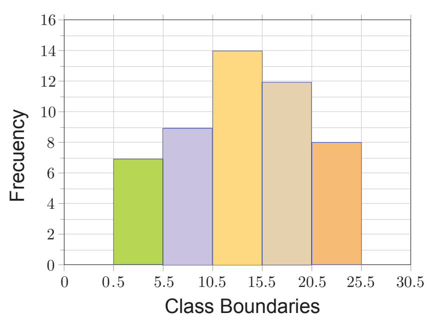


a. Prepare the frequency distribution table of this data.

b. Draw the histogram, ascending cumulative frequency polygon (less than ogive) and descending cumulative frequency polygon (greater than ogive) for this table.

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

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

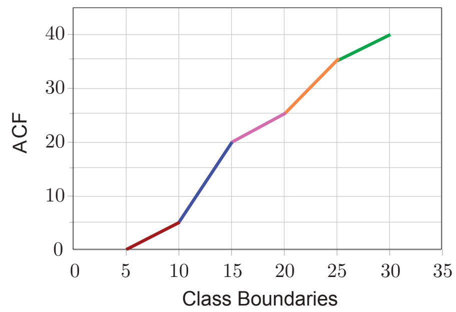


a. Prepare the frequency distribution table of this data.

b. Draw the polygon and ascending cumulative frequency polygon (less than ogive) for this table.

c. Calculate the standard deviation for this table.

27. Consider the following ascending cumulative frequency polygon (less than 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 the given data.

CHAPTER 2

PROBABILITY

LEARNING OBJECTIVES

After completing this chapter, you should be able to:

1. Define and explain the terms space of elementary events, event, mutually exclusive, and Venn diagram.
2. Counting techniques: multiplicative rule, addition rule, permutation and combination.
3. Different approaches to assigning probabilities.
4. Explain what is meant by marginal and conditional probability.
5. Concept of independent events.
6. Explain the total probability and Bayes' formula.

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 PROBABILTY CALCULUS
- SECTION 2.3 CONCEPT OF PROBABILITY FUNCTION
- SECTION 2.4 CONDITIONAL PROBABILTY AND INDEPENDENCE OF EVENT

EXERCISES

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 phenomena O_1, O_2,...and O_k. So that these phenomena occur in n_1, n_2, ... and n_k ways respectively. Then the number of ways that these phenomena occur in one time is $n_1 \times n_2 \times \ldots \times n_k$ ways.

Addition Rule

If we have k phenomena O_1, O_2,...and O_k. So that these phenomena occur in n_1, n_2, ... and n_k ways respectively. Then the number of ways that occur phenomenon O_1 or O_2 or ... or O_k equal to $n_1 + n_2 + \ldots + n_k$ 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?

Solution: 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 $\{0,1,2,3,4,5,6,\ldots\}$, then n factorial, written as $(n!)$, and is defined by the following relations:

we put $0! = 1,\ 1! = 1$ and for $n \geq 2$ we have:

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

If the number n great enough $(n \geq 10)$, then one can use the following relation to calculate $n!$:

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

DEFINITION 2.1.2 (Permutations)

Any an arrangement of r distinct objects, from a set of $n \in \mathbb{N}$ different objects called permutation.

We use the notation $n\,\mathrm{Pr}$ 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:

$$\boxed{n\,\mathrm{Pr} = \frac{n!}{(n-r)!} \quad ;\ 0 \leq r \leq n}$$

Here we note that $n\,\mathrm{P}\,n = n!$

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

Solution: The permutation formula gives:

$$_8\mathrm{P}_3 = \frac{8!}{(8-3)!} = \frac{8!}{5!} = \frac{8 \times 7 \times 6 \times (5!)}{5!} = 8 \times 7 \times 6 = 336.$$

◀

▶ **EXAMPLE 2.1.3** From a class of 10 students, a group of 3 will be chosen one after the other 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:

$$_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10\times 9\times 8\times 7!}{7!} = 720$$

◀

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 chosen r".

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

$$nCr = \frac{n!}{r!\ (n-r)!} \qquad ;\ 0 \le r \le n$$

REMARK 2.1.1

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

We note that $nCr = \frac{n\,\mathrm{Pr}}{r!}$; $\quad 0 \le r \le n$

▶ **EXAMPLE 2.1.4** How many different unordered groups of any three of the six letters l, m, n, o, p and q ?

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 2.1.5** From a class of 10 students, a group of 3 will be chosen at the same time 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 Ω be a given set, then $|\Omega|$ denote the number of all elements in Ω, and this number is called the cardinal number of Ω.

REMARKS 2.1.2

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

2. If the set Ω is infinite, and we cannot number its elements by the number of natural numbers $\mathbb{N}$, then we say that the set Ω is an uncountable set, and we write $|\Omega| = \wp$, we say that the set Ω has continuous capacity.

Section 2.2

DEFINITIONS AND CONCEPTS IN PROBABILTY CALCULUS

TYPES OF EXPERIMENTS

Experiments performed by a person are usually classified in 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 give us regular experiments:

a. The reaction of hydrochloric acid with magnesium pieces.

b. Analysis of pure water with electrical currents.

The following situations are random experiments:

a. Tossing a coin has two possibilities, head (H) or Tail (T). We cannot decide whether any results will appear in advance, but we can determine the set which containing all results of this experiment. This set is $\Omega = \{H, T\}$.

b. Rolling a die and observe the number appears on top, we will have six possibilities that are $1, 2, 3, 4, 5, 6$. We cannot decide whether any results will appear in advance, but we can determine the set which containing all results of this experiment. This set is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

c. Rolling two different dice (at the same time) and observe the number appears on top. We will have thirty six possibilities that are $(1,1), (1,2), \ldots$ and $(6,6)$. We cannot decide whether any results will appear in advance, but we can determine the set which containing all results of this experiment. This set is $\Omega = \{(i,j) \mid i,j = 1,2,3,4,5,6\}$.

d. A football team plays two games and in each game either wins (W) or be draw (D) or losses (L). In this case we will have three possibilities are $W, D \; and \; L$. We cannot decide whether any results will appear in advance, but we can determine the set which containing all results of this experiment. This set is $\Omega = \{W, D, L\}$.

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 $[\Omega, \mathscr{A}, P]$, where:

Ω is the set of all possible results of the random experiment.

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

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

Through this chapter we will quickly and succinctly explain 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 set of all possible results of this random experiment is called the space of elementary events, and denoted by Ω.

▶ **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 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 experiments.

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 get that $|\Omega| = 2$ outcomes.

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

For c) we have $\Omega = \{(1,1),(1,2),\dots,(1,6),(2,1),\dots,(2,6),\dots,(6,6)\}$, so we get that $|\Omega| = 36$ 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 get that $|\Omega| = 9$ outcomes.

◀

▶ **EXAMPLE 2.2.4** We tossing a coin three times and we observe the sequence of heads (H) and tails (T) that appear. Then the space of elementary events for this random experiment is:

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

Now when we consider 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 head is the first toss.

Then we find that:

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

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

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

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

◀

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 Ω 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 Ω consists uncountable number of outcomes, then it is called a continuous space.

EXAMPLES 2.2.5

a. Rolling a die and observe the number appears on top, Then we have the space of elementary events $\Omega = \{1,2,3,4,5,6\}$ is finite. Therefore, the space Ω is discrete.

b. Someone casts a stone in a circular well with a radius of R, Then we have the space of elementary events $\Omega = \{(x,y) \mid x^2 + y^2 < R^2\}$ consists uncountable number of outcomes. Therefore, the space Ω is continuous.

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 Ω. Then a collection $\mathscr{A}$ of subset of Ω is said to be an algebra on Ω if and only if the following condition are verified:

1. $\Omega \in \mathscr{A}$.
2. For any two elements A and $B \in \mathscr{A}$. Then $A \bigcup B \in \mathscr{A}$.
3. For any element $A \in \mathscr{A}$. Then $\overline{A} \in \mathscr{A}$.

REMARKS 2.2.2

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

   For any sequence $A_1, A_2, \ldots, A_n, \ldots \in \mathscr{A}$ we have $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$.

   Then $\mathscr{A}$ is said to be an σ −algebra.

2. The elements of $\mathscr{A}$ (as an algebra or an σ −algebra) are called events.
3. In the triple $[\Omega, \mathscr{A}, P]$ must $\mathscr{A}$ to be an σ −algebra.
4. When Ω is finite or countable infinite (Ω is a discrete space), then any subset of Ω is an event. But this statement is not true if Ω 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 Ω a finite, and in this case one can prove that any algebra on Ω is an σ −algebra on Ω also.

6. When we talk about events, we will always assume that we have a given random experiment.

In the general case, an event consists one or more elementary event, and these elementary events have a common characteristic. Based on the number of elementary events, which the events contain, we can classify two types of events.

DEFINITION 2.2.6 (Simple Event and Compound Event)

An event A is said to be a simple event if contains only one elementary event (outcome).

An event A is said to be a compound event if contains at least two elementary events (outcomes).

▶ **EXAMPLE 2.2.6** Rolling a single die once, so the space of elementary events is:

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

We will determine the events, which contain a certain number.

Events, which contain a certain number are $\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}$, and these events are simple events.

We will determine the events $\{x,y\}$ with $x,y \in \Omega$ and the property $y = x + 2$.

For that, we have the events $\{1,3\},\{2,4\},\{3,5\},\{4,6\}$, and these events are compound events.

◀

▶ **EXAMPLE 2.2.7** Tossing a coin twice, so the space of elementary events is:

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

Then the events $\{HH\}$, $\{TH\}$, $\{HT\}$, $\{TT\}$ are simple events, while the events $\{HH, HT\}$, $\{HH, TH\}$, $\{TH, TT\}$, $\{HH, HT, TT\}$ are compound events.

◀

▶ **EXAMPLE 2.2.8** A box contains a few red and a few green cards. If two cards are randomly drawn (one after the other) and the colors of these cards are observed, how many total outcomes are possible?

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

E_1: Both cards have different colors.

E_2: At least one card is red.

E_3: At most one card is red.

E_4: The first card 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 card is selected, and G is the event that a green card is selected. Then we find:

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

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

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

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

REMARK 2.2.3

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

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

2. If Ω is a space of elementary events of a random experiment, and Ω is finite or countable infinite, then 2^{Ω} is a σ - algebra on Ω, therefore the elements of 2^{Ω} are events. But if Ω is uncountable, as $\Omega = \mathbb{R}$ (or Ω is an interval $I \subset \mathbb{R}$), then can be exist elements of 2^{Ω} which not events. In this case one denote the σ - algebra on Ω by $\Re$ (or $I \bigcap \Re$, where $\Re$ is called Boreal field).

3. In the following study we will take Ω a finite or countable (for simplicity), therefore the elements of 2^{Ω} are events. Therefore (to avoid repetition), when we write A, B and … are events, we mean that we have a random experiment with space of elementary events Ω and σ −algebra $\mathscr{A} = 2^{\Omega}$, and the events A, B and … are from $\mathscr{A}$.

SOME OPERATIONS ON EVENTS

Since a space of elementary events Ω of a randomized experiment is a set, then from the logical operations on sets, we can create new events related to the random experiment.

For some of these operations, one applies some different terminology from set theory in probability theory.

Union of Two Events (This Expression is Metaphorical)

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

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

Where ω refers to any element belong to A or B or to both.

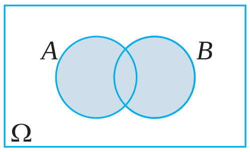


Figure 2.2.1 (Shaded region represents the event $A \cup B$)

EXAMPLE 2.2.9 Let $\Omega = \left\{1,2,3,4,5,6\right\}$ be the space of elementary events of rolling a die, and we consider $A = \left\{1,3\right\}$ and $B = \left\{3,5\right\}$, then we have $A \bigcup B =$ {1,3,5} is the event that we get an odd number.

EXAMPLE 2.2.10 In the random experiment of tossing a coin twice, let A be an event that we get a head in the first toss, and B that we get a head in the second toss. Then we have $\Omega = \left\{HH, HT, TH, TT\right\}$ is the space of elementary events, and $A = \left\{HH, HT\right\}$, $B = \left\{HH, TH\right\}$. Therefore, we have $A \cup B = \{HH, HT, TH\}$ is the event that we get at least one head.

Intersection of Two Events (This Expression is Metaphorical)

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

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

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

Both events A and B are occurring.

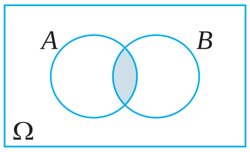


Figure 2.2.2 (Shaded region represents the event $A \cap B$)

▶ **EXAMPLE 2.2.11** 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 a tail in the second toss, we have:

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

Then we have $A \cap B = \{TT\}$ is the event that we get exactly two tails in this experiment.

◀

Complement of an Event

Let A be an event of $\mathscr{A}$. Then the complement of A with respect to Ω is an event that occurs if A does not occur, and is denoted by $\overline{A}$. So we have:

$$\overline{A} = \{\omega : \omega \in \Omega, \omega \notin A\}$$

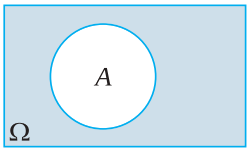


Figure 2.2.3 (Shaded region represents the event $\overline{A}$)

For every event A their corresponds the event $\overline{A}$ consist of all elementary events of Ω which are not in A.

▶ **EXAMPLE 2.2.12** We rolling a die at one time, and we observe the number that appears on top. We will determine the following events:

a. The event that an even or a prime number occurs.

b. The event that a prime number does not occur.

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

A is the event that we get an even number, then we have $A = \{2, 4, 6\}$,

B is the event that we get a prime number, then we have $B = \{2, 3, 5\}$.

Therefore, we have:

For a) the event that an even or a prime number occurs is $A \cup B = \{2, 3, 4, 5, 6\}$,

For b) the event that a prime number does not occur is $\overline{B} = \{1, 4, 6\}$.

◀

Difference Between Two Events (This Expression is Metaphorical)

If A and B are two events of $\mathscr{A}$, 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 occurs or only A must be occurs.

$$A \setminus B = A \cap \bar{B} = \left\{\omega : \omega \in A \;\; and \;\; \omega \notin B\right\}$$

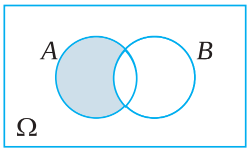


Figure 2.2.4 (Shaded region represents the event $A \setminus B$)

Exactly One of Event

If A and B are two events of $\mathscr{A}$, then:

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

means A or B occurs, but not both together:

$$A \Delta B = \{x : x \in A \bigcap \bar{B} \; or \, x \in B \bigcap \bar{A}\} = \{x : x \in A \bigcup B \; and \, x \notin A \bigcap B\}$$

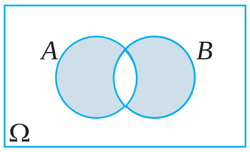


Figure 2.2.5 (Shaded region represents the event $A \Delta B$)

DEFINITION 2.2.7 (Impossible Event)

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

EXAMPLES 2.2.13

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 $A \in \mathscr{A}$ we know that it is surely an outcome of the experiment belongs to $A \bigcup \bar{A}$. Therefore, the event $A \bigcup \bar{A}$ is called a certain event, and since in set theory $A \bigcup \bar{A} = \Omega$, so one denotes the certain event by Ω also.

EXAMPLES 2.2.14

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^{Ω} are:

$$\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\} \quad \Rightarrow \left|2^{\Omega}\right| = 2^3 = 8$$

We note that $\varnothing$ is the impossible event, Ω 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 Events)

Two events A and B of $\mathscr{A}$ are mutually exclusive if they cannot occur at the same time.

That means two events A and B of $\mathscr{A}$ are called mutually exclusive events if $A \cap B = \varnothing$. That is A and B have no elementary events in common.

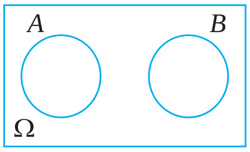


Figure 2.2.6 (A and B are mutually exclusive events)

We remark here that for any event A of $\mathscr{A}$, the events A and $\varnothing$ are mutually exclusive events.

DEFINITION 2.2.10 (Pair Wise Mutually Exclusive Event)

Events $A_1, A_2, A_3, \ldots$ of $\mathscr{A}$ are said to be pair wise mutually exclusive if:

$$A_i \cap A_j = \varnothing \qquad ; \forall\, i \neq j$$

EXAMPLE 2.2.15 One throws a single die and observes the number that appears on top. Let us determine the following events:

A: appearance of an odd number,

B: appearance of an even number,

C: appearance of a prime number,

$A \cap B$, $A \cap C$, $B \cap C$ and $A \cap B \cap C$.

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

For the event A we have $A = \{1, 3, 5\}$,

For the event B we have $B = \{2, 4, 6\}$,

For the event C we have $C = \{2, 3, 5\}$,

For the event $A \cap B = \varnothing$,

For the event $A \cap C = \{3, 5\}$,

For the event $B \cap C = \{2\}$ and

For the event $A \cap B \cap C = \varnothing$.

◀

Section 2.3

CONCEPT OF PROBABILITY FUNCTION

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

DEFINITION 2.3.1 (Probability)
Probability is a 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 conditions 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.

RELATIVE FREQUENCY OF AN EVENT AS AN APPROXIMATION OF PROBABILITY

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 $f_A = \dfrac{n(A)}{N}$ represents the relative frequency of occurrence of A on these trials.

We note here that the value of f_A satisfies the following:

a. $0 \leq f_A \leq 1$ with $f_\varnothing = 0$ and $f_\Omega = 1$

b. For any two mutually exclusive events A and B dependent of this experiment is:

$$f_{A \cup B} = f_A + f_B$$

On the above we can define a real function on the algebra of events $\mathscr{A}$ as follow:

$$P : \mathscr{A} \longrightarrow \mathbb{R} \ ; \ A \mapsto P(A) = f_A$$

Then we have:

a. $P(\varnothing) = 0$

b. For any two mutually exclusive events A and $B \in \mathscr{A}$ is $P(A \cup B) = P(A) + P(B)$

c. $P(\Omega) = 1$

This function gives us approximate value of the probability of A. That means, 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** It has been observed, that 10 of the 500 randomly selected cars in a certain auto factory are defective cars. What is the approximately probability that the next manufactured car at this factory is a defective car?

Solution: Assume that the event under study is B, then we have:

$$N = 500 \quad n(B) = 10$$

Therefore, we can write:

$$\begin{aligned} P\big(\text{next car is a defective}\big) &= P(B) \\ &= \frac{10}{500} = 0.02 \end{aligned}$$

◀

▶ **EXAMPLE 2.3.2** In an 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 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:

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

Note that there are examples (such as the Buffon -Pearson experiment) that prove that this limit may not exist. Therefore, the value of this limit cannot be accepted as a value for the probability of the event A.

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 (Equal choice). 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 Ω, 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).

For a simple event S :	$P(S) = \dfrac{1}{\mid \Omega \mid}$
For a compound event A:	$P(A) = \dfrac{\mid A \mid}{\mid \Omega \mid}$

EXAMPLE 2.3.3 Tossing a fair coin one time. Then:

a. Determine the space of elementary events of this experiment.

b. Calculate the probability of obtaining a head.

c. Calculate the probability of obtaining a tail.

Solution:

For a) We have $\Omega = \{H, T\}$.

For b and c) Because the outcomes are equal likely outcomes, then we have:

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

EXAMPLE 2.3.4 Rolling a fair die one time. Then calculate the probability of obtaining an even number.

Solution:

We have the space of elementary events of this experiment:

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

We suppose that A the event of obtaining an even number. So because the outcomes are equal likely outcomes, we have:

$$P(A) = \frac{\mid A \mid}{\mid \Omega \mid} = \frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2}$$

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 Ω be a space of elementary events of a random experiment, and $\mathscr{A} \subseteq 2^{\Omega}$ is a σ-algebra on Ω. Furthermore, we suppose that P a real set function on $\mathscr{A}$ with the following properties:

a. We have $P(\varnothing) = 0$.

b. For any sequence $A_1, A_2, \dots, A_n, \dots \in 2^{\Omega}$ with $A_i \bigcap_{i \neq j} A_j = \varnothing$, then:

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

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

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

REMARK 2.3.2

If Ω is a space of elementary events of a random experiment, and $\mathscr{A} \subseteq 2^{\Omega}$ is a $\sigma -$ algebra on Ω and P is a probability measure on $\mathscr{A}$. Then the triple $[\Omega, \mathscr{A}, P]$ is called the probability space of the random experiment (or the theoretical mathematical model of the random experiment).

THEOREM 2.3.1

For a random experiment with probability space $[\Omega, \mathscr{A}, P]$, then:

a. For any event A of $\mathscr{A}$ we have $P(\bar{A}) = 1 - P(A)$.

b. For any two events A and B of $\mathscr{A}$ with $A \subset B$, we get $P(B \setminus A) = P(B) - P(A)$.

c. For any two events A and B of $\mathscr{A}$ with $A \subset B$, then $P(A) \leq P(B)$ (monotonicity property of the measure P).

d. For any event A of $\mathscr{A}$ follows from the monotonicity property that:

$$P(\varnothing) = 0 \leq P(A) \leq 1 = P(\Omega)$$

THEOREM 2.3.2

Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment. Now, if Ω is finite (we suppose $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$), then we can calculate the probability of any event $A \in 2^{\Omega}$ by the following relation:

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

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

REMARK 2.3.3

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

a. We have $\overline{A \cup B} = \overline{A} \cap \overline{B}$ for any two sets A and B.

b. We have $\overline{A \cap B} = \overline{A} \cup \overline{B}$ for any two sets A and B.

ADDITIVE LAW IN THE PROBABILITY CALCULATION

Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment, one can prove the following statements.

a. If A and B are any two events of $\mathscr{A}$, then:

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

We note that, if A and B are mutually exclusive events of $\mathscr{A}$, then we get:

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

b. If A, B and C are any three events of $\mathscr{A}$, then:

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

▶ **EXAMPLE 2.3.5** Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment and we consider A and B be events of $\mathscr{A}$ with $P(A) = \frac{3}{8}$, $P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{4}$. Then calculate:

a. $P(A \cup B)$	**b.** $P\left(\overline{A}\right), P\left(\overline{B}\right)$	**c.** $P\left(\overline{A} \cap \overline{B}\right)$
d. $P\left(\overline{A} \cup \overline{B}\right)$	**e.** $P\left(A \cap \overline{B}\right)$	**f.** $P\left(\overline{A} \cap B\right)$

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) $P\left(\overline{A}\right) = 1 - P(A) = 1 - \frac{3}{8} = \frac{5}{8}$, $P\left(\overline{B}\right) = 1 - P(B) = 1 - \frac{1}{2} = \frac{1}{2}$.

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

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

For e) $P\left(A \cap \overline{B}\right) = P(A) - P(A \cap B) = \frac{3}{8} - \frac{1}{4} = \frac{1}{8}$

For f) $P\left(\overline{A} \cap B\right) = P(B) - P(A \cap B) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$.

◀

▶ **EXAMPLE 2.3.6** Three students Ahmad, Bandar and Conan are in a swimming race. Ahmad and Bandar have the same probability of winning and each is twice as likely to win as Conan. Now, if only one person wins, then calculate the probability that Bandar or Conan wins.

Solution: We assume that

A is the event that Ahmed is winning,

B is the event that Bandar is winning,

C is the event that Conan is winning.

Now if we put $P(C) = x$, then we have $P(A) = P(B) = 2x$, and 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 Bandar or Conan wins is:

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

◀

▶ **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:

a. An even number appears.

b. An odd number appears.

c. A prime number appears.

d. An odd number occurs but not prime number.

Solution: We have:

$$\Omega = \{\underset{\omega_1}{1}, \underset{\omega_2}{2}, \underset{\omega_3}{3}, \underset{\omega_4}{4}, \underset{\omega_5}{5}, \underset{\omega_6}{6}\}$$

Now we assume that:

A is the event that we get an even number,

B is the event that we get an odd number,

C is the event that get a prime number.

Then we have $A = \{2,4,6\}$, $B = \{1,3,5\}$ and $C = \{2,3,5\}$, and:

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

If we put:

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

We get:

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

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

$$P\left(\{1\}\right) + P\left(\{3\}\right) + P\left(\{5\}\right) + 2\ P\left(\{1\}\right) + 2P\left(\{3\}\right) + 2P\left(\{5\}\right) = 1$$

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

and hence:

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

Therefore, we have:

For a) $P(A) = \sum\limits_{i\,;\omega_i \in A} P\left(\{\omega_i\}\right) = \mathrm{P}(\{\omega_2\}) + \mathrm{P}(\{\omega_4\}) + \mathrm{P}(\{\omega_6\}) = \dfrac{6}{9}$

For b) $P(B) = \sum\limits_{i\,;\omega_i \in B} P\left(\{\omega_i\}\right) = \mathrm{P}(\{\omega_1\}) + \mathrm{P}(\{\omega_3\}) + \mathrm{P}(\{\omega_5\}) = \dfrac{3}{9}$

For c) $P(C) = \sum\limits_{i\,;\omega_i \in A} P\left(\{\omega_i\}\right) = \mathrm{P}(\{\omega_2\}) + \mathrm{P}(\{\omega_3\}) + \mathrm{P}(\{\omega_5\}) = \dfrac{4}{9}$

For d) $P\left(B \cap \overline{C}\right) = P\ (B) - P(B \cap C)\ = \dfrac{3}{9} - \dfrac{2}{9} = \dfrac{1}{9}$

◀

▶ **EXAMPLE 2.3.8** Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment. If A and B are two events of $\mathscr{A}$, and we suppose that $P\left(A\right) = 0.8,\ P\left(B\right) = 0.55$ and $P(A \cup B) = 0.9$. Then calculate the probability of:

- **a.** Occurrence of A and B,
- **b.** Occurrence of only A and not B,
- **c.** None occurrence of A and none occurrence of B,
- **d.** None occurrence of A and B,
- **e.** Occurrence only A or only B.

