Basic Statistics

Seemon Thomas

Alpha Science

Basic Statistics

Basic Statistics

Seemon Thomas

Alpha Science International Ltd.
Oxford, U.K.

Basic Statistics
178 pgs. | 30 figs. | 12 tbls.

Seemon Thomas
Department of Statistics
St. Thomas College
Arunapuram, Kerala

ALPHA SCIENCE INTERNATIONAL LTD.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

Printed from the camera-ready copy provided by the Author.

ISBN 978-1-84265-849-9

E-ISBN 978-1-78332-030-1

Printed in India

PREFACE

The purpose of this book is to introduce the basic concepts and methods in statistics. It assumes only the knowledge of plus-two mathematics. The book aims to introduce a number of simple but important statistical techniques and is intended for undergraduate students as well as others who want a mathematical introduction to probability and statistics. It is designed for one semester course at the college level.

I feel that the text provides an excellent balance between theory and applications. Each topic discussed is illustrated with worked-out examples, many of which are taken from real life situations. A large number of applied and theoretical exercises are included at the end of each chapter.

Many teachers find teaching Statistics a challenge, and most of us are looking for computer-based resources to enliven our classrooms. In preparing this book I was guided by the thought that now-a-days the study of statistics is impossible without resorting to computers. The computational steps are explained using MS Excel for those problems which require computations. Students are encouraged to use computers to perform calculations. The Internet offers a huge array of teaching resources for statistics and some of them are listed in this book.

I wish to express deep gratitude to Professor A.M. Mathai, Emeritus Professor of McGill University, Canada for the careful reading of the original manuscript and making necessary corrections.

I am grateful to the Management of St. Thomas College Pala and my colleagues in the Department of Statistics for providing me with the much needed moral support and facilities.

I hope that students will find this book interesting and informative.

Seemon Thomas

CONTENTS

1. INTRODUCTION 1

1.1 WHAT IS MEANT BY STATISTICS? 1
1.2 STATISTICS FOR HUMAN WELFARE 2
1.3 A HISTORICAL NOTE 4
1.4 STATISTICS IN INDIA 5
1.4.1 Census in India 6
1.5 JOB CHARACTERISTICS OF A STATISTICIAN 8

2. ORGANIZATION OF DATA 9

2.1 INTRODUCTION 9
2.2 POPULATION AND SAMPLE 9
2.3 CENSUS AND SAMPLE SURVEY 10
2.3.1 Types of Sampling 11
2.4 PRIMARY AND SECONDARY DATA 13
2.5 COLLECTION OF PRIMARY DATA 14
2.6 TYPES OF VARIABLES 16
2.6.1 Qualitative Variables 16
2.6.2 Quantitative Variables 17
2.7 EXERCISES 18

3. PRESENTATION OF DATA 21

3.1 INTRODUCTION 21
3.2 TABULAR PRESENTATION 21
3.2.1 Frequency Table 22
3.3 GRAPHIC PRESENTATIONS OF DATA 25
3.3.1 Stem and Leaf Plot 31
3.3.2 Graphic Presentation of Grouped Frequencies 32
3.4 EXERCISES 39

4. MEASURES OF CENTRAL TENDENCY 43

4.1 INTRODUCTION 43
4.2 THE ARITHMETIC MEAN 43

4.3 THE MEDIAN 51
4.4 THE MODE 53
4.5 THE GEOMETRIC MEAN 55
4.6 THE HARMONIC MEAN 57
4.7 PERCENTILES, DECILES AND QUARTILES 58
4.8 EXERCISES 60

5. MEASURES OF DISPERSION AND SKEWNESS 67

5.1 INTRODUCTION 67
5.2 THE RANGE 68
5.3 THE MEAN DEVIATION 68
5.4 THE VARIANCE AND THE STANDARD DEVIATION 70
5.5 THE INTERQUARTILE RANGE 77
5.6 SKEWNESS IN THE DATA 77
5.7 BOX PLOT 80
5.8 HOW DESCRIPTIVE STATISTICS CAN BE MISUSED? 82
5.9 EXERCISES 83