Solution: We have:

For a) $P\left(A \cap B\right) = P(A) + P(B) - P(A \cup B) = 0.8 + 0.55 - 0.90 = 0.45$

For b) $P\left(A \cap \overline{B}\right) = P\ (A) - P(A \cap B)\ = 0.8 - 0.45 = 0.35$

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

For d) $P\left(\overline{A \cap B}\right) = 1 - P(A \cap B) = 1 - 0.45 = 0.55$

For e) $P\left(\left(A \cap \overline{B}\right) \cup \left(\overline{A} \cap B\right)\right) = P\left(A \cap \overline{B}\right) + P\left(\overline{A} \cap B\right) = 0.35 + 0.10\ = 0.45$

◀

EXAMPLE 2.3.9 Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment. If A, B and C are three events of $\mathscr{A}$ such that $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{8}$ $P(C) = \frac{1}{4}$, where A, B and C are mutually exclusive. Then calculate:

a. $P(A \cup B \cup C)$ **b.** $P\left(\overline{A} \cap \overline{B} \cap \overline{C}\right)$

Solution: We have:

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

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

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 $P(A \cup B) = 0.5$. Then calculate the probability that:

a. both sets are on,

b. the color set is on and the other is off,

c. exactly one set is on,

d. Neither set is on.

Solution: We have:

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

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

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

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

EXAMPLE 2.3.11 Let $[\Omega, \mathscr{A}, P]$ is a probability space of a random experiment. If $P(A) = 0.4$ and $P(A \cup B) = 0.6$. Then calculate the value of $P(B)$ which makes A and B mutually exclusive events.

Solution: We have:

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

According to the requirements of the problem, we have:

$$0.6 - 0.4 - P(B) = P(A \cap B)$$

Since A and B are mutually exclusive, we have $P(A \cap B) = 0$.

So we get that:

$$P(B) = 0.6 - 0.4 = 0.2$$

Section 2.4

CONDITIONAL PROBABILTY AND INDEPENDENCE OF EVENT

Suppose that we have a random experiment with a probability space $[\Omega, \mathscr{A}, P]$. Moreover, let $B \in \mathscr{A}$ be an arbitrary event with $P(B) > 0$. Then the probability that an event A occurs given the event B occurs, written as $P\left(A \mid B\right)$, is defined as:

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

REMARKS 2.4.1

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

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

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

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

This result can be extended as follows:

If $A_1, A_2, \ldots, A_n, \ldots$ are a sequence of pair wise mutually exclusive events. Then we have:

$$\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B = \bigcup_{i=1}^{\infty} \left(A_i \cap B\right)$$

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

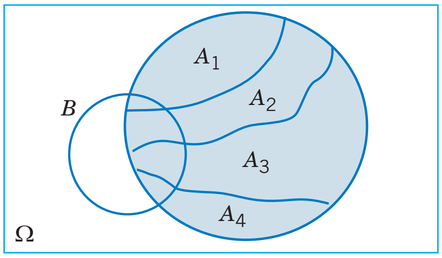


Figure 2.4.1

Note that the sequence $A_1 \cap B$, $A_2 \cap B, \ldots, A_n \cap B, \ldots$ is also sequence of pair wise mutually exclusive events. Therefore, we get the following results:

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

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

4. If two events B and $C \in \mathscr{A}$ with $B \subseteq C$, then we have $B \cap C = B$, therefore:

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

In special case $C = \Omega$ we get that:

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

We find from this last relation that:

$$P(A \cup \bar{A} \mid B) = P(A \mid B) + P(\bar{A} \mid B) = 1 \Rightarrow P(\bar{A} \mid B) = 1 - P(A \mid B)$$

5. According to the definition of conditional probability, if $B \subset A$ then, we get that:

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

6. For any $B \in \mathscr{A}$ with $P(B) > 0$ one can proves that the conditional probability measure $P(\bullet \mid B)$ has all usual properties of probability measure P, but on the reduced space of elementary events B.

▶ **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 is left-handed?

Solution: We suppose that:

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

B is the event, that a randomly selected student is a sophomore.

Then we have:

$$P\left(A\right) = 0.35 \ , \ P\left(B\right) = 0.50 \ , \ P(A \cap B) = 0.05$$

What we want is $P\left(A \mid B\right)$.

Now by the definition we can write:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.50} = 0.1$$

◀

▶ **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, and suppose that:

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

Then if the system has a defect of type 1, what is the probability that it has a defect of type 2 also?

Solution: What we want to calculate is:

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

◀

MULTIPLICATION LAW IN THE PROBABILITY CALCULATION

Suppose that we have a random experiment with a probability space $[\Omega, \mathscr{A}, P]$ and $A, B \in \mathscr{A}$ with $P(B) > 0$. Then by using the formula of conditional probability:

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

We become the following relation:

$$P\left(A \cap B\right) = P\left(B\right)P\left(A \mid B\right)$$

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

$$P\left(A \cap B\right) = P\left(A\right)P\left(B \mid A\right)$$

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 is as follow:

If we have A_1, A_2, ...and $A_n \in \mathscr{A}$ with $P\left(A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right) > 0$. Then we can write:

$$\begin{aligned} P\left(A_1 \cap A_2 \cap \ldots \cap A_n\right) &= P\left(A_1\right) \cdot P\left(A_2 | A_1\right) \cdot P\left(A_3 | A_1 \cap A_2\right) \cdot \ldots \\ &\ldots \cdot P\left(A_{n-1} | A_1 \cap A_2 \cap \ldots \cap A_{n-2}\right) \cdot P\left(A_n | A_1 \cap A_2 \cap \ldots \cap A_{n-1}\right) \end{aligned}$$

This relation is known as the multiplication law in probability (or chain rule in probability).

▶ **EXAMPLE 2.4.3** If a fair die is rolled twice. Now, if we know that a sum greater than 7 is occurred, and more we know that we get the same numbers to the top. Then, what is the probability that we get the pair (5,5)?

Solution: The space of elementary events of this experiment is given by:

$$\Omega=\left\{\begin{array}{l}(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\end{array}\right\} \Rightarrow |\Omega|=36$$

Now, we suppose that:

A is the event that a sum greater than 7 is occurred,

B is the event that we have same numbers to the top is occurred,

C is the event that we get $(5,5)$ to the top is occurred.

Then we can write:

$$A=\left\{\begin{array}{l}(2,6),(3,5),(3,6),(4,4),(4,5),(4,6),\\(5,3),(5,4),(5,5),(5,6),(6,2),(6,3),\\(6,4),(6,5),(6,6)\end{array}\right\},\ B=\{(4,4),(5,5),(6,6)\} \text{ and } C=\{(5,5)\}$$

What we want to calculate is:

$$P(C|A\cap B)=\frac{P(A\cap B\cap C)}{P(A\cap B)}=\frac{P(\{(5,5)\})}{P(\{(4,4),(5,5),(6,6)\})}=\frac{1}{3}$$

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 2.4.4 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:

$$P\left(A \mid B\right) = P\left(A\right)$$

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:

$$P\left(A \cap B\right) = P\left(A \mid B\right) \cdot P\left(B\right) = P\left(A\right) \cdot P\left(B\right)$$

This last equation gives us another way of looking at independent events. It says that if two events A and B are independent then the product of their individual probabilities $P(A) \cdot P(B)$ is equal to the probability of occurrence of the two events $P\left(A \cap B\right)$.

Let's look at independent events with an example.

▶ **EXAMPLE 2.4.5** 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.

a. Are the events A and B independent?

b. 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:

For a) 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:

$$P\left(A \bigcap B\right) = P\left(A\right)P\left(B\right)$$

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

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

So we have $A = \left\{HH, HT\right\}$ and $B = \left\{HH, TH\right\}$. Therefore, we get:

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

Comparing this quantity with:

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

We see that in fact A and B are independent.

For b) The probability that you would get a head on the second flip given that you had a head in the first flip equal to $P\left(B\right) = \frac{1}{2}$.

◀

▶ EXAMPLE 2.4.6 Suppose you flip a coin three times., and assume that A, B and C represents the event you get a head on the first, second and third flip respectively. Then:

a. Are events A, B and C independent?

b. What's the probability of getting a head on the third flip give that the previous two flips were heads?

Solution:

For (a): We know that the space of elementary events for flipping a coin three times is:

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

Therefore, we have:

$$P(A) = P(\{HHH, HHT, HTH, HTT\}) = \frac{4}{8} = \frac{1}{2}$$

$$P(B) = P(\{HHH, HHT, THH, THT\}) = \frac{4}{8} = \frac{1}{2}$$

$$P(C) = P(\{HHH, THH, HTH, TTH\}) = \frac{4}{8} = \frac{1}{2}$$

In addition:

$$P(A \cap B) = P(\{HHH, HHT\}) = \frac{2}{8} = \frac{1}{4}$$

$$P(A \cap C) = P(\{HHH, HTH\}) = \frac{2}{8} = \frac{1}{4}$$

$$P(B \cap C) = P(\{HHH, THH\}) = \frac{2}{8} = \frac{1}{4}$$

Thus, we find that:

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

$$P(A \cap C) = P(A) \cdot P(C)$$

$$P(B \cap C) = P(B) \cdot P(C)$$

On the other hand, we have:

$$P(A \cap B \cap C) = P(\{HHH\}) = \frac{1}{8}$$

Therefore, we find that:

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

So we get that the events A, B and C are independent.

So this shows that getting a head on the third flip is unaffected by having heads on the previous two flips.

For (b): We must calculate $P\left(C \mid \left(A \cap B\right)\right)$. For that we have:

$$
\begin{aligned}
P\left(C \mid \left(A \cap B\right)\right) &= \frac{P\left(C \cap (A \cap B)\right)}{P(A \cap B)} \\
&= \frac{P\left(\{HHH\}\right)}{P\left(\{HHH, HHT\}\right)} \\
&= \frac{1}{2}
\end{aligned}
$$

DEFINITION 2.4.3 (Statistical Independency of Events)

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment, and $A_1, A_2, \ldots, A_n \in \mathscr{A}$ are given events. Then $A_1, A_2, \ldots, A_n$ said to be statistically independent if and only if for any permutation $(i_1\ i_2\ \ldots\ i_k)$ with $k \in \{2, 3, \ldots, n\}$ of the numbers $1, 2, \ldots, n$ we have:

$$P\left(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\right) = P\left(A_{i_1}\right) \cdot P\left(A_{i_2}\right) \cdot \ldots \cdot P\left(A_{i_k}\right)$$

If the last relation satisfy for the special case $k = 2$, then one say that the events $A_1, A_2, \ldots, A_n$ are pair wise independent.

REMARKS 2.4.3

1. It is important that one be 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.

2. If A and B are two independent events, then
   The events A and $\bar{B}$ are independent,
   The events $\bar{A}$ and B are independent,
   The events $\bar{A}$ and $\bar{B}$ are independent.

For two events A and B, the following table will help to distinguish between mutually exclusive and independent events:

	Probabilities	Verbal Description
Mutually Exclusive	$P\left(A \cap B\right) = 0$	Both events cannot happen
Independent	$P\left(A \mid B\right) = P\left(A\right)$ $P\left(A \cap B\right) = P\left(A\right) \cdot P\left(B\right)$	The occurrence of B does not affect the occurrence of A

▶ **EXAMPLE 2.4.7** We have a box with identical balls, but in different colors, white and yellow. Moreover, a part of the white balls is numbered by the number 1 and the rest by the number 2, as well as for the yellow balls. The colors and numbers of these balls are detailed in the table below.

		Number on the ball		
		Have the number 1	**Have the number 2**	Total
Balls	**White**	24	8	32
	Yellow	18	10	28
	Total	42	18	60

After mixing the balls well we randomly pull a ball out of the box. Then:

a. If the ball, which pulled is white, what is the probability that it has the number 2?

b. If the ball, which pulled has the number 1, what is the probability that is yellow?

c. Are the events “Balls” and “Numbers” independent.

Solution: We suppose that:

W is the event that the ball, which pulled is white,

O is the event that the ball, which pulled has the number 1,

Y is the event that the ball, which pulled is yellow,

T is the event that the ball, which pulled has the number 2.

Now, if we use the events above, we get the next table divide by 60:

	O	T	Total
W	0.400	0.133	0.533
Y	0.300	0.167	0.467
Total	0.700	0.300	1.000

Therefore, we have:

For a) $P(T \mid W) = \dfrac{P(T \cap W)}{P(W)} = \dfrac{0.133}{0.533} = 0.2495$

For b) $P(Y \mid O) = \dfrac{P(O \cap Y)}{P(O)} = \dfrac{0.300}{0.700} = 0.4286$

For c) Since $P(T \mid W) = 0.2495 \neq 0.300 = P(T)$, then the events “Balls” and “Numbers” are not independent. ◀

▶ **EXAMPLE 2.4.8** We suppose that, if a student studied, then the probability of passing a certain quiz is 0.99. As well, we suppose that, if the student did not study, then the probability of passing the quiz is 0.05. Now if we know that the probability of studying is 0.7 and a student flunks the quiz, what is the probability that he (or she) did not study?

Solution: We suppose that:

A is the event that the student passing the quiz.

B is the event that the student studying.

Then:

$\bar{A}$ is the event that the student did not passing the quiz.

$\bar{B}$ is the event that the student did not studying.

And from the given information we have:

$$P\left(A \mid B\right)= 0.99,\ P\left(A|\bar{B}\right)=0.05,\ P\left(B\right)= 0.7 \text{ and } P\left(\bar{B}|\bar{A}\right) = ?$$

The probability to be calculated:

$$P\left(\bar{B}|\bar{A}\right) = \frac{P\left(\bar{B}\cap\bar{A}\right)}{P\left(\bar{A}\right)} = \frac{P\left(\overline{B\cup A}\right)}{P\left(\bar{A}\right)} = \frac{1-P\left(A\cup B\right)}{P\left(\bar{A}\right)}$$

But we have:

$$P(A\cup B) = P(A) + P(B)-P(A\cap B)= 0.708+0.3-0.693= 0.715$$

Where we have:

$$P(A\cap B) = P(B)\ P(A \mid B)= 0.7\times 0.99 =0.693$$

$$P(A\cap \bar{B}) = P(\bar{B})\ P(A \mid \bar{B})= 0.3\times 0.05 =0.015$$

And

$$P\left(A|\bar{B}\right) = \frac{P\left(A\cap\bar{B}\right)}{P\left(\bar{B}\right)} = \frac{P\left(A\right)-P\left(A\cap B\right)}{P\left(\bar{B}\right)}$$

So we get:

$$P\left(A\right)= P\left(\bar{B}\right)P\left(A|\bar{B}\right)+ P\left(A\cap B\right) = 0.3\times0.05 + 0.693 = 0.708$$

$$P\left(\bar{A}\right)= 1 - P\left(A\right)= 1 - 0.708 = 0.292$$

Therefore, we become:

$$P\left(\bar{B}|\bar{A}\right) = \frac{1-0.715}{1-0.708} = \frac{0.285}{0.292} = 0.976$$

◀

▶ **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:

a. If he failed chemistry, what is the probability that he failed mathematics?

b. If he failed mathematics, what is the probability that he failed chemistry?

c. What is the probability that he failed mathematics or chemistry?

Solution: We suppose that:

M is the event that the student failed in mathematics.

C is the event that the student failed in chemistry.

So from the given information we have:

$$P(M) = 0.25,\ P(C) = 0.15 \ \ and \ \ P(M \cap C) = 0.10$$

For a) We have: $P\left(M|C\right) = \dfrac{P\left(M \cap C\right)}{P\left(C\right)} = \dfrac{0.10}{0.15} = \dfrac{2}{3}.$

For b) We have: $P\left(C|M\right) = \dfrac{P\left(C \cap M\right)}{P\left(M\right)} = \dfrac{0.10}{0.25} = \dfrac{2}{5}.$

For c) We have: $P\left(M \cup C\right) = P\left(M\right) + P\left(C\right) - P\left(M \cap C\right) = 0.30$

TOTAL PROBABILITY

DEFINITION 2.4.1

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment. Then events $Z_1, Z_2, \ldots, Z_n$ of $\mathscr{A}$ are a partition of Ω, if:

a) $Z_i \neq \varnothing$ for all i

b) $Z_i \cap Z_j = \varnothing$ for all $i \neq j$

c) $\bigcup_{i=1}^{n} Z_i = \Omega$

THEOREM 2.4.1 (Total Probability Formula)

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment. If the events $Z_1, Z_2, \ldots, Z_n$ are events of $\mathscr{A}$, constitute a partition of the space of elementary events Ω such that $P\left(Z_k\right) \neq 0$ for $k = 1, 2, \ldots, n$, then for any event B we have:

$$P(B) = \sum_{k=1}^{n} P(Z_k) P(B|Z_k)$$

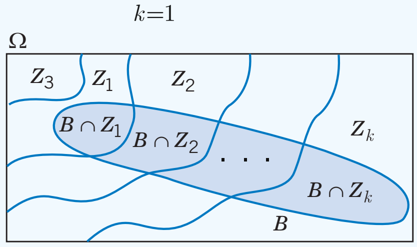


Figure 2.4.2

The above relation is known as the total probability formula.

BAYES' FORMULA

Bayes' theorem is an important element in the probability calculation. It was 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 of the form $P\left(B \mid A\right)$ and $P\left(A \mid B\right)$.

THE DERIVATION OF BAYES' FORMULA FOR TWO EVENTS

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment, and we consider A and B two events of $\mathscr{A}$ with $P(B) > 0$. Then we can write the following relation:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B \mid A)}{P(A) \cdot P(B \mid A) + P(\bar{A}) \cdot P(B \mid \bar{A})}$$

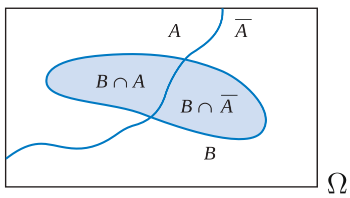


Figure 2.4.3

We want to know:

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

From the multiplication rule in probability we have:

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

In addition, in view of the above Venn diagram we notice that:

$$P(B) = P(B \cap A) + P(B \cap \bar{A}) = P(A) \cdot P(B \mid A) + P(\bar{A}) \cdot P(B \mid \bar{A})$$

Thus we get that:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B \mid A)}{P(A) \cdot P(B \mid A) + P(\bar{A}) \cdot P(B \mid \bar{A})}$$

■

THEOREM 2.4.2 (Bayes' Theorem)

Let $[\Omega, \mathscr{A}, P]$ be a probability space of a random experiment. If the events $Z_1, Z_2, \ldots, Z_n \in \mathscr{A}$ are a partition of a space of elementary events Ω such that $P\left(Z_k\right) \neq 0$ for $k = 1, 2, \ldots, n$, then for any event B such that $P\left(B\right) \neq 0$:

$$P(Z_i \mid B) = \frac{P(Z_i)\ P(B \mid Z_i)}{\sum_{k=1}^{n} P(Z_k)\, P(B \mid Z_k)} \qquad ; i = 1, 2, \ldots, n$$

PROOF

From the definition of conditional probability and the total probability formula we find for any natural number $i \le n$:

$$P(Z_i \mid B) = \frac{P(B \cap Z_i)}{P(B)} = \frac{P(Z_i) \cdot P(B \mid Z_i)}{P(B)} = \frac{P(Z_i) \cdot P(B \mid Z_i)}{\sum_{k=1}^{n} P(Z_k) \cdot P(B \mid Z_k)}$$

■

▶ **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.

a. What percentage of students is expected to pass the course?

b. Given that a student passes the course, what is the probability that he (or she) attended classes on Thursday?

Solution: We define the following events:

Z_1 is the students attend class on Thursday.

Z_2 is the students do not attend class on Thursday.

Therefore, we find that Z_1 and Z_2 are a partition of Ω

B is the students pass the course.

For a) From the text of the problem we have:

$$P(Z_1) = 0.6,\ P(Z_2) = 1 - P(Z_1) = 0.4,\ P(B \mid Z_1) = 0.98,\ P(B \mid Z_2) = 0.2$$

Then by using the total probability formula, we get:

$$P(B) = P(Z_1) \cdot P(B \mid Z_1) + P(Z_2) \cdot P(B \mid Z_2) = 0.6 \times 0.98 + 0.4 \times 0.2 = 0.668$$

For b) By Bayes' theorem, we have:

$$P(Z_1 \mid B) = \frac{P(Z_1) \cdot P(B \mid Z_1)}{P(Z_1) \cdot P(B \mid Z_1) + P(Z_2) \cdot P(B \mid Z_2)} = \frac{0.6 \times 0.98}{0.6 \times 0.98 + 0.4 \times 0.2} = 0.88$$

◀

▶ **EXAMPLE 2.4.11** In a particular population, 30% of people drive cars of type K, 15% of people drive cars of type L and the rest (55%) of people drive cars of type M. It is known that 10% of people driving cars of type K have accidents, 7% of people driving cars of type L have accidents, and 12% of people driving cars of type M have accidents. If we elected randomly one of this population, and we find that he had an accident, what is the probability that this person driving a car of type L?

Solution: We suppose that:

Z_1 is the event that, the person driving a car of type K. Then we have $P(Z_1) = 0.30$.

Z_2 is the event that, the person driving a car of type L. Then we have $P(Z_2) = 0.15$.

Z_3 is the event that, the person driving a car of type M. Then we have $P(Z_3) = 0.55$.

A is the event that, the person had an accident.

Then we have:

$$P(A \mid Z_1) = 0.10\,,\ P(A \mid Z_2) = 0.07 \text{ and } P(A \mid Z_3) = 0.12\,.$$

The probability to calculate is $P(Z_1 \mid A) = ?$, where we have:

$$P(Z_2 \mid A) = \frac{P(Z_2 \cap A)}{P(A)} = \frac{P(Z_2) \cdot P(A \mid Z_2)}{P(A)}$$

But we note that the events Z_1, Z_2 and Z_3 are a partition of the certain event Ω, Therefore, we can calculate $P(A)$ by using the total probability formula. Where have:

$$P(A) = \sum_{i \in I} P(Z_i)\ P(A \mid Z_i) = \frac{30}{100}\frac{10}{100} + \frac{15}{100}\frac{7}{100} + \frac{55}{100}\frac{12}{100} = \frac{1065}{10000} = 0.1065$$

Therefore, we have (by using the Bayes' formula):

$$P(Z_2 \mid A) = \frac{P(Z_2)\ P(A \mid Z_2)}{P(A)} = \frac{(15/100)(7/100)}{1065/10000} = \frac{105}{1065} = 0.0986$$

EXERCISES

1. Determine 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?

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, to determine the space of elementary events for all breakfasts Noura 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 (English characters) followed by three single-digit numbers, how many different license plates are possible?

8. A chairman must choose five secretaries from among 10 applicants and assign them to different stations. How many different arrangements are possible?

9. A health inspector has time to visit seven of the 20 restaurants on a list. How many different routes are possible?

10. 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?

11. 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.

    a. For any given spark plug, what is the probability that it will not fail during the first 20 minutes.

b. What is the probability that none of the four spark plugs will fail during the first 20 minutes of flight?

c. What is the probability that at least one of the spark plugs will fail during the first 20 minutes?

d. 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?

12. Fatin and Ali both drive older cars that don't always work in the wintertime. Suppose that Fatin's car works 90% of the time and Ali's works 75% of the time. What is the probability their cars will both be working on any given winter morning?

13. 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, $P\left(A\right)=0.134$, $P\left(B\right)=0.254$ and $P\left(A\cap B\right)=0.080$. What is the probability that the household selected is better off or educated?

14. 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 0.85.

a. What is the probability that a call is not for medical assistance?

b. Assuming that successive calls are independent, what is the probability that both of two successive calls will be for medical assistance?

c. What is the probability that three consecutive calls are not for medical assistance?

d. What is the probability that of the next 10 calls, at least one is for medical assistance?

15. If you eat at Star Café there's a 50% chance that your food will be cold and a 40% chance your food will taste bad. We assume that these two events are independent. Then:

a. What is the probability that both will occur, your food is cold and it tastes bad?

b. What is the probability that your food is cold or it tastes bad?

16. When spot-checked for safety, automobiles are found to defective tires 25% of the time, defective lights 35% of the time, and both defective tires and lights 10% of the time. Find the probability that a randomly chosen car has defective lights or that its tires are found to be defective.

17. 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. Then:

a. What is the probability that both bids are successful?

b. What is the probability that neither bid is successful?

c. What is the probability that at least one of the bids is successful?

18. There are two traffic lights on the route used by student to go from home to university. Let E denote the event that the student 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:

a. Must stop for at least one light?

b. Doesn't stop at either light?

c. Must stop at exactly one light?

d. Must stop just at the first light?

19. The following table provides a joint probability distribution for the marital status of a specific city adults by gender.

	Never Married M_1	**Married** M_2	**Widowed** M_3	**Divorced** M_4	$P(GM)$
Male (G_1)	0.129	0.298	0.013	0.040	0.480
Female (G_2)	0.104	0.305	0.057	0.054	0.520
$P(MG)$	0.233	0.603	0.070	0.095	1.00

a. Determine the probability that the adult selected is divorced.

b. Determine the probability that the adult selected is male.

c. Determine the probability that the adult selected is divorced and male.

d. Determine the probability that the adult selected is divorced, given that the adult selected is male.

e. Determine the probability that the adult selected is male, given that the adult is divorced.

20. 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?

21. An automobile dealer has kept records on the customers who visited his showroom. 40% 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.

a. What is the probability that a customer entering the showroom will buy an automobile?

b. Suppose a customer visited the showroom and purchased a car, what is the probability that the customer was a woman?

c. Suppose a customer visited the showroom but did not purchase a car, what is the probability that the customer was a man?

22. In a factory for the production of switches there are 4 lines that produce 15%, 25%, 20% and 40%. The percentage of defected switches in the production of these lines are 3%, 2%, 3% and 1% respectively. We select a switch randomly of the total production of the factory, then:

a. What is the probability that the switch is defected?

b. If we found that the switch is valid for work (not defected), what is the probability that this switch produced by the second line?

23. A factory has four machines M_1, M_2, M_3 and M_4. If these machines have the same capacity to produce. Furthermore, we know that, the defective items from these machines are 7%, 5%, 3% and 2% respectively. Now, if an item selected at random, then:

a. Calculate the probability that the selected item is valid for work (not defective).

b. If we find that the selected item is defective, what is the probability that this item was made by machine M_2?

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.
2. Distinguish various probability distributions.
3. Understand relations among probability models. Explain what is meant by expectation.
4. Learn some new distributions.

INTRODUCTION

In this Chapter we will learn two 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, $\ldots n$ 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
- EXERCISES

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 tossing two times, then the space of elementary events Ω 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 Ω as follow:

$$X : \Omega = \{\underbrace{HH}_{\omega_1}, \underbrace{HT}_{\omega_2}, \underbrace{TH}_{\omega_3}, \underbrace{TT}_{\omega_4}\} \longrightarrow \mathbb{R}$$

$$\omega \mapsto X(\omega) = \begin{cases} 0 & \text{for } \omega = TT \\ 1 & \text{for } \omega = TH, HT \\ 2 & \text{for } \omega = HH \end{cases}$$

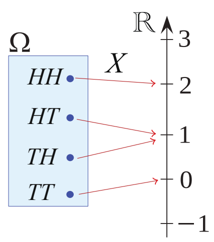


Figure 3.1.1

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 to 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^{Ω^*} is an event belongs to 2^{Ω}, then one say that X is a random variable on Ω.

Refer to the previous example we note that:

$X^{-1}(\{0\}) = \{TT\}$ is an event of 2^{Ω}

$X^{-1}(\{1\}) = \{HT, TH\}$ is an event of 2^{Ω}

$X^{-1}(\{2\}) = \{HH\}$ is an event of 2^{Ω}

So the map X is a random variable on Ω.

However, the correct definition of the random variable as follow.

DEFINITION 3.1.1 (Random Variable)

Let $[\Omega, \mathscr{A}, P]$ a probability space, and X a map (or a function) defined on a space of elementary events Ω with values in $\mathbb{R}$. Now if the inverse image of any event B of $\Re$ is an event in $\mathscr{A}$. This means:

$$X^{-1}(B) \in \mathscr{A} \quad ; \forall\, B \in \Re$$

Then one say that X is a random variable on the probability space $[\Omega, \mathscr{A}, P]$.

REMARKS 3.1.1

1. Usually a random variable (**random variable**) is denoted by capital letters X, Y, Z, …, whereas their values are denoted by small letters such as x, y, z, ... 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 a probability space $[\Omega, \mathscr{A}, P]$ if and only if:

   $$\{\omega \in \Omega\, ; X(\omega) \leq x\} \in \mathscr{A} \quad ; \forall x \in (-\infty, +\infty) = \mathbb{R}$$

   For example, if we go back to the previous example, we find:

   $$\{\omega \in \Omega\, ; X(\omega) \leq x\} = \begin{cases} \varnothing & \textit{for } x < 0 \\ \{TT\} & \textit{for } 0 \leq x < 1 \\ \{TT, TH, HT\} & \textit{for } 1 \leq x < 2 \\ \{TT, HT, TH, HH\} = \Omega & \textit{for } x \geq 2 \end{cases}$$

   But we know that $\varnothing$, $\{TT\}$, $\{TT, TH, HT\}$ and Ω are elements of $\mathscr{A} = 2^{\Omega}$. Therefore, the map X is a random variable on the probability space $[\Omega, \mathscr{A}, P]$.

3. For a random variable X we write as a shortcut $P\left(X = x\right), P\left(X < x\right)$, $P\left(X \leq x\right), P\left(X > x\right)$ and $P\left(X \geq x\right)$ instead of $P\left(\{\omega \in \Omega\, ;\, X(\omega) = x\}\right)$, $P\left(\{\omega \in \Omega\, ;\, X(\omega) < x\}\right)$, $P\left(\{\omega \in \Omega\, ;\, X(\omega) \leq x\}\right)$, $P\left(\{\omega \in \Omega\, ;\, X(\omega) > x\}\right)$ and $P\left(\{\omega \in \Omega\, ;\, X(\omega) \geq x\}\right)$ respectively.

4. Note that if X is a random variable defined on a probability space $[\Omega, \mathscr{A}, P]$, then $Y = aX + b$, for any $a, b \in \mathbb{R}$ and $a \neq 0$, is a random variable on $[\Omega, \mathscr{A}, P]$.

5. According to the values of random variables, they are mostly classified into either discrete or continuous. One has to distinguish well between discrete and continuous

random variables, that is very important in probability theory because different techniques and mathematical calculations are used to describe their distributions and their properties.

DEFINITION 3.1.2 (Distribution Function of a Random Variable)
Let X be a random variable on a probability space $[\Omega, \mathscr{A}, P]$. Now for this random variable we define a real function F_X on $\mathbb{R}$ as follow:

$$F_X : \mathbb{R} \longrightarrow \mathbb{R} \ ; \ x \mapsto F_X(x) := P(X \le x)$$

The function F_X is called the distribution function (D.F.) of X.

REMARKS 3.1.2

1. For any value $x \in \mathbb{R}$ the value $F_X(x)$ is between 0 and 1, i.e.
$$0 \le F_X(x) \le 1 \quad \text{for any } x \in \mathbb{R}$$
2. One can prove that:
   - **a.** The distribution function F_X is a non-deceasing function, i.e. $F_X(x) \le F_X(y)$ for any $x < y$.
   - **b.** $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
   - **c.** The distribution function F_X is right continuous function at each point $x \in \mathbb{R}$.
3. For any two real values a and b with $a < b$, one can prove that:
$$P(a < X \le b) = F_X(b) - F_X(a).$$

▶ **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:

$$F_X(x) = P\left(\{\omega \in \Omega \ ; X(\omega) \le x\}\right) = \begin{cases} P(\varnothing) = 0 & \text{for } x < 0 \\ P(\{TT\}) = 1/4 & \text{for } 0 \le x < 1 \\ P(\{TT, TH, HT\}) = 3/4 & \text{for } 1 \le x < 2 \\ P(\{TT, HT, TH, HH\} = P(\Omega) = 4/4 & \text{for } x \ge 2 \end{cases}$$

This relation can be written and drawn graphically as follow:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 0.25 & \text{for } 0 \le x < 1 \\ 0.75 & \text{for } 1 \le x < 2 \\ 1 & \text{for } x \ge 2 \end{cases}$$

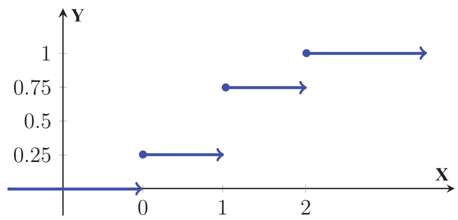


Figure 3.1.2 the graph of the distribution function of X

▶ **EXAMPLE 3.1.2** We rolling a fair die one time, and define a map X on the probability space $[\Omega, \mathscr{A} = 2^{\Omega}, P]$ of this random experiment as follow:

$$X : \Omega = \{1,2,3,4,5,6\} \longrightarrow \mathbb{R} \; ; \; \omega \mapsto X(\omega) = 2\omega$$

Then we find the values set of this map is $\mathbf{X} = \{2,4,6,8,10,12\}$. Moreover, we see that:

$$\{\omega \in \Omega \; ; X(\omega) \leq x\} = \begin{cases} \varnothing & \text{for } x < 2 \\ \{1\} & \text{for } 2 \leq x < 4 \\ \{1,2\} & \text{for } 4 \leq x < 6 \\ \{1,2,3\} & \text{for } 6 \leq x < 8 \\ \{1,2,3,4\} & \text{for } 8 \leq x < 10 \\ \{1,2,3,4,5\} & \text{for } 10 \leq x < 12 \\ \{1,2,3,4,5,6\} = \Omega & \text{for } x \geq 12 \end{cases}$$

We note that X is a random variable while the sets $\varnothing, \{1\}, \{1,2\}, \{1,2,3\}, \{1,2,3,4\}, \{1,2,3,4,5\}$ and Ω are elements of $\mathscr{A} = 2^{\Omega}$. Therefore, the distribution function of X has the following form:

$$F_X(x) = P\left(\{\omega \in \Omega \; ; X(\omega) \leq x\}\right) = \begin{cases} 0 & \text{for } x < 2 \\ 1/6 & \text{for } 2 \leq x < 4 \\ 2/6 & \text{for } 4 \leq x < 6 \\ 3/6 & \text{for } 6 \leq x < 8 \\ 4/6 & \text{for } 8 \leq x < 10 \\ 5/6 & \text{for } 10 \leq x < 12 \\ 1 & \text{for } x \geq 12 \end{cases}$$

◀

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 random variable X is said to be discrete if it takes its values on a countable set, whether this set is finite or infinite.

▶ **EXAMPLES 3.2.1** Note that the following statements are examples of discrete random variables:

1. X is a random variable, which observes the number of cars passing through a street during one hour of the day.
2. Y is a random variable, which observes the number phone calls received in a day by a police station.
3. Z is a random variable, which observes the number of heads in 5 tosses of a coin.
4. W is a random variable, which observes the number of rainy days during a year in Riyadh.
5. V is a random variable, which observes the number of defective parts in a shipment.

◀

▶ **EXAMPLE 3.2.2** A fair 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 results (elementary 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, and 3.

◀

REMARKS 3.2.1
In Example 3.2.2, 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 "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 random variable X	0	1	2	3
$p_x = P(X = x)$	1/8	3/8	3/8	1/8

Here, we note the set of the random variable values is $\{0, 1, 2, 3\}$.

DEFINITION 3.2.2 (Probability Mass Function)

Let X be a discrete random variable on a probability space $[\Omega, \mathscr{A}, P]$. Then the function:

$$p_\bullet : \mathbb{R} \longrightarrow [0\ ,\ 1] \quad ; x \mapsto p_x := P\left(\{\omega \in \Omega \colon X(\omega) = x\}\right)$$

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

1. For a discrete random variable 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.**)).

2. A discrete random variable X with finite number of values $x_1, x_2, \dots, x_n$ can be represented in the following table (called the probability distribution table for X).

Table 3.2.3

x_i the values of the random variable X	x_1	x_2	.....	x_n
$p_i = P(X = x_i)$	p_1	p_2	.....	p_n

3. Assuming that X is a discrete random variable on a space of elementary events Ω with values $x_1, x_2, \dots, x_n$. Then this random variable can be graphically represented by drawing two orthogonal axes XoY, and then drawing a height p_i (called the jump of the random variable) at the value x_i for all possible values of i.

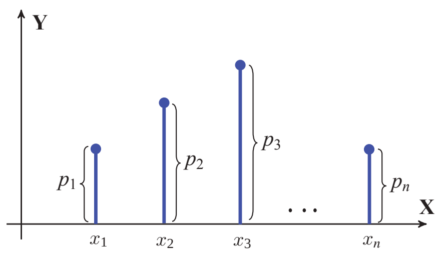


Figure 3.2.1

Note that tabular and graphical representation for a discreet 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.

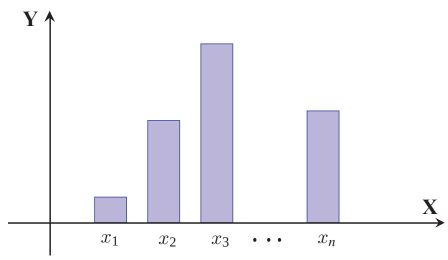


Figure 3.2.2

5. Usually, when x is an integer number, we use the letter k, 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 is thrown repeatedly until a 6, at the top, is obtained.

a. Determine the **p.m.f.** for the number of 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 of throwing the die to get 6. Thus $P(X = 1) = 1/6$, if one 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:

$$P(X = 2) = \left(1 - \frac{1}{6}\right)\left(\frac{1}{6}\right) = \left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = \left(\frac{5}{6^2}\right)$$

Hence, one gets 6 in the third trial if he gets other than 6 in the first two trial and 6 in the third:

$$P(X=3)=\left(1-\frac{1}{6}\right)\left(1-\frac{1}{6}\right)\left(\frac{1}{6}\right)$$
$$=\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)=\frac{5^2}{6^3}$$

and so on, we get:

$$P(X=k)=\underbrace{\left(1-\frac{1}{6}\right)\left(1-\frac{1}{6}\right)\cdots\left(1-\frac{1}{6}\right)}_{k-1\text{ factor}}\left(\frac{1}{6}\right)=\left(\frac{5}{6}\right)^{k-1}\left(\frac{1}{6}\right)=\frac{5^{k-1}}{6^k}\quad;\text{ for } k\in\mathbb{N}$$

Table 3.2.4

k the values of the random variable X	1	2	3	4	…	k	….
$p_k=P(X=k)$	1/6	$5/6^2$	$5^2/6^3$	$5^3/6^4$	…	$5^{k-1}/6^k$	….

For b) To show that $p_\bullet=P(X=\bullet)$ is a 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:

$$P(X=k)=\frac{5^{k-1}}{6^k}\geq 0\qquad ;\textit{for } k=1,2,3,\ldots$$

Also, we have:

$$\sum_{k=1}^{\infty}P(X=k)=\sum_{k=1}^{\infty}\frac{5^{k-1}}{6^k}$$
$$=\frac{1}{5}\sum_{k=1}^{\infty}\left(\frac{5}{6}\right)^k=\frac{1}{5}\left(\frac{5/6}{1-(5/6)}\right)=1$$

This random variable is a famous random variable and is known as "geometric random variable" with probability $p=1/6$. It is sometimes denoted by $X\sim G(p)$.

For c) Now if one wants to calculate that the probability of getting 6 in not more than two trials. This can be expressed as:

$$P(X\leq 2)=P(X=1)+P(X=2)$$
$$=\frac{1}{6}+\frac{1}{6}.\frac{5}{6}=\frac{11}{36}$$

For d) Similarly, the probability of getting 6 in more than 2 trials is:

$$P(X\geq 3)=\sum_{k=3}^{\infty}P(X=k)=1-\sum_{k=1}^{2}P(X=k)$$
$$=1-[P(X=1)+P(X=2)]=1-\left[\frac{1}{6}+\frac{5}{36}\right]=1-\frac{11}{36}=\frac{25}{36}$$

DEFINITION 3.2.3 (Distribution Function of a Discrete Random Variable)

Let X be a discrete random variable on a probability space $[\Omega, \mathscr{A}, P]$., where the set of its values $\mathbf{X} = \{x_i : i \in I\}$. Then the distribution function of X (denoted by F_X) is a real function on $\mathbb{R}$ given as follow:

$$F_X : \mathbb{R} \longrightarrow [0\ ,\ 1]$$

$$x \mapsto F_X(x) := \sum_{i \in I, x_i \leq x} P(X = x_i)$$

REMARKS 3.2.3

a. From the distribution function definition we observe that $0 \leq F_X(x) \leq 1$ always.

b. For each value x of the random variable X, the distribution function F_X has a jump up.

c. 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 $-1,\ -0.5,\ 0,\ 0.75,\ 1,\ 1.5$ with equal probabilities. We will represent this random variable 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	1	2	3	4	5	6
x_k	-1	-0.5	0	0.75	1	1.5
$p_k = P(X = x_k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The graphical representation of X is as in the following Figure.

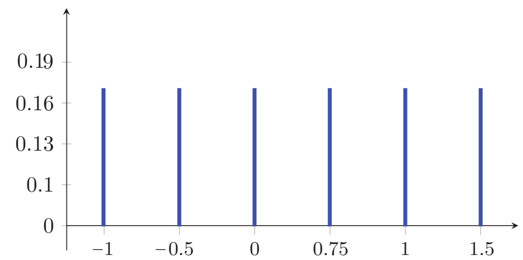


Figure 3.2.3

The probability mass function of this random variable is:

$$P(X = x_k) = \frac{1}{6} \quad \text{for any} \quad k = 1, 2, 3, 4, 5, 6$$

The distribution function of this random variable is:

$$F_X(x) = \sum_{k,\ x_k \le x} P(X = x_k) = \begin{cases} 0 & for\ \ x < -1 \\ \frac{1}{6} & for\ \ -1 \le x < -0.5 \\ \frac{2}{6} & for\ \ -0.5 \le x < 0 \\ \frac{3}{6} & for\ \ 0 \le x < 0.75 \\ \frac{4}{6} & for\ \ 0.75 \le x < 1 \\ \frac{5}{6} & for\ \ 1 \le x < 1.5 \\ \frac{6}{6} = 1 & for\ \ x \ge 1.5 \end{cases}$$

Then the graph of the distribution function F_X has the following form:

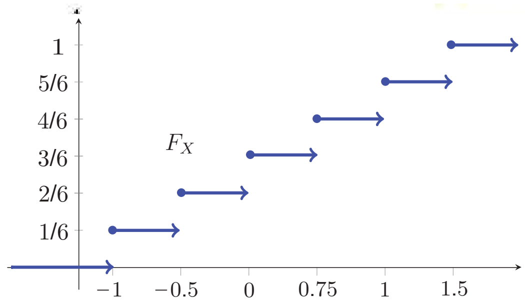


Figure 3.2.4

This random variable is a famous random variable, and it is called a **discrete uniform random variable** (or uniformly distributed random variable) with parameter $k = 6$. We will denote to this type of random variables by $X \sim DU(k)$.

REMARK 3.2.4

One can represent the distribution function of a discrete random variable X by

Table 3.2.6

For	$x < x_1$	$x_1 \le x < x_2$	$x_2 \le x < x_3$	...	$x_n \le x < x_{n+1}$	...
$F_X(x)$	0	p_1	$p_1 + p_2$	...	$p_1 + p_2 + \ldots + p_n$	...

If n (the number of values) is finite, then for $x \ge x_n$ we have $F_X(x) = 1$.

EXAMPLE 3.2.5 Consider tossing four fair different coins (at the same time). Let X be a random variable observes the number of heads on all four coins. Then.

a. Determine the probability space $[\Omega, \mathscr{A}, P]$ for this experiment.

b. What are the possible values for X?

c. Is the random variable X continuous or discrete?

d. Construct the probability distribution table for this random variable and its distribution function.

e. Give the tabular and graphical representation of X.

Solution: We have:

For a) The probability space for this experiment is $[\Omega, \mathscr{A}, P]$, whereas,:

$$\Omega = \left\{ \begin{array}{l} \overbrace{HHHH}^{\omega_1}, \overbrace{HHHT}^{\omega_2}, \overbrace{HHTH}^{\omega_3}, \overbrace{HTHH}^{\omega_4}, \overbrace{THHH}^{\omega_5}, \overbrace{HHTT}^{\omega_6}, \overbrace{HTHT}^{\omega_7}, \overbrace{THHT}^{\omega_8}, \\ \underbrace{HTTH}_{\omega_9}, \underbrace{TTHH}_{\omega_{10}}, \underbrace{THTH}_{\omega_{11}}, \underbrace{HTTT}_{\omega_{12}}, \underbrace{THTT}_{\omega_{13}}, \underbrace{TTHT}_{\omega_{14}}, \underbrace{TTTH}_{\omega_{15}}, \underbrace{TTTT}_{\omega_{16}} \end{array} \right\}, \ \mathscr{A} = 2^{\Omega} \text{ and}$$

$$P(A) = \sum_{i\ ;\ \omega_i \in A} P(\{\omega_i\})\ ;\ \ \forall A \in \mathscr{A}, \text{ where we have } P(\{\omega_i\}) = \frac{1}{16} \text{ for } i = 1, 2, \dots, 16\,.$$

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) To represent X tabular we must determine the probability mass function of X, where we have:

$$P(X = 0) = P(X = 4) = \frac{1}{16},$$

$$P(X = 1) = P(X = 3) = \frac{4}{16},$$

and $P(X = 2) = \dfrac{6}{16}$.

Table 3.2.7-a

k	0	1	2	3	4
$p_k = P\left(X = k\right)$	$1/2^4$	$4/2^4$	$6/2^4$	$4/2^4$	$1/2^4$

The table for the distribution function of the random variable X *as* follows:

Table 3.2.7-b

For	$x < 0$	$0 \le x < 1$	$1 \le x < 2$	$2 \le x < 3$	$3 \le x < 4$	$x \ge 4$
$F_X(x)$	0	$1/2^4$	$5/2^4$	$11/2^4$	$15/2^4$	1

The graphical representation of this random variable X is as follow:

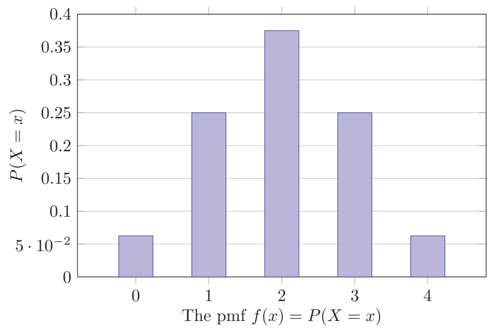


Figure 3.2.5

▶ **EXAMPLE 3.2.6** Consider tossing a fair coin four times. Let X a random variable observes the number of heads that appears up on a coin. Then

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

$$\Omega = \left\{ \begin{array}{l} \overbrace{HHHH}^{\omega_1}, \overbrace{HHHT}^{\omega_2}, \overbrace{HHTH}^{\omega_3}, \overbrace{HTHH}^{\omega_4}, \overbrace{THHH}^{\omega_5}, \overbrace{HHTT}^{\omega_6}, \overbrace{HTHT}^{\omega_7}, \overbrace{THHT}^{\omega_8}, \\ \underbrace{HTTH}_{\omega_9}, \underbrace{TTHH}_{\omega_{10}}, \underbrace{THTH}_{\omega_{11}}, \underbrace{HTTT}_{\omega_{12}}, \underbrace{THTT}_{\omega_{13}}, \underbrace{TTHT}_{\omega_{14}}, \underbrace{TTTH}_{\omega_{15}}, \underbrace{TTTT}_{\omega_{16}} \end{array} \right\}$$

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 3.2.7-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** Consider tossing two fair identical coins at the same time. Let X be a random variable, so that it observes the number of heads on all four coins. Then:

a. Determine the probability space $[\Omega, \mathscr{A}, P]$ for this experiment.

b. What are the possible values for X?

c. Determine the probability mass function of X.

Answers: We have:

For a) The probability space for this experiment is $[\Omega, \mathscr{A}, P]$, where we have:

The all posable results are HH, HT, TH and TT, but because of the symmetry property of coins, we cannot distinguish between the two outcomes HT and TH. Therefore, the space of elementary events is reduce to $\Omega = \{HH, HT, TT\}$:

We take

$$\mathscr{A} = 2^{\Omega} \text{ and } P(A) = \sum_{\omega\ ;\ \omega \in A} P(\{\omega\}) \quad ; \forall\, A \in \mathscr{A}\ ,$$

where we have:

$$P(\{HH\}) = \frac{1}{4}\ ,\ P(\{HT\}) = \frac{2}{4} = \frac{1}{2}\quad ,\quad P(\{TT\}) = \frac{1}{4}$$

For b) The possible values for X are $k = 0,\ 1,\ 2$.

For c) The probability mass function of X is:

$$P(X=0)=P(\{TT\})=\frac{1}{4},\ \ P(X=1)=P(\{HT\})=\frac{1}{2},\ \ P(X=2)=P(\{HH\})=\frac{1}{4}.$$

Table 3.2.8-a

k	0	1	2
$p_k = P(X=k)$	1/4	1/2	1/4

The table for the distribution function of the random variable X *as* follows:

Table 3.2.8-b

For	$x<0$	$0\le x<1$	$1\le x<2$	$x\ge 2$
$F_X(x)$	0	1/4	3/4	1

MATHEMATICAL EXPECTATION, MEAN AND VARIANCE FOR DISCRETE RANDOM VARIABLE

The mean and variance of a random variable are two special cases of the moments (or the mathematical expectation) of the random variables.

DEFINITIONS 3.2.4 (Moments of a Discrete Random Variable)

Let X be a discrete random variable on a probability space $[\Omega, \mathscr{A}, P]$ with values set $\mathbf{X}=\{x_i : i \in I\}$. Then the moment of order k or the mathematical expectation of the random variable X^k (denoted by $\mathrm{E}(X^k)$) is given by:

$$\mathrm{E}(X^k)=\sum_{i\in I} x_i^k\, P\left(X=x_i\right)$$

Special Cases

- If $k=1$, then $\mathrm{E}(X)$ is called the mean (or the mathematical expected value) of X (and can be denoted by μ also). Thus the mean (or the expected value) of X is:

$$\boxed{\mu = E(X)=\sum_{i\in I} x_i\, P(X=x_i)}$$

- If $k=2$, then $\mathrm{E}\left[X-\mathrm{E}(X)\right]^2$ is called the variance of this random variable X (denoted by σ^2 or $Var(X)$). This means that the variance of a random variable X is the mathematical expectation of the random variable $\left[X-\mathrm{E}(X)\right]^2$.
  This is given by

$$\boxed{\sigma^2 = Var(X)=\mathrm{E}\left[X-\mathrm{E}(X)\right]^2=\sum_{i\in I}(x_i-\mu)^2 P(X=x_i)}$$

DEFINITIONS 3.2.5 (Factorial Moments of a Discrete Random Variable)

Let X be a discrete random variable on a probability space $[\Omega, \mathscr{A}, P]$ with values set $\mathrm{X} = \{x_i : i \in I\}$. Then the factorial moment of order k (we denote it by $\mathscr{F}_k(X)$) of this random variable is given by:

$$\mathscr{F}_k(X) = \mathrm{E}\left[X(X-1)(X-2)...(X-k+1)\right]$$
$$= \sum_{i \in I}\left[x_i(x_i-1)(x_i-2)...(x_i-k+1)\right] P\left(X = x_i\right)$$

REMARK 3.2.5

We know that the positive square root of the variance is called the standard deviation (sd) (denoted by σ), thus we have:

$$\boxed{\sigma = +\sqrt{Var(X)}}$$

PROPERTIES OF EXPECTATION AND VARIANCE

Let a and b be constants and X is a discrete random variable, then:

1. For any constant real number b we have $\mathrm{E}(b)=b$.
2. For any two real numbers $a \neq 0$ and b, we have $\mathrm{E}\left(aX+b\right) = a\,\mathrm{E}\left(X\right)+b$.
3. For any real number a, we have $Var(a) = 0$.
4. For any two real numbers $a \neq 0$ and b, we have $Var(aX+b) = a^2 Var(X)$.