6. INTRODUCTION TO PROBABILITY THEORY 91

6.1 INTRODUCTION 91
6.2 RANDOM EXPERIMENT 94
6.2.1 Algebra of Events 96
6.3 PROBABILITY 98
6.3.1 Mathematical or Classical or 'a priori' Definition of Probability .. 98
6.3.2 Empirical or Statistical or Relative Frequency Definition of Probability 99
6.3.3 Axiomatic Definition of Probability 100
6.3.4 Theorems of Probability 101
6.4 PROBABILITY PROBLEMS 106
6.4.1 Probability for Uncountable Outcomes 115
6.4.2 Odds Ratio 116
6.5 SUBJECTIVE PROBABILITY 117
6.6 EXERCISES 119

7. CONDITIONAL PROBABILITY AND BAYES' THEOREM 127

7.1 INTRODUCTION 127
7.2 CONDITIONAL PROBABILITY 129
7.2.1 Independent Events 131
7.2.2 Independent Trials 136

7.3 THEOREM OF TOTAL PROBABILITY AND BAYES' FORMULA 139
7.4 EXERCISES 145

8. INDEX NUMBERS 151

8.1 INTRODUCTION 151
8.2 SIMPLEST TYPES OF INDEX NUMBERS 153
8.3 CONSTRUCTION OF INDEX NUMBERS 154
8.3.1 Consumer Price Index 158
8.3.2 Wholesale Price Index 160
8.3.3 Changing the Base Period or Splicing 162
8.3.4 Tests of Index Numbers 163
8.4 USES OF INDEX NUMBERS 164
8.5 EXERCISES 166

Chapter 1

INTRODUCTION

"The science of statistics is essentially a branch of applied mathematics and may be regarded as mathematics applied to observational data"- Sir R.A Fisher.

1.1 WHAT IS MEANT BY STATISTICS?

We encounter the word 'statistics' frequently in our everyday life. We use statistics as a plural and singular noun. When it is used in the plural sense it refers to a collection of numerical information. Examples include:

- The number of deaths occurred due to road accidents in Kerala during the last year.
- The mean age at marriage of females belonging to Muslim community.
- The batting-average of a particular batsman.

The singular noun 'Statistics' using the upper-case S, is the subject of Statistics and it has a much broader meaning than just collecting and publishing numerical information. Statistics is a systematic and scientific study of data and provides methods for producing and understanding data. Eventhough there is a prolific increase in the use of statistical methods, the subject is still largely unexplored and under used. This is either because of resource constraints or the benefits of the methods are not sufficiently promulgated.

Statistics is the scientific application of mathematical principles to the collection, analysis, presentation and interpretation of numerical data. Statisticians contribute to scientific enquiry by applying their mathematical and statistical knowledge to the design of surveys and experiments; the collection, processing, and analysis of data; and the interpretation of the results.

Statistical studies fall into many categories-*descriptive, visual, deductive, inductive statistics etc.* Descriptive statistics include methods of summarizing and presenting data in an informative way. To most people it is this notion that is conveyed by the word statistics: a summarization and presentation of huge mass of data in the form of tables, graphs, charts, histograms, summary items such as averages etc.

Inductive statistics is concerned with developing and using mathematical tools to make forecasts and inferences on the basis of a random sample or from a small part or subset or sample of observations of the population or system which produced these observations. To understand the principles of inductive and deductive procedures one has to be familiar with the concepts of probability theory. Most often statistical procedures cater to the needs of research scientists in such diverse disciplines as biological sciences, industrial engineering, economics and social sciences to mention just a few.

The term *experiment* is open to a very broad interpretation and covers any type of study, trial, or investigation where data are to be collected and assessed. Experiments can often be simple in nature involving comparison of, for example, two teaching methods, animal growth using two different animal feeds, pH level of water in two rivers and tensile strength of two alloys. To carry out such assessment appropriately requires the use of basic statistical methods. Hence Statistics is an indispensable tool for any scientific experiment– right from the stage of planning the experiment to the stage of drawing inference from the data. Efficient managers of any kind of organization are well aware of the importance of statistical tools to provide accurate and timely information to make wise decisions. Here are a few examples:

- Pharmaceutical companies study the cure rates of diseases using different drugs and different forms of treatment. For example, if you take an aspirin each day, does that reduce your risk of a heart attack?