Using these properties one can prove the following:

5. Variance X can be expressed as $Var(X) = \mathrm{E}\left(X^2\right) - \left(\mathrm{E}(X)\right)^2$

   Its proof goes as follows;

$$Var(X) = \mathrm{E}\left[\left(X-\mathrm{E}(X)\right)^2\right] = \mathrm{E}\left[X^2 - 2\,X\,\mathrm{E}(X) + \left(\mathrm{E}(X)\right)^2\right]$$
$$= \mathrm{E}(X^2) - 2\,\mathrm{E}(X)\,\mathrm{E}(X) + \left(\mathrm{E}(X)\right)^2 = \mathrm{E}(X^2) - 2\left(\mathrm{E}(X)\right)^2 + \left(\mathrm{E}(X)\right)^2$$
$$= \mathrm{E}(X^2) - \left(\mathrm{E}(X)\right)^2$$

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
$p_k = P\left(X = k\right)$	$1/2^4$	$4/2^4$	$6/2^4$	$4/2^4$	$1/2^4$

Now we compute $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$ in the following in the following table:

Table 3.2.9-b

k	0	1	2	3	4
$p_k = P\left(X = k\right)$	$1/2^4$	$4/2^4$	$6/2^4$	$4/2^4$	$1/2^4$
$k \cdot P(X = k)$	0	$4/2^4$	$12/2^4$	$12/2^4$	$4/2^4$
$k^2 \cdot P(X = k)$	0	$4/2^4$	$24/2^4$	$36/2^4$	$16/2^4$

Hence the mean of this distribution is:

$$\mu = \mathrm{E}(X) = \sum_{i=1}^{\infty} x_i\, P(X = x_i) = \sum_{k=1}^{\infty} k\, P(X = k)$$

$$= 0 \times (1 / 2^4) + 1 \times (4 / 2^4) + 2 \times (6 / 2^4) + 3 \times (4 / 2^4) + 4 \times (1 / 2^4) = \frac{32}{16} = 2$$

where the expected value of X^2 (is called second moment of X) is:

$$\mathrm{E}(X^2) = \sum_{i=1}^{\infty} x_i^2\, P(X = x_i) = \sum_{i=1}^{\infty} k^2\, P(X = k)$$

$$= 0^2 \times (1 / 2^4) + 1^2 \times (4 / 2^4) + 2^2 \times (6 / 2^4) + 3^2 \times (4 / 2^4) + 4^2 \times (1 / 2^4) = \frac{80}{16} = 5$$

Now the variance of X is:

$$\sigma^2 = \mathrm{E}(X^2) - \left[\mathrm{E}(X)\right]^2 = \mathrm{E}(X^2) - \mu^2 = 5 - 2^2 = 1$$

The standard deviation for this random variable is:

$$\sigma = +\sqrt{1} = 1\,.$$

◀

▶ **EXAMPLE 3.2.9** Let f be a probability mass function (p.m.f.) for a discrete random variable X defined by the following relation:

$$f(k) = \begin{cases} 2c & for\ k = 10 \\ c & for\ k = 20 \\ c - 0.2 & for\ k = 30 \end{cases}$$

where c is a real constant. Then:

a. Determine the constant c.

b. Determine the distribution function of X.

c. Calculate the mean, the variance and standard deviation for X.

d. Calculate $\mathrm{E}\left(5X + 9\right)$ and $Var\left(3X + 14\right)$

Solution: We have

For a) According to the definition of a discrete random variable with probability mass function f, we can write:

$$\begin{aligned}1 &= \sum_{k\in\mathbb{Z}} f(k) = \sum_{k\in\mathbb{Z}} P(X=k)\\ &= P(X=10)+P(X=20)+P(X=30)\\ &= 2c+c+c-0.2\end{aligned}$$

Thus, we find $c = 0.3$.

For b) The distribution function of the random variable is:

$$F_X(x) = \sum_{k\in\mathbb{Z},\ k\leq x} P(X=k) = \begin{cases} 0 & \text{for} \quad x<10 \\ 0.6 & \text{for} \quad 10 \leq x < 20 \\ 0.9 & \text{for} \quad 20 \leq x < 30 \\ 1 & \text{for} \quad x \geq 30 \end{cases}$$

For c) The mean of this random variable is:

$$\begin{aligned}\mu = \mathrm{E}(X) &= \sum_x x\cdot P(X=x)\\ &= 10\,P(X=10)+20\,P(X=20)+30\,P(X=30)\\ &= (10\times 0.6)+(20\times 0.3)+(30\times 0.1) = 15\end{aligned}$$

Where the variance can be obtained by finding $\mathrm{E}(X^2)$ (the moment of order 2 of X) as follows:

$$\begin{aligned}\mathrm{E}(X^2) &= \sum_x x^2\cdot P(X=x)\\ &= 10^2\,P(X=10)+20^2\,P(X=20)+30^2\,P(X=30)\\ &= (100\times 0.6)+(400\times 0.3)+(900\times 0.1) = 270\end{aligned}$$

Hence the variance is:

$$\sigma^2 = Var(X) = \mathrm{E}(X^2)-\mu^2 = 270-225 = 45$$

Hence the standard deviation of the random variable X is:

$$\sigma = +\sqrt{Var(X)} = \sqrt{45} = 6.71$$

Now, by using the properties of the expectation, we have:

$$\mathrm{E}\left(5X+9\right) = 5\ \mathrm{E}\left(X\right)+9 = (5\times 15)+9 = 84$$

For d) Also, we have

$$Var(3X+14) = 3^2 Var\left(X\right)+Var(14) = (9\times 45)+0 = 405$$

◀

▶ **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 following form:

$$P\left(X=k\right) = \frac{k^3}{9\,c},\ k=1,\,2,\,3, \text{ where } c \text{ is a constant.}$$

a. Determine the value of c.

b. Determine the distribution function of the random variable X.

c. Calculate the mean and the variance of the random variable X.

Solution: We have:

For a) The value of c is the quantity that make $\sum_k P(X=k)=1$, hence:

$$1=\sum_k P(X=k)=P(X=1)+P(X=2)+P(X=3)$$
$$=\frac{1^3}{9c}+\frac{2^3}{9c}+\frac{3^3}{9c}=\frac{36}{9c}=\frac{4}{c}$$

Thus $c=4$.

For b) The distribution function of the random variable X is:

$$F_X(x)=\sum_{k\le x}P(X=k)=\begin{cases}0 & \text{for } x<1\\ \frac{1}{36} & \text{for } 1\le x<2\\ \frac{9}{36} & \text{for } 2\le x<3\\ \frac{9+27}{36}=1 & \text{for } x\ge 3\end{cases}$$

For c) The mean for this discrete random variable is:

$$\mu=\mathrm{E}(X)=\sum_{i=1}^{\infty}x_i\,P(X=x_i)=\sum_{k=1}^{3}k\,P(X=k)$$
$$=1\,P(X=1)+2\,P(X=2)+3\,P(X=3)$$
$$=\left(1\times\frac{1}{36}\right)+\left(2\times\frac{8}{36}\right)+\left(3\times\frac{27}{36}\right)=\frac{98}{36}=\frac{49}{18}=2.7\overline{2}$$

Now for the variance of the random variable X, we find:

$$\mathrm{E}(X^2)=\sum_{i=1}^{\infty}x_i^2\,P(X=x_i)=\sum_{k=1}^{3}k^2\,P(X=k)=1^2\,P(X=1)+2^2\,P(X=2)+3^2\,P(X=3)$$
$$=\frac{1}{36}+4\cdot\frac{8}{36}+9\cdot\frac{27}{36}=\frac{276}{36}=\frac{46}{6}=7.667$$

Hence the variance is:

$$\sigma_X^2=Var(X)=\mathrm{E}(X^2)-\mu^2=\frac{46}{6}-\left(\frac{49}{18}\right)^2=0.2561$$

◀

▶ **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 random variable X as follows:

$X = 0$ if no loss occurs, which happens with probability 1 - (1/400) = 0.9975

$X = -2000$, loss occurs which happens with probability $\frac{1}{400} = 0.0025$.

Now the following 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:

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} x_i\, P(X = x_i) = \sum_{k=1}^{2} k\, P(X = k) = 0 \times 0.995 + \left(-2000 \times 0.0025\right) = -5\ SR$$

Thus, the company must charge 5 SR to break even. Note that the insurance company 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 distribution.

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. X a random variable observes the number of good edited transcripts out of n transcripts that is either in compliance with procedures or it is not.

b. Y a random variable observes the number of correct guesses at 30 true-false questions when you randomly guess all answers

c. Z a random variable observes 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 $[\Omega, \mathscr{A}, P]$. Then one says that X has a **binomial distribution** with parameters $n \in \mathbb{N}$ and $1 > p > 0$, if the probability mass function is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad for\ k = 0, 1, 2, \ldots, n$$

REMARKS 3.2.7

1. We denote the binomial distribution with parameters n and p, by $B(n, p)$. It is usually written as $X \sim B(n, p)$.

2. Let X be a binomial distributed random variable with parameters n and p. Then the distribution function of X is given by:

$$F_X(x) = P(X \leq x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

3. According to Al-Karji (953-1029)-Newton's binomial expansion we note that:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \left[p + (1-p)\right]^n = 1$$

The Mean and the Variance of Binomial Distribution

The mean of the binomial distribution is derived as follows:

$$\mu = \mathrm{E}(X) = \sum_{k=0}^{n} k\, P(X = k) = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k}$$
$$= n\, p \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1}(1-p)^{n-k}$$
$$= n\, p\,(1-p) \sum_{k=1}^{n-1} \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k-1}$$
$$= n\, p\,(1-p) \sum_{j=0}^{m} \binom{m}{j} p^{j}(1-p)^{n-j} = n\, p\,[\,p + (1-p)]^m = np$$

To obtain the variance, we calculate the factorial moment of order 2, so we find:

$$\mathrm{E}[X(X-1)] = \sum_{k=0}^{n} k\,(k-1)\, P(X = k) = \sum_{k=2}^{n} k\,(k-1) \frac{n!}{k!(n-k)!}\, p^k\, (1-p)^{n-k}$$
$$= n\,(n-1)\, p^2 \underbrace{\sum_{k=2}^{n-2} \frac{(n-2)!}{(k-2)!\,\big[(n-2)-(k-2)\big]!}\, p^{k-2}\,(1-p)^{(n-2)-(k-2)}}_{=1}$$
$$= n\,(n-1)\, p^2$$

Thus, the variance of the binomial distribution is (by using **Steiner** formula):

$$\sigma^2 = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \mathrm{E}[X(X-1)] + \mathrm{E}(X) - \big(\mathrm{E}(X)\big)^2$$
$$= n\,(n-1)p^2 + n\,p - (n\,p)^2 = n\,p\,(1-p)$$

Therefore, the mean and the variance of binomial distribution $B(x\,;n\,,p)$ is given by:

$$\boxed{\mathrm{E}(X) = n\,p \quad \text{and} \quad Var(X) = n\,p\,(1-p)}$$

▶ **EXAMPLE 3.2.13** If the mean and the variance of a binomial distribution are 16 and 8 respectively, then:

- **a.** Determine the probability mass function.
- **b.** Calculate the probability $P\big(X = 0\big)$.
- **c.** Calculate the probability $P(X \geq 2)$.

Solution: We have:

For a) We have the mean equals

$$\mathrm{E}(X) = n\, p = 16 \;\Rightarrow n = \frac{16}{p}$$

The variance equals

$$Var(X) = n\, p\,(1-p) = 8 \;\Rightarrow \frac{16}{p}\, p\,(1-p) = 8 \;\Rightarrow p = \frac{1}{2}$$

Therefore, from these we find $n = 32$

The probability mass function is:

$$\begin{aligned} P(X = k) &= \binom{32}{k}(1/2)^k(1/2)^{32-k} \\ &= \frac{32!}{k!(32-k)!}(1/2)^k(1/2)^{32-k} \end{aligned}$$

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 random variable with parameters:

a. $n = 12$, $p = 1/2$

b. $n = 8$, $p = 1/3$

Draw the graph of X for the two cases.

Solution:

For a) Note that for $n = 12$, $p = 1/2$ we have for any $k = 0,1,\ldots,12$ the following:

$$P(X = k) = \binom{12}{k}(1/2)^k(1/2)^{12-k} = \frac{12!}{k!(12-k)!}(1/2)^k(1/2)^{12-k}$$

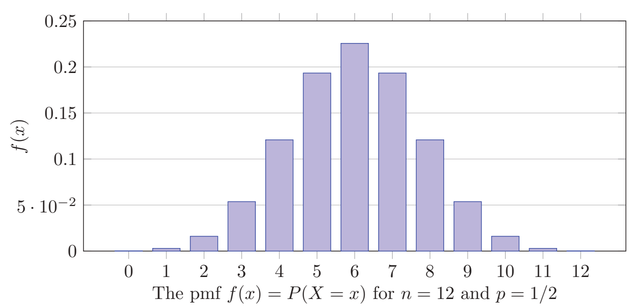


Figure 3.2.6

We note that the **p.m.f.** is symmetric around the mean.

Whereas the distribution function of X is:

$$F_X(x) = P(X \le x) = \sum_{k=0}^{x} \binom{12}{k} (1/2)^k (1/2)^{12-k} \qquad ;x \in \mathbb{R}$$

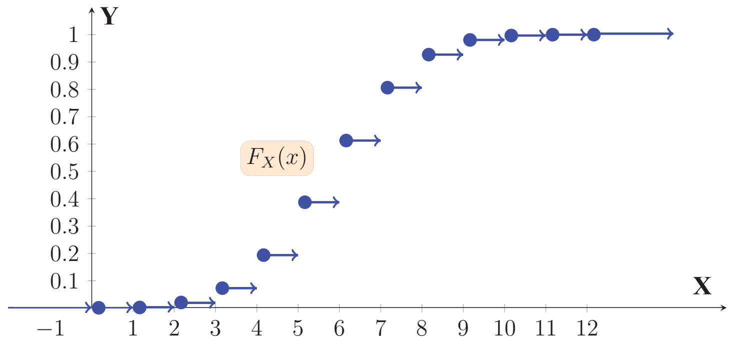


Figure 3.2.7

Note that the distribution function is a non-decreasing step function.

For b) Now for $n = 8,\ p = 1/3$ one has any $k = 0, 1, \dots, 8$:

$$P(X = k) = \binom{8}{k} (1/3)^k (2/3)^{8-k} = \frac{8!}{k!(8-k)!} (1/3)^k (2/3)^{8-k}$$

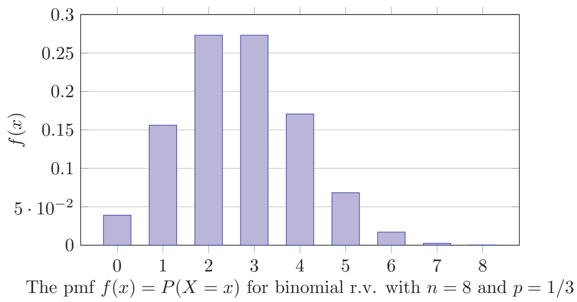


Figure 3.2.8

Note here that this binomial random variable is slightly skewed toward the right side.

Whereas the distribution function of X is

$$F_X(x) = P(X \le x) = \sum_{k=0}^{x} \binom{8}{k} (1/3)^k (2/3)^{8-k} \qquad ;x \in \mathbb{R}$$

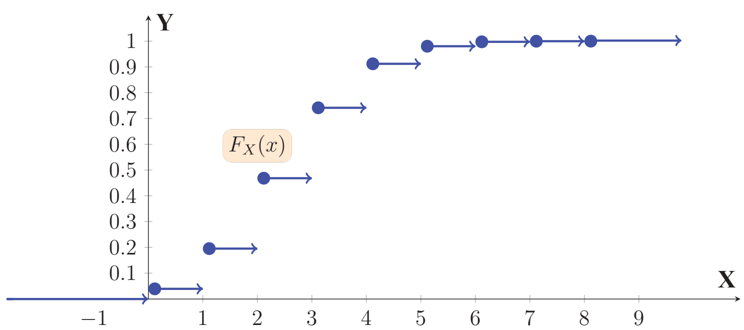


Figure 3.2.9

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 short intervals 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 short 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 the phenomenon occurrence is constant.
3. Each phenomenon is independent of all other phenomena

EXAMPLE 3.2.15 The following situations represent Poisson random variables:

a. X a random variable observes the number of phone calls arriving at a call center within a time unit.

b. Y a random variable observes the number of customers arriving at airline ticket purchasing office during one day.

c. Z a random variable observes the number of cars arriving at a traffic light during a rush hour.

d. U a random variable observes the 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 probability space $[\Omega, \mathscr{A}, P]$. Then one says that X has a **Poisson distribution** with parameter $\lambda > 0$, if the probability mass function is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad ; k = 0, 1, 2, \ldots$$

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:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}\left[1 + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \cdots\right] = e^{-\lambda}e^{\lambda} = 1$$

This is because we have Taylor expansion of the function:

$$f(x) = e^x = 1 + \frac{x^1}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \ldots$$

We sometimes denote to the Poisson distribution with parameter $\lambda > 0$ by $Po(\lambda)$ or simply $X \sim Po(\lambda)$.

The Mean and the Variance of Poisson Distribution

Let X be a Poisson distributed random variable with parameter $\lambda > 0$, then the mean of this random variable is derived as follows:

$$\mu = \mathrm{E}(X) = \sum_{k=0}^{n} k\, P(X = k) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

To obtain the variance, we have:

$$\begin{aligned} \mathrm{E}(X(X-1)) &= \sum_{k=0}^{n} k(k-1)\, P(X = k) = \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = e^{-\lambda} \lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2 \end{aligned}$$

Therefore, the variance of this random variable is:

$$\begin{aligned} \sigma^2 &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= \mathrm{E}(X(X-1)) + \mathrm{E}(X) - \left(\mathrm{E}(X)\right)^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

Therefore, the mean and the variance of the Poisson distribution with parameter $\lambda > 0$ are given by

$$\boxed{\mathrm{E}(X) = \lambda \quad \text{and} \quad Var(X) = \lambda}$$

▶ **EXAMPLE 3.2.16** In a large oil exploration firm company, engineering accidents occur independently at the mean of three per month $\lambda = 3$. Then the probability mass function is:

$$P(X = k) = \frac{3^k e^{-3}}{k!} \qquad ; k = 0, 1, 2, \ldots$$

Thus the instance probability of 4 accidents occurrence in one month equals to:

$$P(X=4)=\frac{3^4\,e^{-3}}{4\,!}=0.168$$

◀

▶ **EXAMPLE 3.2.17** Suppose the number of flaws in a 100–foot roll of paper is Poisson distributed with mean $\lambda = \ 10$.

a. Calculate the probability that there are eight flaws in a 100-foot roll.

b. Calculate the mean and the variance of the random variable.

Solution: Suppose that X is a random denotes the number of flaws in every printed page. Then X has a Poisson distribution with parameter $\lambda = 10$.

For a) The required probability is:

$$P(X=8)=\frac{10^8\,e^{-10}}{8\,!}=\frac{(100\,000\,000)\,(2.71828)^{-10}}{40320}=0.1126$$

For b) For the mean and the variance, we know that:

$$\mathrm{E}(X)\ \text{ and }\ Var\,(X)=\lambda=10\,.$$

◀

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 (an interval 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 set of real numbers, in spite of the fact that they are sometimes approximated by discrete values.

THEOREM 3.3.1

A random variable X defined on the probability space $[\Omega, \mathscr{A}, P]$ is said to be continuous if there exists a real non-negative function $f \geq 0$ on $\mathbb{R}$, for it the following relation is realized on any interval $(a, b) \subseteq \mathbb{R}$:

$$P\left(a \leq X \leq b\right) = \int_a^b f(x)\, dx$$

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:

$$P\left(a \leq X \leq b\right) = \int_a^b f_X(x)\, dx$$

DEFINITION 3.3.2 (Distribution Function of a Continuous Random Variable)

Let X be a continuous random variable on a probability space $[\Omega, \mathscr{A}, P]$ with density function f_X. Then the **distribution function** (DF) **of** X (denoted by F_X) is a real function defined on $\mathbb{R}$ by the following relation:

$$F_X(x) = P\left(X \leq x\right) = \int_{-\infty}^{x} f_X(u)\, du \qquad ;\forall\, x \in \mathbb{R}$$

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:

$$1 = \lim_{x \to +\infty} F_X\,(x) = \lim_{x \to +\infty} \int_{-\infty}^{x} f_X\,(t)\; dt = \int_{-\infty}^{+\infty} f_X\,(x)\; dx$$

This relation means that the area under its curve of f_X and for all values of x equal to one.

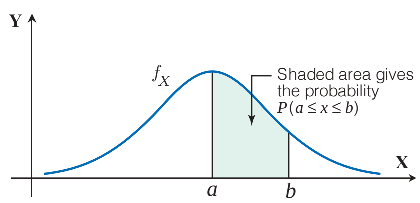


Figure 3.3.1-a

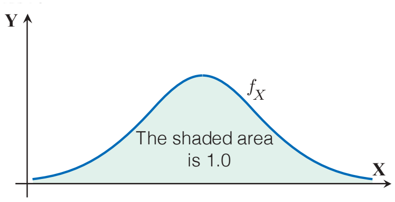


Figure 3.3.1-b

2. One can prove that $P\left(X = x\right) = 0$ for any value $x \in \mathbb{R}$. 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:

$$\boxed{f_X(x) = \frac{d}{dx}\, F_X(x) = F_X'(x)}$$

5. Because of the previous property (2) we find that for any continuous random variable we have:

$$P\left(a \leq x \leq b\right) = P\left(a < x \leq b\right) = P\left(a \leq x < b\right) = P\left(a < x < b\right)$$

▶ **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 $F_X(x) = \dfrac{x^4}{16} \quad ; 0 < x < 2$

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 between 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 $x \geq 0$ the pdf of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx}\left(1 - e^{-5x}\right) = 5e^{-5x}$$

◀

▶ **EXAMPLE 3.3.3** Let X be a continuous random variable with density function f_X given by the following relation:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & for \;\; a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

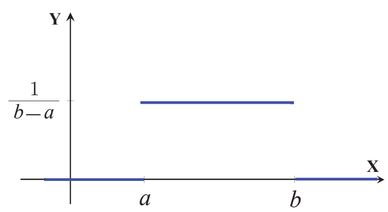


Figure 3.3.2

This function is known as the **p.d.f** of a **continuous uniform random variable** defined on $[a, b]$. The distribution function for this random variable X is:

$$F_X(x) = P\left(X \leq x\right) = \int_{-\infty}^{a} \underbrace{f_X(t)}_{0} \, dt + \int_{a}^{x} f_X(t) \, dt$$

$$= \begin{cases} 0 & for \;\; x < a \\ \dfrac{x-a}{b-a} & for \;\; a \leq x < b \\ 1 & for \;\; x \geq b \end{cases}$$

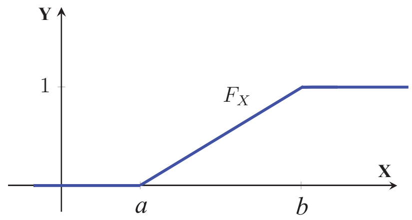


Figure 3.3.3

MATHEMTICAL EXPECTATION, MEAN AND VARIANCE FOR CONTINUOUS RANDOM VARIABLE

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 $[\Omega, \mathscr{A}, P]$ with density function f_X. Then the expected value or mean is defined by:

$$\boxed{\mu = \mathrm{E}(X) = \int_{-\infty}^{+\infty} u\, f_X(u)\; du}$$

Whereas, the variance of X is given by:

$$\boxed{\sigma^2 = Var(X) = \mathrm{E}\big(X - \mathrm{E}(X)\big)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, f_X(x)\; dx}$$

Using the Steiner formula, we get:

$$\sigma^2 = Var(X) = \int_{-\infty}^{+\infty} x^2\, f_X(x)\; dx - \left[\int_{-\infty}^{+\infty} x\, f_X(x)\; dx\right]^2$$

Further, the moment of order k for the continuous random variable X is

$$\boxed{\mathrm{E}(X^k) = \int_{-\infty}^{+\infty} u^k f_X(u)\; du}$$

▶ **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 **p.d.f.** of uniform distribution, we find that the mean of X is:

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{+\infty} x\, f_X(x)\; dx \\ &= \int_a^b \frac{x}{b-a}\; dx = \frac{1}{b-a}\left(\frac{x^2}{2}\bigg|_a^b\right) = \frac{a+b}{2}\end{aligned}$$

The corresponding variance of this random variable is obtained by using Steiner formula, therefore, we must calculate the moment of order 2 of X.

$$\begin{aligned} \mathrm{E}(X^2) &= \int_{-\infty}^{+\infty} x^2 f_X(x)\, dx = \frac{1}{b-a}\int_a^b x^2\, dx \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} \\ &= \frac{(b^2+ab+a^2)}{3} \end{aligned}$$

Therefore, the variance of X is:

$$\begin{aligned} Var(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

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:

$$F_X(x) = 1 - e^{-x}(1+x) \qquad ;x \geq 0$$

We will determine the density function of X, and then we calculate $P(1 \leq X < 2)$.

Solution: It has been noted that $f_X(x) = \frac{d}{dx} F_X(x)$. Therefore, we have:

$$\begin{aligned} f_X(x) &= \frac{d}{dx}[1 - e^{-x}(1+x)] \\ &= e^{-x}(1+x) - e^{-x} = x\ e^{-x} \end{aligned}$$

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:

$$f_X(x) = 2x \qquad ;0 \leq x \leq 1$$

Now we will answer the following:

a. Prove that f_X is a probability density function.

b. Calculate the probability $P\left(-(1/2) \leq X < 1/2\right)$.

c. Determine the distribution function F_X.

d. Calculate the mean, the variance and the standard deviation of X.

e. Draw the graph of the probability density function f_X.

Solution: We have:

For a) By the definition of f_X, we have $f_X(x) = 0$ 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 probability density function.

For b) To calculate $P\left(-(1/2) \leq X < 1/2\right)$, we have:

$$\mathrm{P}\left(-\frac{1}{2} \leq X < \frac{1}{2} \right) = \int_{-1/2}^{1/2} f_X(x)\ dx = \int_{-1/2}^{0} 0\ dx + \int_{0}^{1/2} 2x\ dx = 0 + \left(x^2 \Big|_0^{1/2} \right) = \frac{1}{4}$$

For c) To determine the distribution function for the random variable X, we have

For $x < 0$, $F_X(x) = 0$,

For $0 \leq x < 1$ we have $F_X(x) = \frac{1}{2}(x)(2x) = x^2$

Whereas, for $x \geq 1$ we have $F_X(x) = \frac{1}{2}(1)(2) = 1$

Thus, we get:

$$F_X(x) = \begin{cases} 0 & \textit{for } x < 0 \\ x^2 & \textit{for } 0 \leq x < 1 \\ 1 & \textit{for } x \geq 1 \end{cases}$$

For d) Now the mean of X is:

$$\mathrm{E}(X) = \int_0^1 x\, f_X(x)\ dx = \int_0^1 x\,(2x)\ dx$$

$$= \int_0^1 2x^2\ dx = \left(\frac{2}{3} - 0 \right) = \frac{2}{3}$$

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:

$$Var(X) = E(X^2) - \left[E(X)\right]^2 = \left(\frac{1}{2}\right) - \left(\frac{2}{3}\right)^2$$
$$= \frac{1}{2} - \frac{4}{9} = \frac{1}{18} = 0.056$$

The standard deviation is:

$$\sigma = \sqrt{Var(X)} = \sqrt{0.056} = 0.24$$

For e) The graph of the density function $f_X(x) = 2x$ for $0 \le x \le 1$ is in the following Figure.

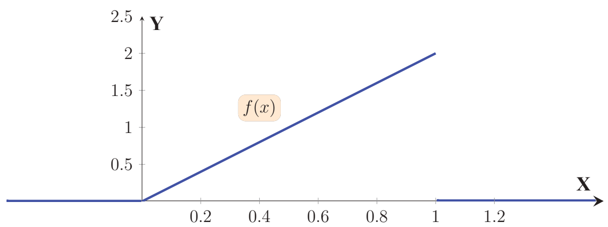


Figure 3.3.4

SOME SPECIAL DISTRIBUTIONS

Below we will present another two 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, λ.

DEFINITION 3.3.3 (Exponential Distributed Random Variable)

A continuous random variable X on a probability space $[\Omega, \mathscr{A}, P]$ is said to have an **exponential distribution** with parameter $\lambda > 0$ (or it is exponentially distributed) if its probability density function is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \textit{for } x \geq 0 \\ 0 & \textit{for } x < 0 \end{cases}$$

Usually the parameter λ 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 is given by the following relation:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \textit{for } x \geq 0 \\ 0 & \textit{for } x < 0 \end{cases}$$

One may draw, for example, the graph of the probability density function and the distribution function for the exponential distribution with mean $\mu = 4$ or $\lambda = 0.25$ are as follows:

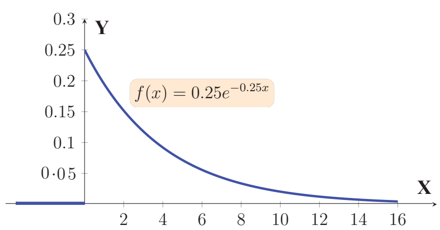


Figure 3.3.5

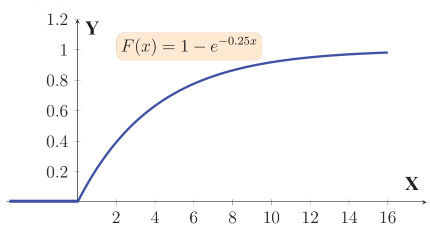


Figure 3.3.6

Mean and Variance of Exponential Random Variable

Let X be an exponential random variable with parameter $\lambda > 0$. Then the mean (or the average time between events) of X is:

$$\mathrm{E}(X) = \int_{-\infty}^{0} x\, f_X(x)\; dx + \int_{0}^{\infty} x\, f_X(x)\; dx = \int_{0}^{\infty} x\,\lambda\, e^{-\lambda x}\; dx = \frac{1}{\lambda}$$

To evaluate the variance, we first calculate $\mathrm{E}(X^2)$, that is:

$$\mathrm{E}(X^2) = \int_{-\infty}^{0} x^2\, f_X(x)\; dx + \int_{0}^{\infty} x^2\, f_X(x)\; dx$$
$$= \int_{0}^{\infty} x^2\,\lambda\, e^{-\lambda x}\; dx = \frac{2}{\lambda^2}$$

Thus, the variance is:

$$Var(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Therefore, the standard deviation is:

$$\sigma = +\sqrt{Var(X)} = \frac{1}{\lambda}$$

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. Then answer the following:

a. What is the probability that a refrigerator can be used for less than 6 years?

b. What is the probability that this refrigerator can be used for more than 18 years?

c. What is the variance and the standard deviation of this random variable?

Solution: The random variable X has an exponential distribution with mean $\mu = \frac{1}{\lambda} = 15$.

Thus the corresponding pdf of the life length of these refrigerators is:

$$f_X(x) = \frac{1}{15} e^{\frac{-x}{15}} \quad for\ x \geq 0$$

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:

$$P(X \geq 18) = 1 - F_X(x) = 1 - \left(1 - e^{-18/15}\right) = 0.3012$$

For c) The variance of this random variable is:

$$Var(X) = \frac{1}{\lambda^2} = \frac{1}{(1/15)^2} = 225$$

And from there we find, the standard deviation is:

$$\sigma = \frac{1}{\lambda} = \frac{1}{(1\,/\,15)} = 15$$

THE NORMAL DISTRIBUTION

In this part of the book we present the normal distribution and direct use of *Z*-Table for calculating the probabilities of the normal distribution. The normal distribution is sometimes known by the Gaussian distribution. It 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.

The following statements represent some random variables that usually follow normal distributions:

1. Human height, weight or age.
2. Human intelligence (IQ) in most communities.
3. Strength of a steel girder or steel bars.
4. Number 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 $[\Omega, \mathscr{A}, P]$ is said to have a **normal distribution** with parameters $\mu \in \mathbb{R}$ (location parameter) and $\sigma > 0$ (scale parameter) if its probability density function is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad ; x \in (-\infty\ ,\ \infty)$$

One denote the distribution by the symbol $N\left(\mu, \sigma^2\right)$, and write for this **r.v.** $X \sim N\left(\mu, \sigma^2\right)$.

The graph of the density and the distribution function for the normal distribution with several values for mean μ and standard deviation σ are as follows:

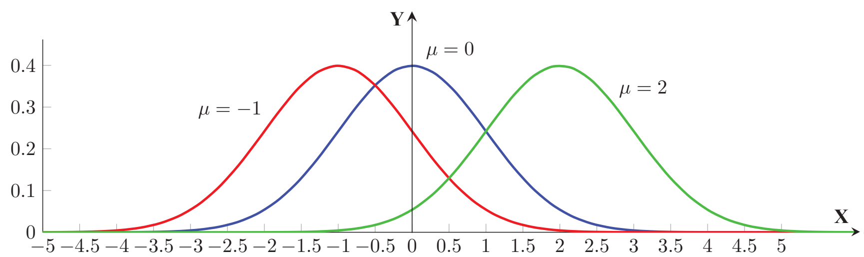


Figure 3.3.7 (The density function for a normal distribution for $\sigma = 1$)

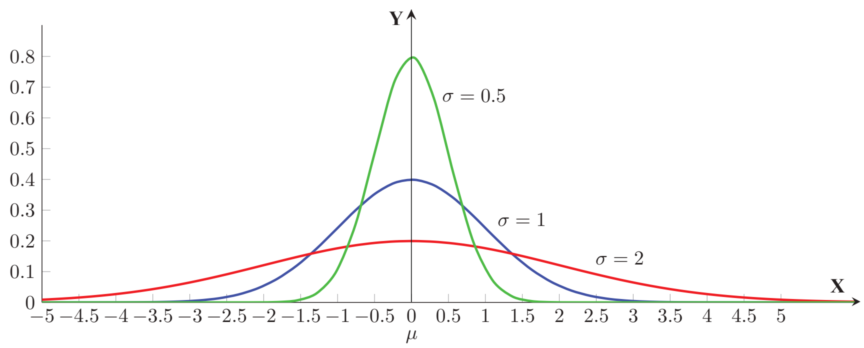


Figure 3.3.8 (The density function for a normal distribution for $\mu = 0$)

REMARKS 3.3.3

1. The normal p.d.f. curve, for any values of the mean and the standard deviation is bell-shaped.
2. The p.d.f. has a single peak at the exact center of the p.d.f. curve. This value of the random variable represents the mean of the distribution.
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 call it "standard normal distribution". The p.d.f. of the standard normal distribution (denoted by $\varphi(x)$) has the following relation and graph:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad ; x \in (-\infty\ ,\ \infty)$$

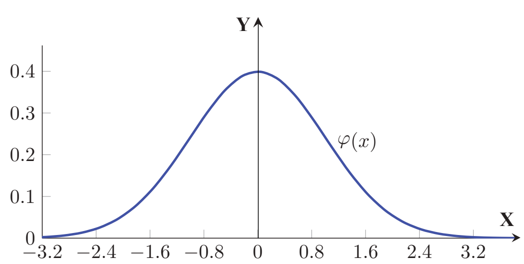


Figure 3.3.9

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 is given by the following relation:

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \qquad ; x \in \mathbb{R}$$

The graph of the normal distribution function for $\mu = 2$ and $\sigma = 2$ is as follow:

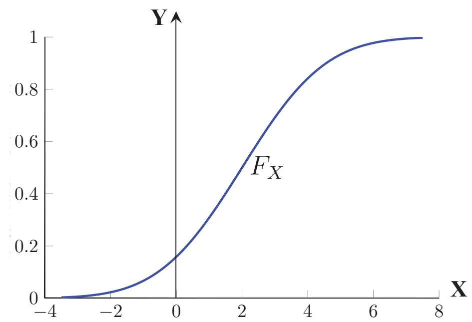


Figure 3.3.10 (the normal distribution function for $\mu = 2$ and $\sigma = 2$)

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:

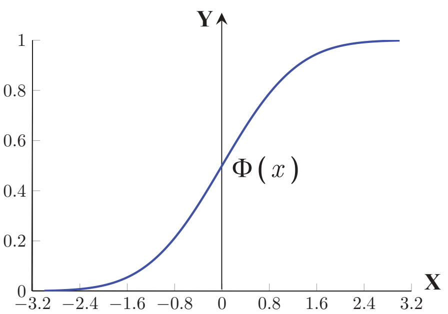


Figure 3.3.11 (The standard normal distribution function)

STANDARDIZING NORMALLY DISTRIBUTED RANDOM VARIABLE

In fact, standardizing normal distribution simplifies the computation of areas under normal density function 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 applying the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$

By standardizing a normal distribution, we eliminate the need to consider μ and σ. Hence we have a standard frame of reference for all calculation of probabilities.

Note that the area under the **p.d.f.** curve before and after the y-axes equals 0.5. The area under the curve of $\varphi(x)$ is tabulated for different values of z (z is the value of the random variable Z) and known by the normal distribution table or the z-table, or the integration from $-\infty$ to any value z of the density function $\varphi(z)$.

The standard normal distribution function $\Phi(z)$ is given by:

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt \qquad ; z \in \mathbb{R}$$

This value $\Phi(z)$ represents the area under the normal pdf curve until z, and is represented by the graph in the following.

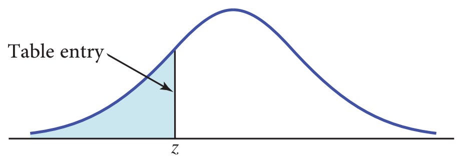


Figure 3.3.12

▶ **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 equals 80 and standard deviation equals to 5.

a. What proportion of GAT scores falls below 75?

b. What proportion of GAT scores falls between 76 and 82?

Solution: We have:

For a) By standardizing the normal random variable X we get:

$$P(X < 75) = P\left(\frac{X-\mu}{\sigma} < \frac{75-\mu}{\sigma}\right)$$
$$= P\left(\underbrace{\frac{X-80}{5}}_{Z} < \underbrace{\frac{75-80}{5}}_{z=-1}\right) = P(Z < -1) = \Phi(-1) = 0.1587$$

This value is obtained from the standard normal distribution table or the z-table for $z=-1$, that is the second column adjacent to the $z=-1$ value. This means that 15.87% (of the scores lies below 75).

For b) By standardizing the normal random variable X we get:

$$P(76 < X < 82) = P\left(\underbrace{\frac{76-80}{5}}_{z_1=-0.8} < \underbrace{\frac{X-\mu}{\sigma}}_{Z} < \underbrace{\frac{82-80}{5}}_{z_2=0.4}\right)$$
$$= P\left(-0.8 < Z < 0.4\right) = P(Z < 0.4) - P(Z < -0.8) = \Phi(0.4) - \Phi(-0.8)$$
$$= 0.6554 - 0.2119 = 0.4435$$

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%.

z	0.00	0.01	⋯	0.07	0.08	0.09
0.1	0.5398	0.5438		0.5675	0.5714	0.5753
0.2	0.5793	0.5832		0.6964	0.6103	0.6141
0.3	0.6179	0.6217		0.6443	0.6480	0.6517
0.4	0.6554	0.6591		0.6808	0.6844	0.6879
0.5	0.6915	0.6950		0.7157	0.7190	0.7224

z	0.00	0.01	⋯	0.07	0.08	0.09
−1.0	0.1587	0.1562		0.1423	0.1401	0.1379
−0.9	0.1841	0.1814		0.1660	0.1635	0.1611
−0.8	0.2119	0.2090		0.1922	0.1894	0.1867
−0.7	0.2420	0.2389		0.2206	0.2177	0.2148

◀

▶ **EXAMPLE 3.3.9** Use the standard normal distribution 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^{th} percentile or the z -value that covers p of the area under the normal distribution 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 standard normal distribution table (or the z-table). We note that this probability is; less than 0.5, so the value of z has to be in the left half of the normal distribution 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 -1.7 and the fifth value of the column gives the fraction part of the z-values that is 0.04. Thus $z = -1.74$.

z	0.00	0.01	⋯	0.04	…	0.09
−2.0	0.0228	0.0222		0.0207		0.0183
−1.9	0.0287	0.0281		0.0262		0.0233
−1.8	0.0359	0.0351		0.0329		0.0294
−1.7	0.0446	0.0436		0.0409		0.0367
−1.6	0.0548	0.0537		0.0505		0.0455

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 for 0.80 in the standard normal table is 0.7995 which has $z = 0.8$ of the first column and on the sixth column of the same z-row the value of $z = 0.84$. Thus we take $z = 0.84$.

z	0.00	0.01	···	0.04	···	0.09
0.5	0.6915	0.6950		0.7054		0.7224
0.6	0.7257	0.7291		0.7389		0.7549
0.7	0.7580	0.7611		0.7704		0.7852
0.8	0.7881	0.7910		0.7995		0.8133
0.9	0.8159	0.8186		0.8264		0.8389

For c) Also, for $P(Z > z) = 0.4090$, we can write:

$$0.4090 = P(Z > z) = 1 - P(Z \leq z)$$

This implies.

$$P(Z \leq z) = 1 - 0.4090 = 0.9510$$

Thus, the value is in the second half and in the standard normal distribution table gives $z = 0.23$

z	0.00	0.01	0.02	0.03	···	0.09
0.1	0.5398	0.5438	0.5478	0.5517		0.5753
0.2	0.5793	0.5832	0.5871	0.5910		0.6141
0.3	0.6179	0.6217	0.6255	0.6293		0.6517
0.4	0.6554	0.6591	0.6628	0.6664		0.6879
0.5	0.6915	0.6950	0.6985	0.7019		0.7224

▶ **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. Calculate 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 standard normal distribution table, this corresponds to the probability of 0.0668.

EXERCISES

1. Let X be a discrete random variable representing the sum of the two numbers on throwing two balanced (or fair) dice for one time only (make the solution for different and identical dice). then

   a. Find the possible values of the random variable X.

   b. What is the probability mass function $P\left(X = \bullet\right)$?

   c. Calculate the distribution function $F_X\left(x\right)$.

   d. Calculate the mean and variance for the random variable X.

2. We tossing a fair coin three times, and we suppose that is the $[\Omega, \mathscr{A}, P]$ probability space of this random experiment. If $X : \Omega \to \mathbb{R}$ a random variable on the probability space $[\Omega, \mathscr{A}, P]$ defined by:

$$X(\omega) = \begin{cases} 0 & for\ \omega = HHH, HHT, HTH, THH \\ 1 & for\ \omega = TTT, TTH, THT, HTT \end{cases}$$

   a. What is the name of this random variable?

   b. Determine the event $\{\omega \in \Omega\ ; X(\omega) \leq x\}$.

   c. Determine the distribution function F_X and draw it.

3. We rolling two identical fair dice (at the same time) and we take the summation of numbers of outcomes. Then:

   a. Determine the probability space of this random experiment.

   b. Let $X : \Omega \to \mathbb{R}$ be a map defined by the following relation:

$$X\left(\omega\right) = \begin{cases} -2 & \text{if } \omega\ =\ 9, 12 \\ -1 & \text{if } \omega \text{ is a prime number} \\ +1 & \text{otherwise} \end{cases}$$

      1. Prove that the given map X is a random variable on $[\Omega, \mathscr{A}, P]$.
      2. Represent this random variable tabular and graphical.
      3. Determine the distribution function of this random variable and sketch it.

4. Let X be a discrete random variable with probability mass function

$$P\left(X = k\right) = c\frac{k}{7},\ k = 2, 3, 4, 5.,\ \text{Then:}$$

   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).

e. Calculate E(3X-a) for any real number a.

5. We have a random experiment with a space of elementary events $\Omega = \{a, b, c, d, e, f\}$, and we take $\mathscr{A} = 2^{\Omega}$ and $P(A) = \dfrac{|A|}{|\Omega|}$. Now, if $X : \Omega \to \mathbb{R}$ is a map on Ω defined by the following relation:

$$X(\omega) = \begin{cases} 1 & \textit{for } \omega = a, e \\ 0 & \textit{otherwise} \end{cases}$$

a. Is this map a random variable on the probability space $[\Omega, \mathscr{A}, P]$? Why?

b. Determine $F_X(x)$ for all $x \in \mathbb{R}$.

c. Calculate the probability $P(-0.5 < X \leq 0.5)$.

6. Consider rolling a balanced die twice and let the random variable 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 (a).

c. Calculate the mean, variance, and standard deviation of the random variable X.

7. A discrete random variable has the following probability distribution.

k	0	1	2	3	4	5
$P(X = k)$	0.10	0.30	c	0.10	$c - 0.35$	0.05

a. Determine the value of c.

b. Represent this random variable tabular and graphical.

c. Determine the distribution function of this random variable and sketch it.

8. Consider a discreet random variable X with 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 distribution function F_X.

b. Draw the density and distribution function for this variable.

c. Calculate the mean, variance and standard deviation for the random variable X.

d. Calculate $\mathrm{E}(3X + 2)$ and $Var(3X - a)$ for any real number a.

9. Consider the **Bernoulli** random variable with parameter $1 > p > 0$ such that:

$$P(X = x) = \begin{cases} p & for\ x = 1 \\ q = 1 - p & for\ x = 0 \end{cases}$$

a. Make the graphical representation of this random variable for $p = 0.7$.

b. Determine the distribution function F_X and draw it for $p = 0.7$.

c. Calculate the mean, variance and standard deviation of this random variable.

10. A box contains 10 equally likely chips numbered from 1 to 10. Let random variable X be the total of two chips drawn at random and without replacement. Then determine the probability mass function and distribution function of X.

11. Let X be a random variable with probability mass function:

$$P(X = x) = c\left(\frac{4}{9}\right)^k \qquad ; k = 1, 2, 3$$

a. Determine the value of the constant c.

b. Determine the distribution function of X.

c. Calculate the mean and variance of X.

12. Let X be a random variable with probability mass function given by:

$$P(X = k) = \begin{cases} 0.1, & for \quad k = 4 \\ 0.3, & for \quad k = 5 \\ 0.3 & for \quad k = 6 \\ 0.2, & for \quad k = 8 \\ 0.1, & for \quad k = 9 \end{cases}$$

a. Represent this random variable graphically.

b. Calculate the probabilities $P(X \leq 6.5)$, $P\left(X > 6.1\right)$ and $P\left(5 < X < 8\right)$.

c. Determine the distribution function F_X .

d. Draw the graph of the distribution function F_X .

e. Calculate the mean for this random variable.

f. Calculate the variance for this random variable.

g. If we define new random variable $W = 7 - 4X$, then calculate $\mathrm{E}\left(W\right)$ and $Var\left(W\right)$.

13. Assume that the probability a baby born is a girl in a maternity hospital, is 0.51 and let the random variable 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 birth are boys?

14. Suppose that $P(X = n) = \dfrac{1}{n(n+1)}$ for any integer $n \geq 1$. Then determine the distribution function F_X and calculate the mean $\mathrm{E}(X)$ (if it exists).

15. Let X be a random variable with the discrete distribution function:

$$F_X(x) = \begin{cases} 0 & \mathit{for} \quad x < 0 \\ 1-p & \mathit{for} \quad 0 \leq x < 1 \\ 1-p^2 & \mathit{for} \quad 1 \leq x < 2 \\ 1-p^3 & \mathit{for} \quad 2 \leq x < 3 \\ \vdots & \quad \vdots \qquad \vdots \end{cases}$$

a. Then determine the probability mass function $P(X = \bullet)$.

b. Calculate the mean, variance and the standard deviation of X.

16. Consider the **Hyper Geometric** random variable with parameter $N, M \text{ and } n$:

$$P(X = x) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \quad ; k = 0, 1, \cdots, N;$$

and realize $k \in \{\max(0, n+k-N), \cdots, \min(n, k)\}$

where $n = 1, 2, \ldots N$ and $M = 0, 1, \cdots, N \;; N \in \mathbb{N}$. Then calculate the mean, variance of the random variable.

17. Let X be a random variable 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.

18. Let X be a discrete uniformly distributed random variable with the following density function

$$P\left(X = x\right) = 0.2 \text{ for } x = 5,\ 6,\ 7,\ 8,\ 9.$$

a. Show that X has a true discrete density function.

b. Calculate the mean of X.

c. Calculate the variance and the standard deviation of X.

19. Consider a random variable X with density function:

$$f_X(x) = \begin{cases} 3\,c\,x^2 & \mathit{for} \quad 0 \leq x \leq 2 \\ 0 & \text{otherwise,} \end{cases}$$

where c is real constant. Then:

a. Calculate the value of the constant c.

b. Draw the graph of the density function F_X.

c. Determine the distribution function F_X.

d. Draw the graph of the distribution function for X.

e. If you compare the graph of F_X with the graph of f_X. What do you note?

20. Let X be a continuous random variable with **p.d.f.**

$$f_X(x) = \begin{cases} \frac{1}{2} + \frac{1}{4}(x-3) & \textit{for } 1 \le x < 3 \\ \frac{1}{2} - \frac{1}{4}(x-3) & \textit{for } 3 \le x \le 5 \\ 0 & \textit{otherwise} \end{cases}$$

Then:

a. Draw the graph of **p.d.f.**.

b. Determine the distribution function of X.

c. Calculate the mean and variance of X.

21. The life length of an electronic device (in hour) is represented by the random variable X. Previous experience has shown that:

$$P(X > x) = \left(1 - \frac{x}{99}\right)e^{-x/99} \qquad ; x \ge 0.$$

Then:

a. Determine the distribution function and probability density function of X.

b. Calculate $P(20 < X < 35)$.

22. Let the time for a student to finish the aptitude test of NCAHE (in hours) is a continuous random variable X with

$$f_X(x) = \begin{cases} k(x-1)(2-x) & \textit{for } 1 \le x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Then:

a. Calculate the value of the constant k.

b. Derive the distribution function F_X.

c. Calculate the mean and variance for X.

d. What is the probability that a student can finish the test in 90 minutes?

23. Determine which of the following is a distribution function:

$$F(x) = \begin{cases} \frac{1}{2} e^{x} & \text{for } x < 0 \\ 1 - \frac{3}{4} e^{-x} & \text{for } x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{x}{1+x} & \text{for } x \geq 0 \end{cases}$$

24. Let F_X be a distribution function of a continuous random variable X such that:

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sqrt{x} & \text{for } 0 < x \leq 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$

a. Determine the density function f_X

b. Calculate the mean and variance of X.

c. Calculate $P(0.125 < X < 0.25)$.

d. Calculate $E(3X + 20)$ and $Var(5X + 20)$.