- Insurance companies use statistical analysis to determine premium for life/ health/ automobile/ wealth insurance policies. Tables are available showing estimates that a 30 year old man or 55 year old lady has how many years remaining in the sense of expected life span.

- Industrial and scientific experiments can often be simple in nature involving comparison of, for example, two advertising strategies, toxic chemical removal by two waste management strategies, tensile strength of two nickel-titanium alloys, and animal growth using two different feeds.

To carry out such assessment appropriately requires the use of statistical tools.

1.2 STATISTICS FOR HUMAN WELFARE

Medicine
The search for improved medical treatments rests on careful experiments that compare promising new treatments with the current state of the art. Statisticians work with medical teams to design experiments and analyze the complex data they produce.

Environment
Studies of the environment require data on the abundance and location of plants and animals, on the spread of pollution from its sources, and on the possible effects of changes in human activities. Endangered species of fish and other wildlife can be identified through the effective use of Statistics.

Industry
The future of industries depends on improvement in the quality of goods and services and the efficiency with which they are produced and delivered. Improvement should be based on data, rather than guesswork. More companies are installing elaborate systems to collect and act on data to better serve their customers.

Market research
Are viewer tastes in television programs changing? Which are the promising locations for a new retail outlet? What is the ideal time for releasing a new film? Market researchers conduct their own surveys to answer questions such as these. Statisticians design surveys that gather data to answer these type of questions.

Sports
Everybody is heard of the Duckworth-Lewis method in cricket. This method was devised by two British statisticians, Frank Duckworth and Tony Lewis. It is a mathematical formulation designed to calculate the target score for the team batting second in a one-day cricket match interrupted by weather or other circumstance. In this method the scoring potential of the team is expressed as a function of wickets and overs. The performance of an athlete or a player is always ranked using statistical tools.

Government surveys
How much is the production of food grains this year? What do we export to China, and what do we import? Are rates of violent crime against women increasing in Kerala? The government wants data on issues such as these to guide policy, and statistics agencies of the government provide them data by conducting various surveys. The Organisation for Economic Co-operation and Development (OECD) now provides a comprehensive range of governmental, social, and economic data for developed nations. For developing nations, the World Bank's development of poverty indicators and measures of business and economic conditions have contributed greatly to public debate and analysis of national and international policies affecting poor people across the world.

The role of Statistics in shaping governmental policies has now expanded. Social reforms are usually initiated as a result of statistical analyses of factors such as crime rates and poverty levels. In today's world the exercise of effective citizenship increasingly requires a public that is competent to evaluate arguments grounded in numerical evidence. To the extent the public lacks the skills to critically evaluate the statistical analyses that shape public policy, more crucial decisions that affect our daily lives will be made by administrators who have these statistical skills or by those who would use their mastery of these skills to serve their own supporters or special interest ends. As educated and responsible members of the society we must sharpen our ability to recognize distorted data; in addition we must also learn to interpret undistorted data intelligently. Statistics have powerful and far-reaching effects on everyone, yet most people are unaware of their connection-from the foods they eat to the medicines they take-and how statistics improve their lives.

The impact of Statistics has led to the development of new disciplines like econometrics, industrial quality control, psychometry and bioinformatics. A statistician can combine his interest with almost every field of human activity. In this text, we will explore some aspects of descriptive statistics and elementary probability theory. Apart from

job-motivated reasons, the study of Statistics will equip you to analyse any information more critically so that you are less susceptible to deceptive claims. An understanding of statistical methods presented in the text will help you to draw right conclusions from data and make wise decisions.

The first World Statistics Day was celebrated on 20th October 2010 (20.10. 2010). The three key words chosen to highlight World Statistics Day were Service, Professionalism and Integrity. The next World Statistics Day will be celebrated in 2015. 2013 is the International Year of Statistics, a worldwide event supported by more than 1,700 organizations!