25. A probability density function of a continuous random variable X is given by:

$$f_X(x) = \begin{cases} ax + b & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

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 $P(0.2 < X < 0.9)$.

26. Let X be a continuous uniform random variable X with parameter $a = 5$ and $b = 10$, that means it has probability density function:

$$f_X(x) = \begin{cases} 0.2 & \text{for } 5 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

a. Then calculate the mean and variance.

b. Calculate $E(10X + 12)$.

27. Let X be a continuous random variable with probability density function

$$f_X(x) = \begin{cases} \frac{c}{x^4} & \text{for } x > 1 \\ 0 & \text{otherwise} \end{cases}$$

Then:

a. Calculate the value of c.

b. Calculate $\mathrm{E}\left(3X+2\right)$.

c. Calculate $P(-3 < X < 7)$ and $P(X > 1000)$.

28. The probability density function for the inverse exponential random variable X with parameter $\lambda > 0$ is:

$$f_X(x) = \begin{cases} \dfrac{\lambda}{x^2} e^{-\frac{\lambda}{x}} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Then:

a. Find the distribution F_X and calculate the mean of X.

b. Draw the graph of the density and distribution for $\lambda = 1$.

29. We consider a tossing four identical fair coins at the same time. Now, let X be a random variable observe the number of heads that appears. Then:

a. Determine the space of elementary events of this random experiment.

b. Compare this result with the example 3.2.5. What do you notice?

c. Determine the set $\mathbf{X}$ of possible values of the random variable X.

d. Compare the set $\mathbf{X}$ for this random variable with the values set of X in the example 3.2.5. what do you notice?

e. Determine the probability mass function $P\left(X = \bullet\right)$.

f. Represent the random variable X tabular and graphical.

g. Is this random variable a famous random variable, and what is his name?

h. Determine the distribution function F_X and sketch it.

i. Calculate $\mathbf{E}(X)$, $\mathbf{E}(X^2)$, $\mathbf{var}(X)$ and the standard deviation of X.

30. Consider rolling two fair different dice one time only (and at the same time), and let X be a discrete random variable observe the minimum of two numbers that appears. Then:

a. Determine the set $\mathbf{X}$ of possible values of the random variable X.

b. Determine the probability space of this random experiment, and calculate values of the probability mass function $P\left(X = \bullet\right)$.

c. Represent the random variable X tabular and graphical.

d. Is this random variable a famous random variable, and what is his name?

e. Determine the distribution function F_X and sketch it.

f. Calculate $\mathbf{E}(X)$, $\mathbf{E}(X^2)$, $\mathbf{var}(X)$ and the standard deviation of X.

31. We consider the experiment of rolling two identical balanced dice at the same time and for one time only, and let X be a random variable observe the sum of the apparent numbers. Then:

a. Determine the probability space of this random experiment.

b. Compare these results of this experiment with the results of the previous exercise. What do you note?

c. Determine the set $\mathbf{X}$ of possible values of the random variable X.

d. Represent the random variable X tabular and graphical.

e. Compare the set $\mathbf{X}$ for this random variable with the values set of X in exercise (5). What do you notice?

f. Determine the probability mass function $P\left(X=\bullet\right)$.

g. Is this random variable a famous random variable, and what is his name?

h. Determine the distribution function F_X and sketch it.

32. It was found that, in maternity hospital in Saudi Arabia, the number of twin births is approximately 1 in 95. Let X be the random variable observes the number of births in that hospital until the first twins are born. Determine the probability mass function and the distribution function of the random variable X.

(Hint: The sum of infinite geometric series $\sum_{n=0}^{\infty} r^n = \dfrac{1}{1-r}$ for $\left|r\right| < 1$).

33. Let weights of sheep (in a cattle farm) be normal distributed with mean (μ) 45 Kg and standard deviation (σ) 5 Kg. We select a sheep randomly, then:

a. What is the probability that the sheep has weight greater than 48 Kg?

b. What is the probability that it has weight between 44.25 and 46.75 Kg?

34. The circumference of a certain type of palm trees in Saudi Arabia farms follows normal distribution with mean 100 cm and standard deviation 5. Calculate the proportion of trees with circumference less than 95 cm.

35. It was found that the cholesterol level in a population of 14-years school boys is approximately normal distribution with mean 160 mg/dL and standard deviation 25 mg/dL.

a. What is the form of the **p.d.f.** for the random variable 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?

36. Lifetimes of a certain type (brand) of cars tires is approximately normally distributed with mean 42500 miles and standard deviation 3200 miles.

a. What is the probability that the lifetime is greater than 50000 miles?

b. What range do the middle 90% of tire lifetimes?

c. What is the probability that $P(X < 3200)$?

d. For a population consists of 2000 tires, what is the number of tires whose lifetimes between 30000 miles and 45000 miles.

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.100,000. 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 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

EXERCISES

Section 4.1

DEFINITIONS AND CONCEPTS

In the first chapter we had presented some basic concepts in statistics. For example: Descriptive Statistics, Inferential Statistics, Population, Sample, Parameter, Statistic, Variables and etc....

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.

A set representing all the elements of the Population under study will be marked with Ω also, and in this study of statistics one say that Ω is a **sample space**.

Several times in this book 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. If we select 50 students at random and 27 of them are female, then a natural estimate is $\frac{27}{50} = 0.54$ or 54%. The question that arises here is: How much confidence we can place in this estimate?

The answer to this question 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.

Now 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. But 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 of this chapter we assume that all samples are either random or at least represent the population well.

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 which have the same size, it will have the same probability of selection.

REMARK 4.1.1

Let Ω be a sample space of a given population described by a random variable X. Then one denotes a simple random sample (mathematically) with a random vector of form:

$$\mathcal{X} = (X_1, X_2, \ldots, X_n)$$

Where all random variables $X_1, X_2, \ldots$ and X_n independent of each other, as well as all of them have the same distribution of the random variable X. For example, if the population is normal with mean μ and standard deviation σ, then the random variable X will have a normal distribution $N(\mu, \sigma^2)$, so all random variables $X_1, X_2, \ldots and\ X_n$ will have the normal distribution $N(\mu, \sigma^2)$ also.

DEFINITION 4.1.1 (Estimator)

An estimator is a statistic (function of the random sample) whose value depends on the particular sample is drawn of a population.

One expresses it mathematically as follows:

Let Ω be a sample space, and $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ is a simple random sample of Ω. Then a statistic T is a function g on the sample $\mathcal{X}$, that means:

$$T := g(\mathcal{X}) := g(X_1, X_2, \ldots, X_n)$$

For example, if we have $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ a simple random sample of a population Ω. Then:

a. $g(\mathcal{X}) = g(X_1, X_2, \ldots, X_n) = \dfrac{X_1 + X_2 + \ldots + X_n}{n} = \bar{X}$ is an estimator, and it is called the mean of the sample $\mathcal{X}$ (one denotes the value of $\bar{X}$ by $\bar{x}$).

b. $g(\mathcal{X}) = g(X_1, X_2, \ldots, X_n) = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S^2$ is an estimator, and it is called the variance of the sample $\mathcal{X}$ (one denotes the value of S^2 by s^2).

REMARKS 4.1.2

1. The value of the estimator is called the estimated value and it is used to predict the value of a population parameter.

2. There are many ways to determine an estimator of a parameter 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 μ of the normal population has the statistic $\overline{X}$ as an estimator.

b. The variance σ^2 of the normal population has the statistic S^2 as an estimator.

c. The proportion p of the Bernoulli population has the statistic $\overline{X}$ as an estimator. For this case $\overline{X}$ represents the relative frequency of successes in the sample $\mathscr{X}$. Therefore, in this special case one denote $\overline{X}$ by $\hat{P}$, and we will denote the value of $\hat{P}$ by the symbol $\hat{p}$. Note that, if we have k successes in the sample $\mathscr{X} = (X_1, X_2, \dots, X_n)$, then we get that $\hat{p} = \dfrac{k}{n}$.

The following chart shows the steps required to draw the statistical inference (estimation) of an unknown parameter.

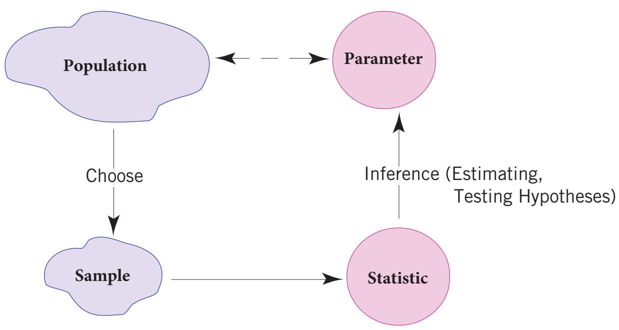


Figure 4.1.1

> **DEFINITION 4.1.2 (Standard Normal Probabilities)**
> Let $0 < \alpha < 1$ be a real number. Then the standard normal probability (z-value or z-score), and it symbolizes by z_α, it is that value on the real axis, for which is the area between the left part of the curve of $\varphi(x)$ and the straight $x = z_\alpha$ equal to α. Where $\varphi(x)$ is the density function of standard normal distribution.

▶ **EXAMPLE 4.1.1** We consider $z_\alpha = 0.57$. Then the area between the left part of the curve of $\varphi(x)$ and the straight $x = z_\alpha$ equal to shaded area in the following figure.

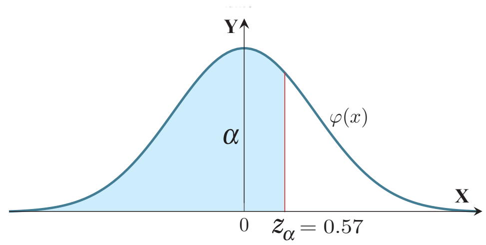


Figure 4.1.2

This shaded area is $P(Z < z_\alpha) = \Phi(z_\alpha) = \Phi(0.57)$, and by using the table of standard normal distribution:

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

We find $\Phi(0.57) = 0.7157$.

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 α in the two tails of the standard distribution 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 variable.

The z-value that cuts off a left tail of area α is denoted z_α, therefore the z-value that cuts off a right tail of area α is $z_{1-\alpha}$. Here we note that $z_\alpha = -z_{1-\alpha}$.

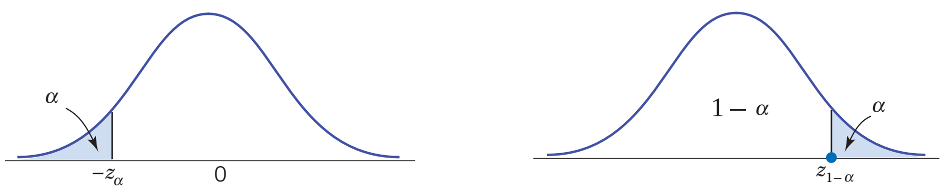


Figure 4.1.3-a (Areas under the one-tail standard normal density function)

Thus the numbers ∓ 1.96 in the example are $\mp z_{1-(\alpha/2)} = \mp z_{1-0.025} = \mp z_{0.975}$, which is for $\alpha = 1 - 0.95 = 0.05$.

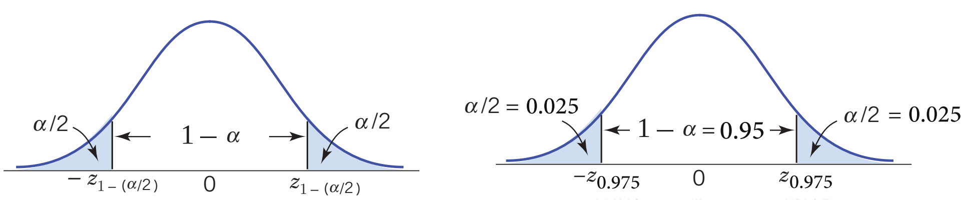


Figure 4.1.3-b (Areas under the two-tail standard normal density function)

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:

1. It gives the probability of getting a particular $\bar{X}$ given the mean μ and the standard deviation σ 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 μ and one can use the sample proportion to estimate the population proportion.

It is known that $\bar{X}$ is random variable, therefore, it has a probability distribution. This distribution is called the sampling distribution of $\bar{X}$.

Next, we present the following theorem.

THEOREM 4.1.1 (The Central Limit Theorem)

Let Ω be the sample space of a statistical population, which described by a random variable X with mean μ and standard deviation σ, and $\mathscr{X} = (X_1, X_2, \ldots, X_n)$ is a simple random sample of Ω. Then for sufficiently large sample size n, the sampling distribution of $\bar{X}$ follows normal distribution with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$ (it is called standard error also).

▶ **EXAMPLE 4.1.2** We suppose that the number of defects in newly manufactured bulbs in each working shift is a normal distributed random variable X with a mean $\mu = 3$ and standard deviation $\sigma = 2.5$ (here μ and σ are known). Now, consider that in one of the working shifts a sample of 225 new bulbs were tested. Then:

a. Determine the sampling distribution of $\bar{X}$ based on samples of size 225.

b. Calculate the mean and the standard deviation of $\bar{X}$.

c. Calculate 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) The mean of $\bar{X}$ is $\mu_{\bar{X}} = \mu = 3$ and the standard deviation (standard error) is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{225}} = 0.17$$

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 $\hat{P}$

We suppose that the Stat101 course has 200 Saudi (S) and Non-Saudi (NS), and the proportion p (with $0 < p < 1$) of NS is not known. Here we note that, we have a Bernoulli population with parameter p (unknown). Now assume 20 students were randomly selected from the section with replacement (one by one), and 6 students out of the 20 students were NS. In this sample, the proportion of NS is $6/20 = 0.3$. This quantity is known as the sample proportion.

Next, one is interested to find the sample distribution of the proportion $\hat{P}$. One may consider a Bernoulli random variable such that $X = 1$ if for NS and $X = 0$ for *S*. Thus one may view $\hat{P}$ as the mean of 20 *X*'s. Thus by the central limit theorem and for large *n*, we get that $\hat{P}$ is normally distributed with mean $\mu_{\hat{P}} = p$ and standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$.

The previous results are more accurate when we have $n \geq 30$ and $np \geq 5$, $n(1-p) \geq 5$. If this condition is not satisfied, then the proportion p has unknown distribution (related to binomial distribution) with the same mean and variance.

REMARK 4.1.3

Let $\mathcal{X} = (X_1, X_2, ..., X_n)$ be a simple random sample of a Bernoulli population with sample space Ω, and Suppose that k of $X_1, X_2, ..., X_n$ 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 $p = 0.20$, then calculate the following:

a. The sampling distribution of $\hat{P}$ based on 225 observations,

b. The mean and the standard error of $\hat{P}$.

c. The probability that $\hat{P} < 0.25$

Solution: We have:

For a) According to the central limit theorem we note that, the sampling distribution of $\hat{P}$ is approximate to the normal distribution.

For b) According to the central limit theorem we have:

$$\mu_{\hat{P}} = p = 0.20$$

and:

$$\sigma_{\hat{P}} = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{0.20\,\,(1-0.20)}{225}} = 0.0267$$

For c) The probability:

$$P\,(\hat{P} < 0.25) = P\left(\underbrace{\frac{\hat{P} - \mu_{\hat{P}}}{\sigma_{\hat{P}}}}_{Z} < \underbrace{\frac{0.25 - 0.20}{0.0267}}_{z=1.87} \right)$$

By using the z-table, one has $P\,(\hat{P} < 0.25) = P\,(Z < 1.87) = 0.9693$.

Section 4.2

ESTIMATION OF THE POPULATION MEAN

DEFINITION 4.2.1 (Point Estimation)

The point estimation is an estimate of the population parameter by a single number. This single number is a value of a suitable statistic for the population parameter.

EXAMPLES 4.2.1 Let Ω be a sample space of the men population in a certain city, and $\mathcal{X} = (X_1, X_2, ..., X_n)$ is a simple random sample of Ω. The following data represents the hemoglobin level in (g/dl -gram/deciliter) 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

Determine the point estimation of the overall mean of the hemoglobin level for men in the city.

Solution: We know that $\bar{X}$ is an estimator for the mean of population μ, therefore we have:

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{1}{50}(17.0 + 17.7 + 15.9 + ... + 17.3 + 15.8) = 16$$

As the estimated value of the mean of hemoglobin level for the men in this city. So we have:

$$\hat{\mu} = \bar{x} = 16 \text{ g/dl}$$

◀

When we estimate an unknown parameter by a value of a point estimator, we do not expect the resulting value to exactly equal the parameter (mean or standard deviation or ...), but we expect that it will be “close” to it. To be more specific, we sometimes try to find an interval about the value of 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 parameter (mean or standard deviation or ...).

DEFINITION 4.2.2 (Interval Estimation)

The interval estimation of a population parameter is a determination of an interval that is predicted to contain the value of the population parameter with certain probability.

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 for the mean 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 determined. Whether the confidence 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 interested population, and a confidence interval were computed for each sample. Then a 95% confidence level implies that 95% of the confidence intervals, that are determined, 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 Ω the sample space of a normal population with mean μ (is unknown) and standard deviation σ (is known and equal to σ). Now let's take $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ a simple random sample, and suppose we want to utilize this sample to obtain a $100(1-\alpha)\%$ confidence interval for the population mean μ. To obtain such an interval, we start with the sample mean $\overline{X}$, which is the point estimator of the population mean μ.

We now make use of the sampling distribution of the statistic $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ and the probability statement as shown in below:

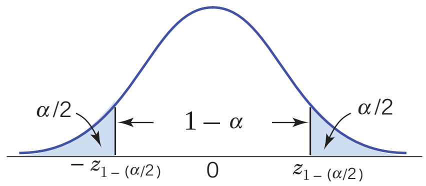


Figure 4.2.1

$$P\left(-z_{1-(\alpha/2)} < \underbrace{\frac{\bar{X}-\mu}{\sigma \,/\, \sqrt{n}}}_{Z} < z_{1-(\alpha/2)}\right) = P\left(-z_{1-(\alpha/2)} < Z < z_{1-(\alpha/2)}\right) = 1-\alpha$$

Then a $100(1-\alpha)\%$ confidence interval for the population mean μ in this case is given by the following interval:

$$\boxed{\left(\bar{x} - z_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}}\right)}$$

Calculate $\bar{x}$ the value of the statistic $\bar{X}$ from the sample, so we become a $100(1-\alpha)\%$ confidence interval for the population mean μ by the interval in which its bound values are:

$$\bar{x} \mp z_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}}$$

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, but we will not deal with such a study.

The following table summarizes different cases.

Table 4.2.1

Case	Population	Sample size	Standard deviation	Confidence interval
1	normal	any sample size	known	$\bar{x} \mp z_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}}$
2	any population	large $(n \geq 30)$	known	$\bar{x} \mp z_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}}$
3	any population	large $(n \geq 30)$	unknown but can by replaced by the sample standard deviation s	$\bar{x} \mp z_{1-(\alpha/2)}\frac{s}{\sqrt{n}}$

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 δ_μ) that is:

$$\delta_\mu = z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} \quad \textbf{or} \quad \delta_\mu = z_{1-(\alpha/2)} \frac{s}{\sqrt{n}}$$

This means that if we use the value $\overline{x}$ instead of μ, we will make a maximum error of $\pm\delta_\mu$ 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:

a. Determine 99% confidence interval of the mean μ drying time of this brand of paint.

b. Determine The marginal error δ_μ.

Solution:

For a) We have the confidence level 99%, therefore $\alpha = 1 - 0.99 = 0.01$, and then $\alpha / 2 = 0.005$. So is $z_{1-(\alpha/2)} = z_{0.995}$, but we find in z-Table $z_{0.9949} = 2.57$ and $z_{0.9951} = 2.58$, therefore, we take the average of the previous two values for value $z_{0.995}$, so we have:

$$z_{0.995} = \frac{2.57 + 2.58}{2} = 2.575$$

Now, from the given sample, we have $n = 15$ and $\overline{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = 3.8$ hours, and because our population is normal with known standard deviation $\sigma = 0.96$, we find that 99% confidence interval for μ is:

$$\overline{x} \mp z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} = 3.8 \mp (2.575)\frac{0.96}{\sqrt{15}} = 3.8 \mp 0.638$$

then, 99% confidence interval for μ 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 δ_μ is given by:

$$\delta_\mu = \mp z_{1-(\alpha/2)} \frac{\sigma}{\sqrt{n}} = \mp(2.575)\frac{0.96}{\sqrt{15}} = \mp 0.638 \text{ hours}$$

◀

▶ **EXAMPLE 4.2.3** A random sample of 120 students from a large university yields mean GPA (Grade Point Average) 2.71 with standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university.

Solution: For confidence level 90%, we have $\alpha = 1 - 0.9 = 0.10$, therefore we get that $\alpha / 2 = 0.05$, so $z_{1-(\alpha/2)} = z_{0.95}$. From Z-Table, we have $Z_{0.95} = \dfrac{1.64 + 1.65}{2} = 1.645$.

Since our sample is large ($n = 120$) and the standard deviation is known to be $\sigma = 0.51$, we use the following formula for determination the 90% confidence interval:

$$\bar{x} \mp z_{1-(\alpha/2)} \frac{s}{\sqrt{n}} = 2.71 \mp (1.645)\frac{0.51}{\sqrt{120}} = 2.71 \mp 0.0766$$

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 recoded. 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 μ for all cars of this model.

Solution: For confidence level 95%, we have $\alpha = 1 - 0.95 = 0.05$, therefore we get $\alpha / 2 = 0.025$, so $z_{1-(\alpha/2)} = z_{0.975}$. From z-Table, we have $z_{0.975} = 1.96$. Since our sample is large ($n = 36$) and the standard deviation for the population is unknown, we must use the standard deviation for the sample $s = 3$ instead of the standard deviation σ of population. Therefore, we get that the 95% confidence interval for the mean mileage of this model of cars as follow:

$$\bar{x} \mp z_{1-(\alpha/2)} \frac{s}{\sqrt{n}} = 18 \mp (1.96)\frac{3}{\sqrt{36}} = 18 \mp 0.98 = (17.02, 18.98)$$

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. Here it is worth noting that the answers to the questions: "How many subjects do I need?" and "How many subjects can I afford to get?" are linked to the companion questions. In fact, 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 with certain confidence. Therefore, for confidence level $1-\alpha$ and by using the relationship between the sample size and the margin error δ_μ we get:

$$\boxed{n = \left(\frac{\sigma\, z_{1-(\alpha/2)}}{\delta_\mu}\right)^2}$$

where δ_μ is the desired margin error, σ 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 $\delta_\mu = 0.5$. Therefore, the required sample size is given by:

$$n = \left(\frac{\sigma\, z_{1-(\alpha/2)}}{\delta_\mu}\right)^2 = \left(\frac{(0.96) z_{0.975}}{0.5}\right)^2 = \left(\frac{(0.96)(1.96)}{0.5}\right)^2 = 14.162 \approx 15 \text{ records.}$$

◀

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 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 the parameter p (the parameter of the Bernoulli population). Suppose that the researcher has collected a random sample of size $n = 1000$ Saudi adult and found $k = 320$ diabetics among the sample. Then for this sample one gets that the estimated value of the proportion of population is $\frac{k}{n} = \frac{320}{1000} = 32\%$.

EXAMPLE 4.3.1 The 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 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 approximately 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 (as we did for normal population). 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 is randomly selected Saudi adult,
- Success (it denoted by S) is the selected adult has diabetes,

- Failure (it denoted by F) is the selected adult has not diabetes,
- p is the proportion of the adult who have diabetic disease,
- n is the number of the independent Bernoulli trails conducted in the study, that is the number of persons in the sample,
- k is the number of success, that is the number of diabetic persons in the sample.

For a simple random sample $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ of a Bernoulli population Ω (that means $X_1, X_2, \ldots$ and X_n are independent Bernoulli random variables), we now for n sufficiently large, then the statistic $\hat{P} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ is approximately normal distributed with mean $n\,p$ and the standard deviation $\sqrt{n\,p(1-p)}$. Therefore, the following statistic is approximately standard normal distributed:

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)\,/\,n}}$$

Now, when n is sufficiently large, then a $100(1-\alpha)\%$ confidence interval for the population proportion p given by the following interval:

$$\boxed{\left(\hat{p} - z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = (\hat{p} - \delta_p\ ,\ \hat{p} + \delta_p)}$$

This means that if we use the value $\hat{p}$ instead of p, we will make a maximum error of $\pm\delta_p$ with probability equal to $1-\alpha$.

DETERMINING OF MARGIN OF ERROR

The quantity $z_{1-(\alpha/2)}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ is called the margin error (or the maximum error) of the population proportion estimation with $100(1-\alpha)\%$ confidence level, and we denoted it by δ_p. That means:

$$\boxed{\delta_p = z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

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. Then:

a. Determine the point estimation of the proportion of the diabetic Saudi adults.

b. Determine a 90% confidence interval for p.

c. If the point estimation is used to estimate p, determine the margin error of the estimate with 90% confidence.

d. 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) We have $k = 320$. Then the estimated value 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%, then we get $\alpha = 1 - 0.90 = 0.10$, then $\frac{\alpha}{2} = 0.05$, so $z_{\alpha/2} = z_{0.95} = 1.645$. More than that we have $n = 1000$ and $k = 320$. So we get $\hat{p} = 0.32$. Therefore, 90% confidence interval is determined by the following relation:

$$\hat{p} \mp z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.32 \mp 1.645\sqrt{\frac{0.32(0.68)}{1000}} = 0.32 \mp \ 0.02427$$

So we have $(0.296, 0.344)$ is the 90% confidence interval for p.

For c) The margin error when we use the point estimation for the population proportion in this case is δ_p=0.02427

For d) From part (b), the researcher can be 90% confidence that the percentage of the diabetic's Saudi adults is between 29.6% and 34.4%. 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%.