1.3 A HISTORICAL NOTE

Collection of data began as early as when man started keeping records. The data on population is important to a state or country and it is stated that ancient Babylonians collected data on population. The word statistics was derived from the latin word 'status' meaning state. In the Holy Bible, we read about census while narrating the birth of Jesus Christ. In the first or second century A.D. the magistrates in Rome were asked to prepare registers of the population and wealth by which the state could determine the availability of adult males for military service and amount of tax. In Kautilya's Artha Shastra, there is reference to collection of such data in ancient India. The sphere of data collection now encompasses a variety of fields and the word statistics is thus used today in a much broader context.

At the start of 19th century there occurred a burst of interest in numerical data on a wide variety of topics. In response, the first statistical society organized by Adolphe Quetelet and Charles Babbage was formed in February 1834.

Lambert Adolphe Jacques Qutelet (22 February 1796 - 17 February 1874) was a Belgian astronomer, mathematician, statistician and sociologist. He founded and directed the Brussels Observatory and was influential in introducing statistical methods to the social sciences. Quetelet also founded several statistical journals and societies, and was especially interested in creating international cooperation among statisticians.

Charles Babbage, FRS (26 December 1791 - 18 October 1871) was an English mathematician, philosopher, inventor and mechanical engineer who originated the concept of a programmable computer. Considered a "father of the computer" Babbage is credited with inventing the first mechanical computer that eventually led to more complex designs.

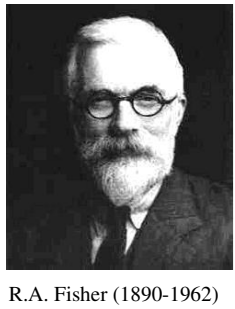
Modern statistical theory had its origins in a diverse collection of practical and theoretical problems. These included:

- Games of chance, which gave rise to initial statements of a theory of probability in the 1650s by Blaise Pascal and Pierre de Fermat;
- Astronomical and geodetic observations, used to calculate the orbits of planets and comets and determine the shape of the earth, with the practical goal of enabling accurate navigation at sea.

1.4 STATISTICS IN INDIA

During the 1920s and until the mid-1930s almost all the statistical work done in India was done single-handedly by Professor Prasanta Chandra Mahalanobis. The early statistical studies in India included analyses of data on stature of Anglo-Indians, meteorological data, rainfall data, data on soil conditions etc. Some of the findings of these early studies were of great impact in the control of floods, development of agriculture, etc., and led to the recognition of Statistics as a key discipline. Mahalanobis' influence was so persuasive that students of Physics began to take interest in Statistics. Subhendu Sekhar Bose was the most notable of them. Later, several talented young scholars including J.M. Sengupta, H.C. Sinha, R.C. Bose, S.N. Roy, K.R. Nair, K. Kishen and C.R. Rao, joined to form an active group of statisticians in India. In 1931, having established a reputation as a consulting statistician, Mahalanobis set up the *Indian Statistical Institute*, then one of the few centres in the world to impart formal instruction in statistical theory and methods to researchers across diverse fields. Theoretical research in Statistics began to flourish in the Institute. Design and analysis of agricultural experiments also bloomed and led to some international contacts, notably with *Sir Ronald Aylmer Fisher* who is known as the *father of Statistics*. India's first computer was installed at the Indian statistical institute, Kolkata in 1956. It was not only the first computer of India but also Asia's first computer outside Japan.The Indian Statistical Institute, Kolkota is now one of the leading academic institutions in the world. It has a long and proud tradition of excellence in training, teaching and research in a number of academic disciplines including statistics, mathematics, computer science, economics, biology, geology, physics and social science. What began in 1931 with a solitary human 'computer' working part-time, now comprises over 250 faculty members and over 1,000 supporting staff and several modern-day personal computers, workstations, minicomputers, supermini computers and mainframe computers!