◀

▶ **EXAMPLE 4.3.3** A random sample of 125 individuals working in a large city indicated that 42 are dissatisfied with their working conditions. Determine a 95% lower confidence bound 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 95% lower bound is given by:

$$\hat{p} - z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.336 - 1.96\sqrt{\frac{0.336(1-0.336)}{125}} = 0.2532$$

That is, we can be 95% certain that over 25.32% 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 $\alpha = 0.01$. Therefore, we have:

$$z_{1-(\alpha/2)} = z_{0.995} = \frac{2.57 + 2.58}{2} = 2.575$$

On the other hand, we have $k = 82$. Therefore, we get:

$$\hat{p} = \frac{82}{100} = 0.82 \Rightarrow 1 - \hat{p} = 1 - 0.82 = 0.18$$

Then 99% confidence interval estimate of the population proportion p is given by the following relation:

$$\hat{p} \mp z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.82 \mp 2.575\sqrt{\frac{0.82(0.18)}{100}} = 0.82 \mp 0.0989 = \left(0.721\ , 0.919\right)$$

Therefore, the 99% confidence interval estimate of the population proportion of all the students at the university who are nonsmokers is $\left(0.721\ , 0.919\right)$. 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 ± 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 the following relation:

$$\hat{p} \mp z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \mp 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

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 of p (the proportion of the population in favor) is:

$$0.46 \mp 1.96\sqrt{\frac{(0.46)(1-0.46)}{n}}.$$

Since the margin of error is ± 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 we have:

$$n = (1.96)^2\frac{(0.46)(0.54)}{(0.03)^2} = 1060.3$$

That is, approximately 1060 people were sampled, and 46% (nearly 488 persons) were in favor of President Bush's handling of the economy.

SAMPLE SIZE DETERMINATION WHEN ESTIMATING p

From the previous example, we noticed that, it can be determining the sample size for estimating the population proportion by the meaning of the maximum error, where:

$$\hat{p} \mp z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \mp \delta_p$$

Then:

$$\delta_p = z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Reformulate this relation; we get the sample size in terms of some specific value of the maximum error δ_p as:

$$\boxed{n = \left(\frac{z_{1-(\alpha/2)}}{\delta_p}\right)^2\hat{p}(1-\hat{p})}$$

Sample size should be an integer value, otherwise we approximate it to the bigger integer.

Now, when the information about the value of $\hat{p}(1-\hat{p})$ are not known, then we use the following relation for the sample size n:

$$\boxed{n = 0.25\left(\frac{z_{1-(\alpha/2)}}{\delta_p}\right)^2}$$

Because we know that $\hat{p}(1-\hat{p})$ has maximum value when $\hat{p}(1-\hat{p}) = 1/4$, this is because if we take the function $f(\hat{p}) := \hat{p}(1-\hat{p})$, then for $f'(\hat{p}) = 0 \Rightarrow \hat{p} = 1/2$.

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:

$$n = 0.25\left(\frac{z_{1-(\alpha/2)}}{\delta_p}\right)^2 = 0.25\left(\frac{1.96}{0.01}\right)^2 = 9604$$

That is, the sample size needs to be 9604 to ensure with 95% confidence that the maximum error will not exceed 0.01.

For a constant value of α we can see as the error δ_p decreases, the sample size n increases.

For a constant value of δ_p we can see as the value of α decreases, the sample size n increases also.

Some common sample sizes in practice are summarized below.

Table 4.3.1

δ_p	α	n
0.05	0.05	385
0.04	0.05	601
0.03	0.05	1068
0.01	0.05	9605

δ_p	α	n
0.05	0.01	666
0.04	0.01	1041
0.03	0.01	1849
0.01	0.01	16641

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:

$$\boxed{\left(\hat{p} + z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) - \left(\hat{p} - z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 2\,z_{1-(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

It can be shown that the product $\hat{p}(1-\hat{p})$ is always less than or equal to 0.25, 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 following statement:

$$\boxed{\text{Length of 100(1-}\alpha\text{)\% confidence interval} \le \frac{z_{1-(\alpha/2)}}{\sqrt{n}}}$$

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 $\frac{z_{1-(\alpha/2)}}{\sqrt{n}} = L$ will be sufficient.

That is n must be chosen n so that $\sqrt{n} = \frac{z_{1-(\alpha/2)}}{L}$. Upon squaring both sides, we see that n must be such that:

$$\boxed{n = \left(\frac{z_{1-(\alpha/2)}}{L}\right)^2}$$

▶ **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:

$$n = \left(\frac{z_{1-(\alpha/2)}}{L}\right)^2 = \left(\frac{1.645}{0.03}\right)^2 = 3006.69 \approx 3007$$

This show that the company need at least 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. δ_p is given by 0.05. For the 95% confidence interval, we see that $z_{1-(\alpha/2)} = z_{0.975} = 1.96$ and thus:

$$n = \left(1.96\sqrt{0.5(1-0.5)}\ /\ 0.05\right)^2 = 384.16$$

For the 99% confidence interval we use $z_{1-(\alpha/2)} = z_{0.995} = 2.575$ and get:

$$n = \left(2.575\sqrt{0.5\,(1-0.5)}\ /\ 0.05\right)^2 = 663.06$$

Section 4.4

INTRODUCTION TO HYPOTHESES TESTING

One of a statistician's most important jobs is to draw inference about the population based on the samples taken from the same population. Most of the statistical inference centers around the population parameters, for example, the mean or proportion of the population. 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.

Now, 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.

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 an issue (or statement) related to the statistical population, based on insufficient evidence from the sample 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“ 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 we 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 μ 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 $H_0 : \mu = 47$ mpg.

For b) From experience with this manufacturer, we have every reason to believe that the advertised mileage is too high. If μ 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 statistic whose calculated value from the given sample is used to make a decision on the hypothesis test.

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	$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$
2	any population	large ($n \geq 30$)	known	$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$
3	any population	large ($n \geq 30$)	unknown, use the sample standard deviation instead of	$Z = \dfrac{\bar{X} - \mu}{s / \sqrt{n}} \sim N(0,1)$

ERRORS AND LEVEL OF SIGNIFICANCE

The error of first type:

If the null hypothesis is true, and based on the sample data we reject this hypothesis, then we have made an error. This error is called " error of first type (Type I error). The probability that we make this error is denoted by α and is called "the significance level" of the test.

The error of second type:

If the null hypothesis is false, and based on the sample data we accept this hypothesis, then we have made an error. This error is called "error of second type (Type II error). The probability that we make this error is denoted by β. The quantity $1-\beta$ is called "the power of the test".

Table 4.4.2

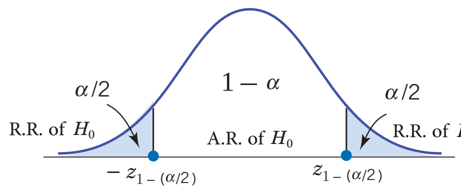
		Actual Situation	
		H_0 is true	H_0 is false
Decision	Do not reject H_0	Correct decision	Type II error $P(\text{Type II error}) = \beta$
	Reject H_0	Type I error $P(\text{Type I error}) = \alpha$	Correct decision

DEFINITION 4.4.5 (The Critical Regions and values)

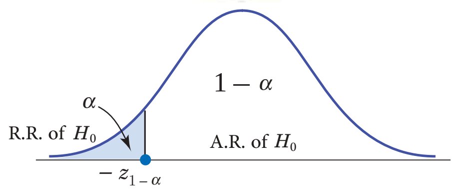
- The critical region is a region that produced by the value(s) that corresponds to the rejection of the null hypothesis at some chosen level of significance.
- The value(s) that determine the critical region(s) is called a critical value(s). The critical values in our study are $z_{1-\alpha}$, $z_{1-(\alpha/2)}$, $-z_{1-\alpha}$ or $-z_{1-(\alpha/2)}$.

The shaded area under the standard normal distribution curve is equal to the level of significance. For a specific value of the significance level, one calculate the critical values from the standard normal distribution table (this calculated values are known as the tabulated value also). If the absolute value of the statistic is larger than the tabulated value, then statistic value is in the critical region.

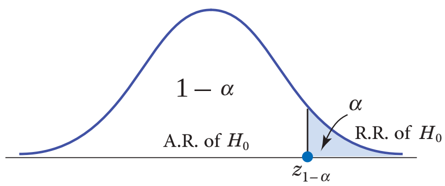
The statistical tests use one tailed or two tailed critical regions depending on the nature of the null hypothesis and the alternative hypothesis. The possible critical regions are the parts of the real axis under the shaded areas (see figure (4.4.1) below):



Two – tailed critical region



Left – tailed critical region



Right – tailed critical region

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 Ω as a sample space of a population with mean μ and standard deviation σ, and $\mathcal{X} = (X_1, X_2, \ldots, X_n)$ is a simple random sample of Ω.

In this section, we will study 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 $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$, because $\overline{X}$ is $N(\mu, \sigma^2 / n)$ distributed, while for the case (iii), we use $Z = \dfrac{\overline{X} - \mu}{s / \sqrt{n}} \sim N(0,1)$ approximately (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 μ. 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:

$$H_0 : \mu = \text{specified value}(\mu_0) \quad \text{versus} \quad H_1 : \begin{cases} \mu > \text{specified value}(\mu_0), \\ \mu < \text{spiecified value}(\mu_0), \\ \mu \neq \text{specified value}(\mu_0). \end{cases}$$

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:

$$z_0 = \begin{cases} \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} & ; \text{when } \sigma \text{ known} \\[2em] \dfrac{\overline{x} - \mu_0}{s\ /\ \sqrt{n}} & ; \text{when } \sigma \text{ unknown and } n \geq 30 \end{cases}$$

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 α. This is because the critical region is chosen in such way that the probability will be α 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 μ is different from the stated value μ_0 (pre-specified value of the mean).

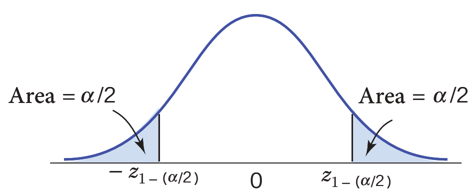


Figure 4.5.1-a

ii. right-tailed test $(H_1 : \mu > \mu_0)$

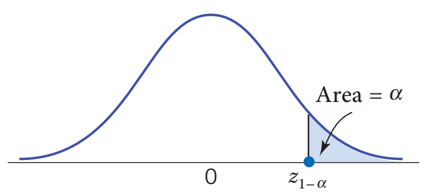


Figure 4.5.1-b

iii. left-tailed test $(H_1 : \mu < \mu_0)$

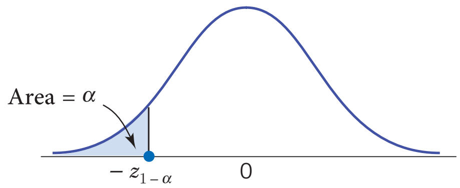


Figure 4.5.1.c

4. **Decision:**

If the test statistic value falls in the critical region, we reject H_0. When this occurs, some statisticians say the results are statistically significant, also giving the α 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. This means that accepting the null hypothesis does not mean that it is necessarily true, but that we do not have enough evidence from the sample data to reject it. The rejection of the null hypothesis is a declaration that it is falls.

p -VALUE 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 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:

$$H_0 : \mu = \text{specified value}(\mu_0) \quad \text{versus} \quad H_1 : \begin{cases} \mu > \text{specified value}(\mu_0), \\ \mu < \text{specified value}(\mu_0), \\ \mu \neq \text{specified value}(\mu_0). \end{cases}$$

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:

$$z_0 = \begin{cases} \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} & ;\text{ when } \sigma \text{ known} \\[2ex] \dfrac{\overline{x} - \mu_0}{s\,/\,\sqrt{n}} & ;\text{ when } \sigma \text{ unknown and } n \geq 30 \end{cases}$$

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-value $= 2\,P\,(Z >|\ z_0\ |)$

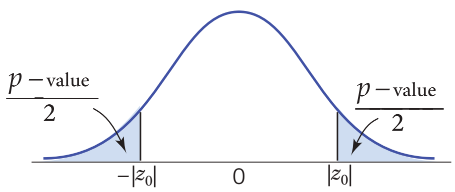


Figure 4.5.2-a

   b. For right-tailed test $(H_1 : \mu > \mu_0)$ we have: p-value $= P(Z > z_0)$

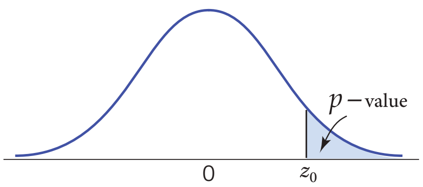


Figure 4.5.2-b

   c. For left-tailed test $(H_1 : \mu < \mu_0)$ we have: p-value $= P\,(Z < -\mid z_0 \mid)$

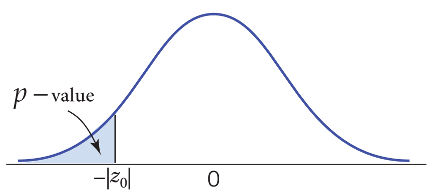


Figure 4.5.2-c

4. **Decision:**

   The decision rule is to reject H_0 when p-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 $\overline{x} = 22.6$ 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:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \Rightarrow z_0 = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{22.6 - 20}{2.5 / \sqrt{49}} = 7.28$$

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):

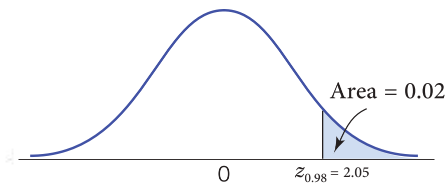


Figure 4.5.3

4. **Decision:**

   The test statistic value falls in the critical region, so we reject H_0 at the significance level of α. Therefore, 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 $(H_1 : \mu > 20)$, therefore, we have:

$$p\text{-value} = P(Z > z_0) = 1 - P(Z \leq 7.28) \approx 1 - 1 = 0 < \alpha = 0.02$$

So we reject H_0. There is sufficient evidence to conclude the typical amount spent per customer is more than \$20 at the significance level $\alpha = 0.02$.

◀

▶ EXAMPLE 4.5.2 Suppose an editor of a publishing company claims that the mean time to write a textbook is less than 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 13.5

months. Assume also that the standard deviation is known to be 3.6 months and the time to write a textbook is normally distributed. Then by using a 0.025 level of significance, would you conclude the editor's claim is true?

Solution: In this example, there population is normal with known standard deviation $\sigma = 3.6$. The sample has a size of 16 with average $\overline{x} = 13.5$ months. Then, we follow the usual test steps.

1. **Identifying the null hypothesis H_0 and the alternate hypothesis H_1:**

$$H_0 : \mu = 15 \quad versus \quad H_1 : \mu < 15$$

2. **Selecting the test statistic and determine its value under the null hypothesis H_0:**

   In this case, we use the following value for the test statistic $Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$:

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{13.5 - 15}{3.6 / \sqrt{16}} = -1.67$$

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):

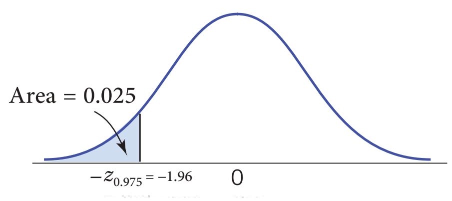


Figure 4.5.4

4. **Decision:**

   The test statistic value falls in the acceptance region, so we don't reject H_0 at the significance level of $\alpha = 0.025$. (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 not true at $\alpha = 0.025$.

By using the p-value approach, we have a left-tailed test $(H_1 : \mu < 15)$, then we have:

$$p\text{-value} = P(Z < -|\, z_0 \,|) = P(Z < -1.67) = 0.0475 > \alpha = 0.025$$

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 the significance level $\alpha = 0.05$?

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 $\overline{x} = 3.2$ sets. Then, we follow the usual test steps.

1. **Identifying the null hypothesis H_0 and the alternate hypothesis H_1:**

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu \neq 3.$$

2. **Selecting the test statistic and determine its value under the null hypothesis H_0:**

   In this case, we use the following value for the test statistic $Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$:

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{3.2 - 3}{1 / \sqrt{100}} = 2.0$$

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):

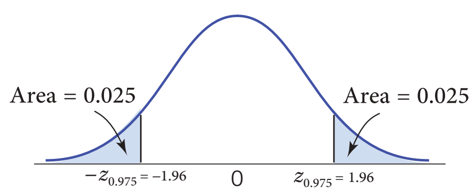


Figure 4.5.5

4. **Decision:**

   The test statistic value falls in the critical region, so we reject H_0 at the significance level of $\alpha = 0.05$. Then, we do have enough evidence to conclude that the average number of TV sets in U.S. homes differs from three at the significance level $\alpha = 0.05$.

By using the p-value approach, we have a two-tailed test $H_1 : \mu \neq 3$, then we have:

$$p\text{-value} = 2\ P(Z > |\ z_0\ |) = 2[1 - P(Z \leq 2.0)] = 2(1\text{-}0.9772) = 0.0456 < \alpha = 0.05$$

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 the significance level $\alpha = 0.05$.

◀

▶ 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 $\overline{x} = 1.170$ and $s = 0.111$. Now, by using $\alpha = 0.05$ significance level, do the data indicate that the mean bran content of all boxes of this brand of cereal differs from 1.2 ounces?

Solution

The population in this example need not to be normal since the sample size in 60. The sample mean and standard deviation are $\overline{x} = 1.17$ 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 following value for the test statistic $Z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$:

$$z_0 = \frac{\overline{x} - \mu_0}{s / \sqrt{n}} = \frac{1.17 - 1.2}{0.111 / \sqrt{60}} = -2.094$$

3. **Determine the critical region:**

The test is 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 determined by $\pm z_{1-(\alpha/2)} = \pm z_{0.975} = \pm 1.96$ (see below):

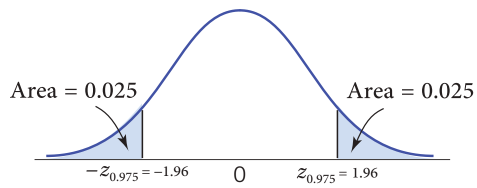


Figure 4.5.6

4. **Decision:**

The value of test statistic falls in the critical region, so we reject H_0 at the significance level of $\alpha = 0.05$. 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 at the significance level $\alpha = 0.05$.

One may use the p-value approach, since the test is two-tailed ($H_1 : \mu \neq 1.2$), then:

$$p - \text{value} = 2\,P(Z > | z_0 |)] = 2[1 - P(Z \leq 2.094)] = 2(1\text{-}0.9817) = 0.0366 < \alpha = 0.05$$

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 p the parameter (unknown) of the population proportion. 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:

$$H_0 : p = \text{specified value}(p_0) \quad \text{versus} \quad H_1 : \begin{cases} p > \text{specified value}(p_0), \\ p < \text{specified value}(p_0), \\ p \neq \text{specified value}(p_0). \end{cases}$$

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:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$$

   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.

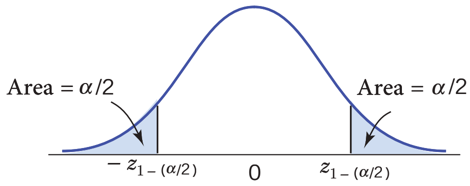


Figure 4.6.1-a

ii. right-tailed test $(H_1 : p > p_0)$

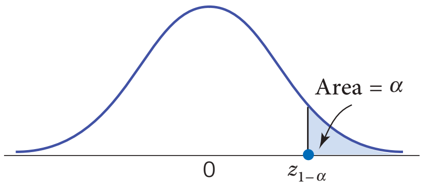


Figure 4.6.1-b

iii. left-tailed test $(H_1 : p < p_0)$

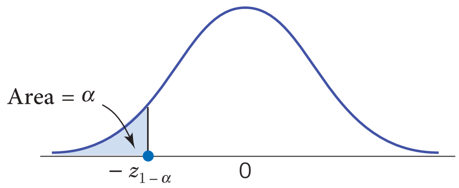


Figure 4.6.1-c

4. **Decision:**

   Again as the population mean hypotheses testing, if the test statistic value 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 PROPORTION

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** The quality control inspector claims that more than 2 percent of the produced computer chips produced by a certain computer company are defective. An electronics company, impressed by that claim, has purchased a large quantity of chips. To determine if the inspector'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 $\alpha = 0.05$ level of significance) the inspector's claim?

Solution: In this example we have:

$$p_0 = 0.02, \quad n = 400, \quad \hat{p} = \frac{13}{400} = 0.0325, \quad \alpha = 0.05$$

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 test statistic value z_0 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, then the critical region is the part of the real axis under the shaded area (see below):

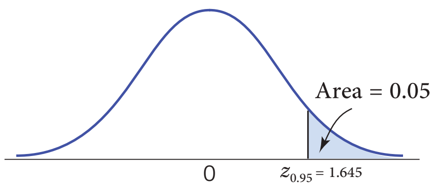


Figure 4.6.2

4. **Decision:**

   The test statistic value falls in the critical region, reject H_0, then the test supports the inspector's claim at $\alpha = 0.05$ 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 support the inspector's claim at $\alpha = 0.05$ 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

$$p_0 = 0.04, \ n = 500, \ \hat{p} = \frac{16}{500} = 0.032, \ \alpha = 0.05$$

Following the test steps described before, we have:

1. **Identifying the null hypothesis H_0 and the alternate hypothesis H_1:**

$$H_0 : p = 0.04 \quad \text{versus} \quad H_1 : p \neq 0.04$$

2. **Selecting the test statistic and determine its value under the null hypothesis H_0:**
   The statistic value 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:

$$\pm z_{1-(\alpha/2)} = \pm z_{0.975} = \pm 1.96 \text{ (see below)}$$

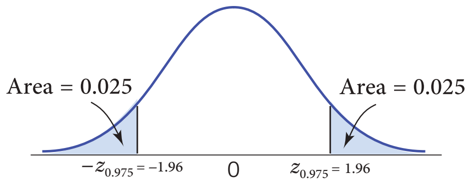


Figure 4.6.3

4. **Decision:**
   The test statistic value does not fall in the critical region, so, we cannot reject H_0, this is not significant evidence, at the $\alpha = 0.05$ 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.

▶ **EXAMPLE 4.6.3** An insurance company advertised that it handles 85% of all cars insurance claims within 15 days. In a sample of 260 cars insurance claims, 208 were handled within 15 days. Do you support the advertisement at the 2.5% level of significance?

Solution: In this example we have

$$p_0 = 0.85, \ n = 260, \ \hat{p} = \frac{208}{260} = 0.80, \ \alpha = 0.025$$

Following the test steps described before, we have:

1. **Identifying the null hypothesis H_0 and the alternate hypothesis H_1:**

$$H_0 : p = 0.85 \quad \text{versus} \quad H_1 : p < 0.85$$

2. **Selecting the test statistic and determine its value under the null hypothesis H_0:**

   The statistic value can be calculated using the available random sample and the specified value of the population proportion under the null hypothesis H_0 as:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.80 - 0.85}{\sqrt{\frac{0.85(0.15)}{260}}} = -2.257884 \approx -2.26$$

3. **The critical region:**

   The test in this case is left-tailed $(H_1 : p < 0.85)$, then the critical regions are the parts of the real axis under the shaded areas, which determined by $-z_{1-\alpha} = -z_{0.975} = -1.96$ (see below)

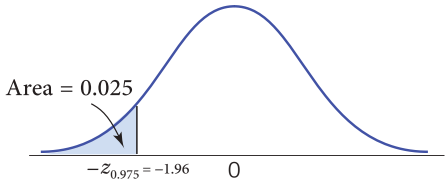


Figure 4.6.4

4. **Decision:**

   The test statistic value falls in the critical region, so, we reject H_0, this is not sufficient evidence, at $\alpha = 0.025$ level of significance, to support the advertisement.

On the other hand**:**

$$p - value = P(Z \leq -2.26) = 0.0119 < \alpha = 0.025$$

This is means that the company's claim is incorrect (exaggerated) at the 2.5% level of significance.

EXERCISES

1. 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. Then calculate the following:
   - **a.** The standard error of the proportion for this sample.
   - **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.

2. A random sample is drawn from a population of known standard deviation 11.3. Then:
   - **a.** Construct a 90% confidence interval for the population mean based on the information given.
     1. $n = 36$, $\bar{X} = 105.2$ and $s = 11.2$
     2. $n = 100$, $\bar{X} = 105.2$ and $s = 11.2$
   - **b.** Compare between the results in (1) and (2).

3. A random sample is drawn from a population of known standard deviation 22.1. Then:
   - **a.** Construct a confidence interval for the population mean based on $n = 36$, $\bar{X} = 182.4$ and $\sigma = 9$, when:
     1. $1 - \alpha = 0.9$
     2. $1 - \alpha = 0.95$
   - **b.** Compare between the results in (1) and (2).

4. 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 95% confidence level, what is the confidence interval for the population mean?

5. 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.

6. 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?

7. 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. Determine a 95% and 99% confidence intervals estimate of the mean length of time that these components function.

8. 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.

9. 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 for the parameter of population of interest?
   b. Construct 90% confidence interval of the population proportion.

10. 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?

11. 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.

12. 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?

13. 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 0.8 ounces. Using a 0.10 level of significance, can it be concluded that the process is out of control (not equal to 16 ounces)?

14. In the past the average waist size of adult males in a town has been 36 inches with a standard deviation of 3 inches. If you collect a random sample of size 36 men, and you found that the average waist size is 37.5 inches. Use significance level $\alpha = 0.01$ to

determine whether the average waist size of males in the town is now greater than 36 inches.

15. A tire manufacturing plant produces 15 tires per hour. This yield has an established variance of 2 (σ =1.414 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.

16. 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 153 minutes and a standard deviation 7.5 minutes. At the 0.05 level of significance, test the consumer group's belief.

17. 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 in the liter, with the overall standard deviation of 0.92 mg in the liter. A water purifying company claims that the mean level of oxygen in the water is 5 mg in the liter. Conduct a hypothesis test to determine whether the mean oxygen level is less than 5 mg in the liter. Use 5% level of significance.

18. 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 30.45 seconds, can we conclude, at the 5% level of significance, that the authorities are incorrect? What about at the 1% level of significance?