P.C. Mahalanobis (1893-1972)



R.A. Fisher (1890-1962)

National Sample Survey (NSS) was set up in 1950 on the recommendations of National Income Committee, chaired by late Mahalanobis to fill up large gaps in statistical data for computation of national income aggregates, especially in respect of unorganized / household sector of the economy. Initially, statistical work of NSS, except for fieldwork, used to be carried out by Indian Statistical Institute under the guidance of Professor Mahalanobis, while NSS Directorate was created and assigned the fieldwork. NSS was reorganized as *National Sample Survey Organization* (NSSO) in March 1970. In India and abroad NSSO has become synonymous with reliable estimates on various aspects of economic and social life in India. The association of Professor Mahalanobis

with the then Prime Minister Jawaharlal Nehru led to the beginning of Five Year Plans in the country and developed a model in planning known as the 'Mahalanobis model'. *Professor Mahalanobis* is called the *father of Indian Statistics*. Mahalanobis also founded "*Sankhya*: The Indian Journal of Statistics". Professor Mahalanobis was born on 29th June 1893 and June 29 is being celebrated as National Statistics Day every year.

The *Central Statistical Organization* (CSO) is responsible for coordination of statistical activities in the country, and evolving and maintaining statistical standards. Its activities include National Income Accounting; conduct of Annual Survey of Industries, Economic Censuses and its follow up surveys, compilation of Index of Industrial Production, as well as Consumer Price Indices for Urban Non-manual Employees, Human Development Statistics, Gender Statistics, imparting training in Official Statistics, Five Year Plan work relating to development of Statistics in the States and Union Territories; dissemination of statistical information, work relating to trade, energy, construction, and environment statistics, revision of National Industrial Classification, etc. The responsibility to equip the country's large set of statistical personnel with newer practices in the official statistics and data management has been entrusted to the National Academy of Statistical Administration (NASA) which is functioning under the overall guidance of the CSO. The CSO is located in New Delhi.

On realizing the importance of official statistics Government of India has set up the *Ministry of Statistics and Programme Implementation* (MOSPI) for coordinating the statistical activities in the country. It consists of the National Statistical Organization (NSO) and the Programme Implementation Wing, and is headed by Minister of State (Independent Charge). At the executive level, it is headed by Secretary to Government of India, who is also the Chief Statistician of India.

CSO and NSSO come under NSO. The Programme Implementation Wing of the Ministry, headed by the Principal Advisor, consists of three divisions, namely, Member of Parliament Local Area Development Division (MPLAD Division), Infrastructure and Project Monitoring Division (IPMD), and Twenty Point Programme Division (TPP Division).

1.4.1 Census in India

A population census is the process of collecting, compiling, analyzing and disseminating demographic, social, cultural and economic data relating to all persons in the country, at a particular time in ten years interval. Conducting population census in a country like India, with great diversity of physical features, is undisputedly the biggest administrative exercise. The wealth of information collected through census on houses, amenities available to the households, socio-economic and cultural characteristics of the population makes Indian census the richest and the only source for planners, research scholars, administrators and other data users. The planning and execution of Indian census is challenging and fascinating. The ministry of Home Affairs is responsible for coordination of census activities in India.

India is one of the very few countries in the World, which has a proud history of holding census after every ten years. The Indian census has a very long history behind

it. The earliest literature 'Rig Veda' reveals that some kind of population count was maintained during 800-600 B.C. Kautilya's Arthasastra, written around 321-296 B.C., laid stress on census taking as a measure of State policy for purpose of taxation. During the regime of Mughal King Akbar the Great, the administrative report 'Ain-e- Akbari' included comprehensive data pertaining to population, industry, wealth and many other characteristics. In ancient Rome, too, census was conducted for purpose of taxation. The census of 1881 which was undertaken on 17th February, 1881 by W.C. Plowden, Census Commissioner of India, was a great step towards a modern synchronous census. Since then, censuses have been undertaken uninterruptedly once in every ten years.

The last census of India was conducted in 2011. The first phase of the census called Houselisting and Housing Census was conducted between April to June, 2010. A Schedule was canvassed during this phase to collect information on housing and amenities available to the households. The major departure of the recent census from earlier censuses was canvassing a National Population Register (NPR) at the time of Houselisting and Housing Census. The NPR would be a register of usual residents of the country. It will be a comprehensive identity database that would help in providing the benefits and services under the Government programmes to improve planning and help to strengthen security of the country. The information collected through NPR will be used for providing a Unique Identity Number after a detailed procedure. The second phase of Census 2011 was conducted from 9th to 28th February 2011 with 5 days revision round from 1st to 5th March, 2011 so that the population figures with reference to reference date, that is, 00.00 hours of 1st March, 2011 was obtained. During the second phase, Household Schedule containing 29 questions was canvassed.

The Indian census has not been a mere statistical operation and the data collected is not only properly scrutinized at different levels but also presented with cross classification of various parameters for interpretation and analysis in an interesting manner. It may be seen from the history of Indian census that how the changes have taken place from one census to other depending upon the need of the time, country and also demand of the data users and development of technology. The Indian census is well recognized for the data it reveals. Problems relating to political, social and cultural reasons also makes it challenging. The Indian census is the most credible source of information on Demography (Population characterstics), Economic Activity, Literacy and Education, Housing and Household Amenities, Urbanisation, Disability, Fertility and Mortality etc. India's population in 1901 was about 238.4 million, which has increased by more than four times in 110 years to reach a population of 1,210 million in 2011.

The delimitation/reservation of constituencies- Parliamentary/ Assembly/ Panchayats and other Local Bodies is also done on the basis of the demographic data thrown up by the Census. Census is the basis for reviewing the country's progress in the past decade, monitoring the ongoing schemes of the Government and most importantly, plan for the future. That is why the slogan: "Our Census-Our Future".

1.5 JOB CHARACTERISTICS OF A STATISTICIAN

Statisticians provide crucial guidance in determining what information is reliable and which predictions can be trusted. They often help search for clues to the solution of a scientific mystery and sometimes keep investigators from being misled by false impressions.
A statistician will:

- Use data to solve problems in a wide variety of fields.
- Apply mathematical and statistical knowledge to social, economic, medical, political, and ecological problems.
- Work individually and/or as part of an interdisciplinary team.
- Travel to consult with other professionals or attend conferences, seminars, and continuing education activities.
- Advance the frontiers of statistics, mathematics, and probability through education and research.

Statistician is the most prevalent title in use, but other titles include:

- Business Analyst
- Professor
- Economist
- Software Engineer
- Mathematician
- Risk Analyst
- Quality Analyst
- Investigator
- Pharmaceutical Engineer
- Researcher
- Data Analyst
- Project Manager.

Chapter 2

ORGANIZATION OF DATA

"Statistics is both a science and an art"- L.H.C.Tippet.

2.1 INTRODUCTION

Statistical practice covers a large number of statistical methods, and the key challenge to the statistician or experimenter is to design the collection of data so that it can be effectively converted into useful knowledge. Which statistical procedure to use, however, depend on the study objective, the study design, the nature and level of measurement of the variables and the amount of data collected. In the Indian Statistical Congress, Fisher was reported to have said:
"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

2.2 POPULATION AND SAMPLE

We all apply sampling principles in our lives. For example, to begin the day you turn on the shower and put your hand in the shower to sample the temperature. If temperature is low then you may add more hot water. As a second example, suppose you are at a bakery and wish to buy a certain sweet item. After tasting a small piece of it, you decide whether to purchase the item or not. In both the examples you make a decision and select a course of action based on a sample.

The aim of a statistical investigation is to explore certain characteristics of a group of items which may be living things or objects. The total collection of observations or measurements that are of interest to the investigator constitute the *population of observations*. The word population does not carry the usual dictionary meaning but refers simply to the set of observations relevant to a given discussion or the final reference set. The aim of the study may be to determine some characteristics of the population. For example, if the government wants to study the problems of HIV infected people, then the HIV infected people constitute the population. As another example, suppose an ornithologist is interested in the migration patterns of birds in Kerala, then all the birds in Kerala will represent the population of interest to him. It does not include birds that are native to Punjab and do not migrate to Kerala.