19. Consider the following hypothesis test: $H_0 : p = 0.8$ *versus* $H_1 : p > 0.8$ A sample of 400 elements of the population provided a sample proportion of 0.853.

a. Using $\alpha = 0.05$, what is the conclusion based on classical hypothesis test?

b. Using $\alpha = 0.01$, what is the conclusion based on p -value?

20. 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.

21. 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.

22. 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 0.025.

23. 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?

24. A random survey of 80 death row inmates revealed that the average length of time on death row is 17 years with a standard deviation of 6.5 years. If you wish to test the hypothesis that the population average time on death row could likely be 16 years.

a. Is this a test of means or proportions?

b. State the null and alternative hypotheses.

25. A bank wishes to estimate the mean balances owed by customers holding Master Card. The population standard deviation is estimated to be \$300. If a 95 percent confidence interval is used and an interval of \$70 is desired, how many cardholders should be sampled?

26. A group of statistics students decided to conduct a survey at their university to find the mean amount of time students spent studying per week.

a. Assuming a standard deviation of 8 hours, what is the required sample size if the error should be less than ½ hour with a 96% level of confidence?

b. Assuming a standard deviation of 5 hours, what is the required sample size if the error should be less than ½ hour with a 95% level of confidence?

27. A survey of 144 TV devices revealed that the average price of a TV was \$375 with a standard error of \$25.

a. What is the 96% confidence interval to estimate the true cost of the TV?

b. What is the 99% confidence interval to estimate the true cost of the TV?

c. If 90% and 95% confidence intervals were developed to estimate the true cost of the TV, what similarities would they have?

d. If 96% and 98% confidence intervals were developed to estimate the true cost of the TV, what differences would they have?

28. 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 $\sigma = 4.5$.

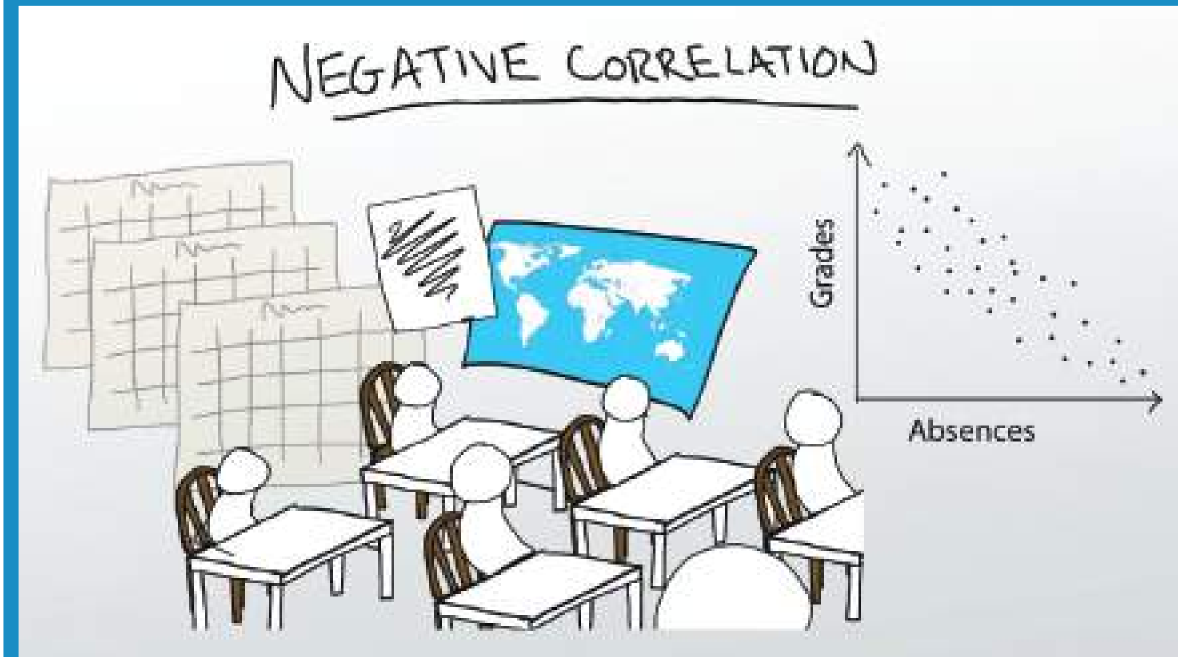
a. Determine a 95% confidence interval for the mean weight of all such runners.

b. Based on this confidence interval, does a test of

$$H_0 : \mu = 61.3 \ kg$$
$$H_A : \mu \neq 61.3 \ kg$$

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 SIMPLE LINEAR CORRELATION
- SECTION 5.2 SIMPLE LINEAR REGRESSION
- EXERCISES

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, 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.

Section 5.1
SIMPLE LINEAR CORRELATION

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 $(x_1, y_1), (x_2, y_2), \ldots$and (x_n, y_n) where x_i is the value of X for the i^{th} observation, y_i is the value of Y for the i^{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 (pairs) (x_1, y_1), $(x_2, y_2), \ldots$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 this representation (see the following figure 5.1.1).

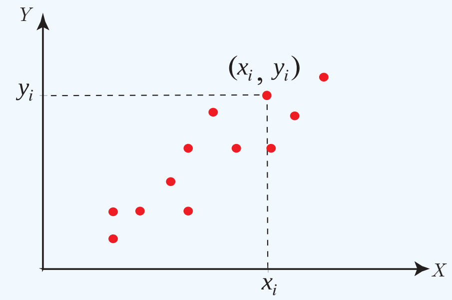


Figure 5.1.1 (Scatter diagram)

CORRELATION COEFFICIENT

In fact, there are many measures to show the correlation between two phenomena represented by two variables 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 one of these formulas that you may encounter. The most common formula for computing a product-moment correlation coefficient ($\mathbf{r}$) is given by the following definition.

DEFINITION 5.1.2 (Pearson's Correlation Coefficient)

Let (x_1, y_1), (x_2, y_2),...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:

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Or using the following relation:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

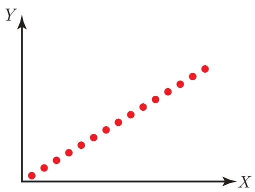
How to interpret the correlation coefficient?

The sign and the absolute value of a correlation coefficient ($\mathbf{r}$) describe the direction and the magnitude of the relationship between two variables or two phenomena.

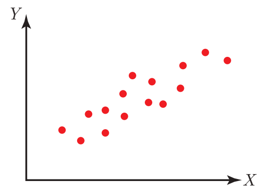
- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of the correlation coefficient, the greater the correlation between the two variables (or phenomena).
- The strong linear relationship is indicated by a correlation coefficient, that is close to ± 1 or equal to ± 1, and when the correlation coefficient ($\mathbf{r}$) is equal to ± 1, then one say 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 ($\mathbf{r}$) 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).

- A positive correlation means that if one variable gets bigger value, the other variable tends to get bigger value also, 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.

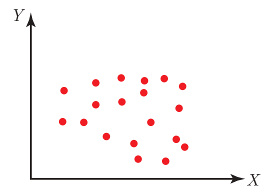
The scatterplots below show how different patterns of data produce different degrees of linear correlation.



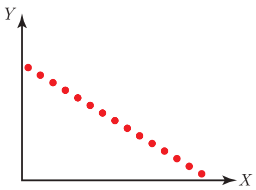
Maximum positive correlation
(**r** = 1.0)



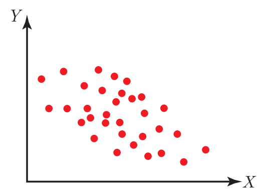
Strong positive correlation
(**r** = 0.80)



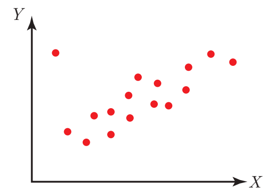
Very Weak correlation
(**r** = 0.25)



Maximum negative correlation
(**r** = -1.0)



Weak negative correlation
(**r** = -0.45)



Strong correlation & outlier
(**r** = 0.7)

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 with unclear direction, the correlation is equal to zero or very close to zero.
- Correlation is affected by outliers. Compare the second scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 0.80 to 0.70).

- 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.
- The square value of correlation coefficient ($\mathbf{r}$) is called the coefficient of determination and one denoted it by $\mathbf{r}^2$.

ASSESSMENT OF CORRELATION STRENGTH

The Relationship between the two variables X and Y (or phenomena)	The Range of $\mathbf{r}$
No linear or $S_X = 0$ or $S_Y = 0$	$\mathbf{r} = 0$
Very weak	$0 < \lvert \mathbf{r} \rvert \leq 0.30$
Weak (an acceptable degree of linearity)	$0.30 < \lvert \mathbf{r} \rvert \leq 0.50$
Moderately strong linear	$0.50 < \lvert \mathbf{r} \rvert \leq 0.70$
Strong (the linearity very clear)	$0.70 < \lvert \mathbf{r} \rvert \leq 0.86$
Very Strong (high degree of linearity)	$0.86 < \lvert \mathbf{r} \rvert < 1$
Complete (all points are located on one straight)	$\lvert \mathbf{r} \rvert = 1$

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

a. Represent the given data on the scatter plot.

b. Is there a linear relationship (linear association) between X and Y? Is it positive or negative?

c. Calculate the correlation coefficient ($\mathbf{r}$).

Solution: We have:

For a) The scatter plot for the given data is:

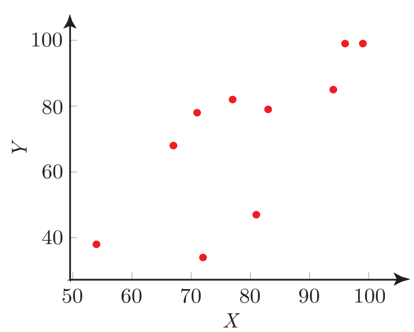


Figure 5.1.3

For b) The scatter plot 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.

For c) To calculating the coefficient of correlation ($\mathbf{r}$) we will create the following table:

Table 5.1.1

i	x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
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
Total	794	709	0	0	1818.4	5040.9	2272.4

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{794}{10} = 79.4 \text{ , } \overline{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{709}{10} = 70.9$$

$$\sum_{i=1}^{10}(x_i - \overline{x})^2 = 1818.4 \text{ , } \sum_{i=1}^{10}(y_i - \overline{y})^2 = 5040.9 \text{ } and \text{ } \sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y}) = 2272.4$$

Then the correlation coefficient is:

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}((x_i - \overline{x}^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{2272.4}{\sqrt{1818.4}\sqrt{5040.9}} = 0.75056 \approx 0.75$$

Alternatively, we can use the relation:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Where we have:

i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$
1	77	5929	82	6724	6314
2	54	2916	38	1444	2052
3	71	5041	78	6084	5538
4	72	5184	34	1156	2448
5	81	6561	47	2209	3807
6	94	8836	85	7225	7990
7	96	9216	99	9801	9504
8	99	9801	99	9801	9801
9	83	6889	79	6241	6557
10	67	4489	68	4624	4556
Total	794	64862	709	55309	58567

Therefore, we have:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$= \frac{585670 - 794 \times 709}{\sqrt{648620 - (794)^2}\ \sqrt{553090 - (709)^2}} = 0.75056$$

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 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 (ε) captures the amount of variation that is not predicted by the slope and intercept terms. The regression coefficient ($\mathbf{r}^2$) 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 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 X.

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 SR 2 per topping. If we let $X =$ toppings and $Y=$ price of pizza, then $Y = 10 + 2X$. If we order a 3-topping pizza, then $Y = 10 + 2\left(3\right) = 16$ 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:

$$\boxed{Y = a + bX + \mathcal{E}}$$

Where:

- $\mathcal{E}$ is a normal random variable with zero expectation $\mathrm{E}(\mathcal{E}) = 0$. This term $(\mathcal{E})$ 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).

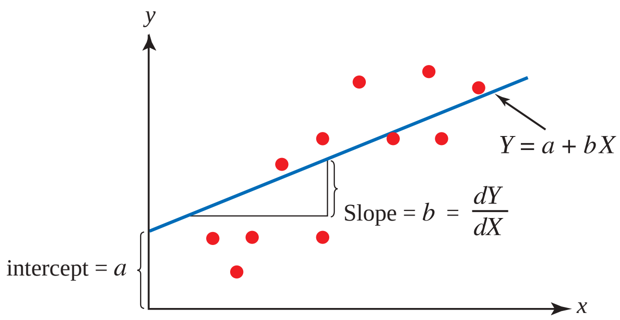


Figure 5.2.1

THE METHOD OF LEAST SQUARES FOR ESTIMATING *a* and *b*

Now, when we have a sample (x_1, y_1), (x_2, y_2), … and (x_n, y_n), where x_1, x_2, … and x_n are values of X with mean $\bar{x}$ and standard deviation S_X, and y_1, y_2, … and y_n are values of

Y with mean $\overline{y}$ and standard deviation S_Y. Then the least squares method is used to find the 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 squared errors equal zero), this can be seen as the expected value of the random term $E(\mathcal{E}) = 0$. So, the estimated regression line for the given sample can be obtained (without proof) to be:

$$\boxed{\hat{Y} = \hat{a} + \hat{b}\,X}$$

where the coefficients $\hat{a}\, and\, \hat{b}$ cab be estimated as:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{Or by the relation} \quad \hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} \quad \text{Or by the relation} \quad \hat{a} = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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 (**r**). 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 $y_1, y_2, \ldots$ and y_n values about their mean $\overline{y}$) that is explained by (or attributed to) the linear relationship between X and Y. Some other way to calculate the coefficient of determination as:

$$\boxed{\mathbf{r}^2 = \frac{SSR}{SS_{tot}} = 1 - \frac{SSE}{SS_{tot}}}$$

Where:

The total sum of squared deviations (Total Variation) is:

$$SS_{tot} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

The sum of squared regression error (Explained Variation) is:

$$SSR = \sum_{i=1}^{n} (\overline{y} - \hat{y}_i)^2$$

The sum of squared error (or residuals) (Unexplained Variation) is:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and

$$SS_{tot} = SSR + SSE$$

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \le \mathbf{r}^2 \le 1$. If $\mathbf{r}^2 = 0$, the least squares regression line has no explanatory value. If $\mathbf{r}^2 = 1$, 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.

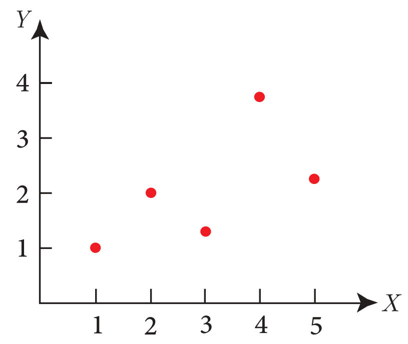


Figure 5.2.2

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.

a. Calculate the correlation coefficient between X and Y,

b. Estimate the simple linear regression line $\hat{Y} = \hat{a} + \hat{b}X$

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
------	$\sum x_i = 15$	$\sum y_i = 10.3$	$\sum x_i^2 = 55$	$\sum y_i^2 = 25.815$	$\sum x_i \cdot y_i = 35.15$

Then the linear correlation coefficient is given by:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$= \frac{5(35.15) - (15)(10.3)}{\sqrt{[5(55) - (15)^2][5(25.815 - (10.3)^2)]}} = 0.63$$

b. Linear regression interested in 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:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

And

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x}$$

In order to use these form to calculate the estimates of $\hat{b}$ and $\hat{a}$ we need to contract the following table:

Table 5.2.2

i	x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
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
Total	15	10.3	0	0	10	4.25

From the table, we have:

$$\overline{x} = \frac{15}{5} = 3 \text{ and } \overline{y} = \frac{10.3}{5} = 2.06$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{4.25}{10} = 0.425$$

And

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} = 2.06 - (0.425)(3) = 0.785$$

Hence, the estimated simple linear regression model is:

$$\hat{Y} = 0.785 + 0.425\,X$$

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	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	ε_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
Total	15	10.3	10.3	0	2.79075

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 point (1,1) is very near the regression line; its error of prediction is small. By contrast, the point (4,3.75) is much higher than the regression line and therefore its error of prediction is large.

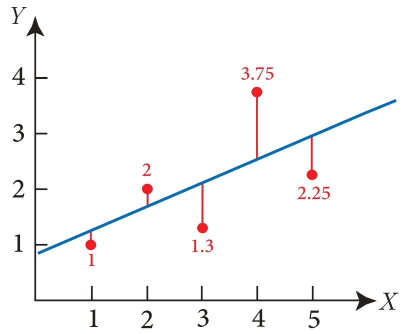


Figure 5.2.3

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 y_i is the value y_i minus the predicted value $\hat{y}_i$ (the value on the line). The above table 5.2.4 shows the predicted values $\hat{y}_i$ and the errors of prediction $\varepsilon_i = y_i - \hat{y}_i$. For example, the first point (1,1) has a y_1 of 1.00 and its predicted value (is called $\hat{y}_1$) has 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 table shows the squared errors of prediction. The sum of the squared errors of prediction shown in the above 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 represents the lot size and Y the number of Man-hours labor for recent production runs. The data is given in the table below.

The values of X	30	20	60	80	40	50	60	30	70	60
The values of Y	73	50	128	170	87	108	135	69	148	132

a. Construct the scatter diagram.

b. 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 has the form:

$$Y = a + b\,X + \mathcal{E}$$

is appropriate to describe the relationship between X and Y?

c. Estimate the parameters of the linear regression line $Y = a + b\,X$ 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 ($\mathbf{r}^2$) and hence deduce the simple linear correlation coefficient ($\mathbf{r}$) 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:

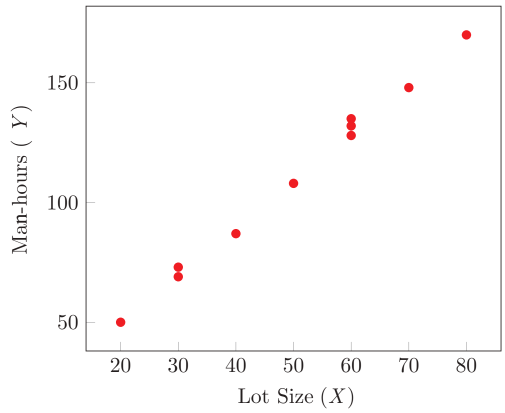


Figure 5.2.4

The scatter plot 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 plot of 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 $Y = a + b\,X + \mathcal{E}$ 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

i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$
6	50	2500	108	11664	5400
7	60	3600	135	18225	8100
8	30	900	69	4761	2070
9	70	4900	148	21904	10360
10	60	3600	132	17424	7920
Total	500	28400	1100	134660	61800

From the previous table, we have:

$$\sum_{i=1}^{10} x_i = 500, \sum_{i=1}^{10} y_i = 1100, \sum_{i=1}^{10} x_i^2 = 28400, \sum_{i=1}^{10} y_i^2 = 134660 \; and \; \sum_{i=1}^{10} x_i \cdot y_i = 61800$$

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.

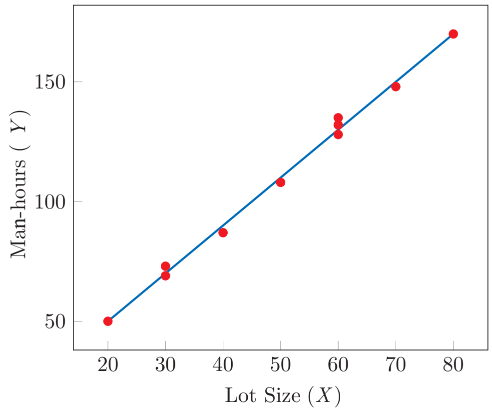


Figure 5.2.5

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 ($\mathbf{r}^2$) we need to get the sum of squared errors as:

Table 5.2.4-b

i	y_i	$\hat{y}_i$	$(y_i - \overline{y})$	$(y_i - \overline{y})^2$	$(y_i - \hat{y})$	$(y_i - \hat{y})^2$	$\hat{y}_i - \overline{y}$	$(\hat{y}_i - \overline{y})^2$
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
Total	1100	1100	0	13660	0	60	0	13600

From the table 5.2.4-a and table 5.2.4-b, we find that:

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{500}{10} = 50 \quad and \quad \overline{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{1100}{10} = 110$$

- The total sum of squared variations $= SS_{tot} = \sum_{i=1}^{10} \left(y_i - \overline{y}\right)^2 = 13660$
- The sum of squared regression error $= SSR = \sum_{i=1}^{10} \left(\hat{y}_i - \overline{y}\right)^2 = 13600$
- The sum of squared error $= SSE = \sum_{i=1}^{10} \left(y_i - \hat{y}_i\right)^2 = 60$

It is clear that $SS_{tot} = SSR + SSE$.

The coefficient of determination is:

$$\mathbf{r}^2 = \frac{SSR}{SS_{tot}} = \frac{13600}{13660} = 0.9956076 \approx 99.6\%$$

Or

$$\mathbf{r}^2 = 1 - \frac{SSE}{SS_{tot}} = 1 - \frac{60}{13660} = 1 - 0.00439 = 0.99561 \approx 99.6\%$$

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 $\mathbf{r} = \pm\sqrt{\mathbf{r}^2}$ where the sign is chosen as the sign of the slope estimation $\hat{b}$.

In Our example, the linear correlation coefficient is:

$$\mathbf{r} = \pm\sqrt{\mathbf{r}^2} = \sqrt{0.99561} = 0.9978$$

this value of the linear correlation coefficient can be also obtained by using Pearson Correlation Coefficient to get the same result.

EXERCISES

1. When the dependent and independent variables are plotted, the graph is called

   A. Scatter diagram **B.** Bar chart
   C. Pie chart **D.** Histogram

2. When a variable X is used to predict a value x of the variable, then X is called:

   A. Independent **B.** Correlation
   C. Dependent **D.** Determination

3. Which of the following statements regarding the coefficient of correlation is not true?

   a. It ranges from –1 to +1 inclusive.
   b. It measures the strength of the relationship between two variables.
   c. A value of 0 indicates two variables are not related linearly.
   d. None of the above.

4. What does a coefficient of correlation of 0.80 infer?

   a. Almost no correlation because 0.80 is close to 1.
   b. 80% of the variation in one variable is explained by the others.
   c. The coefficient of determination is 0.64.
   d. Coefficient of non-determination is 0.20.

5. What is the range of values for a coefficient of determination?

   A. 0 to +1 **B.** –1 to +1 inclusive
   C. –3 to +3 inclusive **D.** Unlimited range

6. If the correlation between two variables is close to one, the association is:

   A. very 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 linearly 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 $\mathbf{r} = -1$, 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 $\mathbf{r} = 0.65$, 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 $\mathbf{r} = 0.89$?
A. 0.94 **B.** 0.79
C. 0.89 **D.** 0.06

12. Which values of **r** indicates a moderately strong linear correlation?
A. – 0.55 **B.** 0.38
C. – 0.48 **D.** 0.69

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 $\hat{Y} = 1.202 + 1.133X$. 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:

   a. Predict the value of the dependent variable given a value of the independent variable.

   b. Predict the value of the independent variable given a value of the dependent variable.

   c. Measure the association between two variables.

   d. all of the above.

17. In the regression line equation, $\hat{Y} = 10 + 20X$ the value of 20 indicates:

   a. The y intercept.

   b. For each unit increase in X, Y increases by 20.

   c. For each unit increase in Y, X increases by 20.

   d. None of the above.

18. In the equation $\hat{Y} = a + bX$, what is $\hat{Y}$?

   a. Slope of the linear regression line.

   b. Y intercept.

   c. Predicted value of Y, given a specific X value.

   d. Value of Y when $X = 0$.

19. Assume the regression line equation is $\hat{Y} = 10 + 20X$. What does the value of 10 in the equation indicate?

   a. Y intercept.

   b. For each unit increased in Y, X increases by 10.

   c. For each unit increased in X, Y increases by 10.

   d. None of the above.

20. Given the following five points: (–2,0), (–1,0), (0,1), (1,1), and (2,3).

   a. What is the slope of the linear regression line?

   b. What is the Y intercept?

21. 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:

$$\sum(x_i - \overline{x})^2 = 498, \ \sum(y_i - \overline{y})^2 = 61.2 \ \text{ and } \ \sum(x_i - \overline{x})(y_i - \overline{y}) = -164$$

Then:

a. The coefficient of correlation ($\mathbf{r}$) equals to:

A. – 0.939 **B.** 0.25 **C.** 0.68 **D.** – 0.81

b. Based on the value of the coefficient of correlation ($\mathbf{r}$), the linear relationship between X and Y is:

A. Weak **B.** Moderate **C.** Strong **D.** Complete

22. 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

a. What is the independent variable (X)?

b. 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 (linear line) equation for the data?

f. What is the meaning of a negative slope?

23. 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?

24. 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 at 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?

25. 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 correlation and determination.

c. Interpret the results in (a) and (b).

26. 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

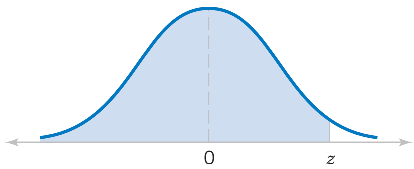
a. Estimate the linear regression line.

b. Calculate the coefficients of correlation and determination.

REFERENCES

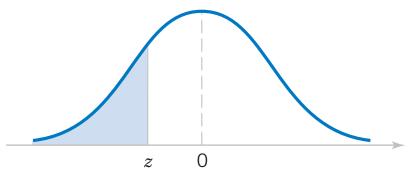
1. Bernstein, Stephen, and Ruth Bernstein. Schaums Outline of Elements of Statistics I: Descriptive Statistics and Probability. McGraw Hill Professional, 1999.
2. Brase C.H. and Brase, C.P. Understanding Statistics, Fifth edition, Brooks/Cole, Canada, 2010.
3. Chase W. and Bown, F. General Statistics, 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									



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

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

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