A population can be *finite* or *infinite*. Sometimes it is economically, physically or for some other reason, almost impossible to examine each and every item in a population. In such a situation the only possibility is to examine a subcollection of units from the population. A subcollection of items drawn from the population constitutes a *sample*. The number of units included in the sample is called the *sample size*. For example, suppose one wants to test the quality of match sticks in a match box containing 50 match sticks. It would not be practical to test all the sticks because the sticks that are tested are not suitable for re-use. So we might pick two of these sticks and test them. The 50 sticks constitute the population and the two sticks the sample.

2.3 CENSUS AND SAMPLE SURVEY

In a statistical survey we gather data or information by interviewing people or by inspecting items or in many other different ways. If in a study or survey each and every item of the population is examined, then it is called a *census survey*; otherwise it is called a *sample survey*.

The public opinion survey is very common in advanced countries, especially in the USA. The various political parties which are contesting the elections would be interested in knowing how the voters would cast their votes on the day of polling. A sample survey of voters belonging to different regions, religions, social status or other groupings is conducted. Such a sample survey is called a *Gallup poll* which is named after the American statistician George Gallup who first introduced this kind of public opinion survey in the USA.

George Gallup (1901-1984)

In 1936, his new organization achieved national recognition by correctly predicting, from the replies of only 5,000 respondents, the result of that year's presidential election, in contradiction to the widely respected Literary Digest magazine whose much more extensive poll based on over two million returned questionnaires got the result wrong. Not only did he get the election right, he correctly predicted the results of the Literary Digest poll as well using a random sample smaller than theirs but chosen to match it.

Census surveys are costly, extend over a long period and require a huge number of enumerators and supervisory staff. However sample surveys are quick, economical and can be made sufficiently reliable by giving proper training to a comparatively smaller group of enumerators. In the case of sample surveys, it may happen that the sample need not reflect all the characteristics of the population. Hence in sample surveys the results may differ from what would be obtained if the whole population had been surveyed. It is called *sampling error* and it can be reduced by increasing sample size and adopting suitable sampling procedures. But in the case of census survey one has to deal with massive data, large number of enumerators and various kinds of errors may creep in at different stages of data handling process. These errors are

called *non-sampling errors*. Examples of non-sampling errors include non-response or wrong response, imperfect measuring instrument or questionnaire, inaccurate recording, clerical errors in copying materials, computational errors in data processing etc. It is possible to reduce non-sampling errors to a great extent by using better organization and suitably trained personnel for collection and processing of data. Non-sampling error is present in sample surveys also but it can be detected and eliminated to a great extent as the volume of data and the number of persons handling the data are small. Non-sampling errors tend to increase with the increase in sample size. Hence a sample survey can give a greater accuracy than a census survey when the population size is large.

2.3.1 Types of Sampling

Like anything else sampling may be done well or badly, and it is desirable to understand the theoretical basis of sampling. Since the purpose behind taking a sample is to get information about the population, the sample must be taken in such a way that it is a representative of the population. If correct statistical principles are employed it is generally possible to design samples that will be sufficiently accurate for the purposes at hand. The nature of any given sampling procedure should depend on the level of accuracy required, as well as on the type of population from which the sample is drawn. Any type of sampling in which the sampling units selected depend on personal discretion or judgement of the investigator is called *judgement* or *purposive sampling*.

Any kind of bias in sampling can be eliminated by using a random mechanism for selecting units into the sample. To illustrate the bias in sampling let us consider the opinion surveys conducted by TV channels by inviting SMS. The viewers who send SMS constitute the sample. This sample is a biased one since some of the viewers are lethargic or ignorant or not interested in sending SMS. Moreover, some others may have missed the announcement and thus are unable to send their opinion.

Any type of sampling in which every unit of the population has a definite, preassigned chance of being selected is called *probabilistic* or *random sampling*.

Simple Random Sampling

Simple random sampling without replacement (SRSWOR) is a method of selecting n units out of the N units. The number of possible subsets of n distinct units from a set of N distict units, with $n \leq N$ is the number of combinations of N taking n at a time. This can be shown to be the following:

$$\binom{N}{n} = \frac{N(N-1)(N-2)\dots(N-n+1)}{n!} = \frac{N!}{n!(N-n)!}$$

where

$$m! = 1 \times 2 \times \dots \times m.$$

For example, $3! = 1 \times 2 \times 3 = 6$ and by convention $0! = 1$. Thus there are $\binom{N}{n}$ samples possible. If all these $\binom{N}{n}$ samples have equal chances of being selected then the sampling scheme is called a *simple random sampling without replacement*. We can prove that taking one subset of n units, giving equal chances for all subsets to be selected,

is equivalent to taking item one by one without replacement. Suppose that the units in the population are numbered from 1 to N. Select n units one by one from N units by a random mechanism which give an equal chance of selection to any number in the population not already drawn. For example, suppose we want a simple random sample of size 2 from a population having 8 units. To do this, prepare 8 identical chits with one number from 1 to 8, mix them thoroughly and select one chit randomly. Corresponding unit is selected for the sample. Next, without replacing the first chit, select a second one so that the remaning seven units have an equal chance of entering the sample. This method of selection of sample is called *lottery method*. As the lottery method is quite tedious, especially when the population size is large, we often select a sample by using random number tables.

The *random number table* is a collection of random digits. The term 'random' means that these digits are so arranged that each digit has an equal chance of occurrence. In this method the members of the population are numbered from 1 to N and n numbers are selected from one of the tables in any convenient systematic way. When using the table of random numbers, one may use them as numbers of any desired size (00 to 99 or 000 to 999 etc.). The starting point is picked arbitrarily and the required number of digits are obtained by reading across the columns in the random number table. Random numbers greater than N are ignored and continue until we obtain n random numbers. The **RAND** function in MS Excel can be used for generating required number of random numbers.

In SRSWOR the unit that has been drawn is removed from the population for all subsequent draws. At any draw, if the unit selected is replaced into the population before the next draw, then it is called a *simple random sampling with replacement* (SRSWR). In SRSWR all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn.

Stratified Random Sampling

Suppose that the population of size N is heterogeneous and can be divided into 'k' different strata of respective sizes $N_1, \ldots, N_k, \sum_{i=1}^{k} N_i = N$ such that the units within each stratum are more or less homogeneous. Usually the 'stratifying factor' may be sex, age group, education level, economic status, physical dimension, geographical region and so on. Now simple random samples of sizes $n_i, i = 1, \ldots, k$ are drawn respectively from the k distinct strata. Usually the sample size from each strata is proportional to the stratum size. The resulting sample constitutes a stratified random sample of size $n = \sum_{i=1}^{k} n_i$. Stratified sampling helps in precisions as it minimizes non-sampling errors.

Systematic Sampling

Suppose that the N units in the population are numbered 1 to N in some order. To select a sample of n units, we take a unit at random from the first k units (k is taken as the integer nearest to $\frac{N}{n}$) and every k^{th} unit thereafter. The selection of first unit determines the whole sample. This type of sampling is called systematic sampling. Such a sample is sometimes drawn from an alphabetical list of names or from a list prepared in accordance with a numerical, or other order. The sampling intervals (every k^{th} unit) must not coincide with constantly recurring characteristics in the listing of the items.

Cluster Sampling

It is useful in such cases when the complete list of primary units of the population is not available. For example, if we want to collect data from the primary school students of Kerala we select a few primary schools first and then the complete enumeration of students in the school is done. Note that the list of primary school students may not be available but the list of primary schools in Kerala may be available. The main advantage of cluster sampling is to reduce costs.

There are other types of sampling such as multistage sampling, sequential sampling, quota sampling etc. Since these topics are outside the purview of this text we are not discussing them here.

Simple Random Sampling Versus Other Sampling Schemes

When deciding which sampling plan to use, the investigator must consider the efficiency of the scheme. It has been noted that a stratified sample yields more reliable results (that is, its sampling error is smaller) than does a random sample of the same size if the population is non-homogeneous and stratification is properly done. Cluster sampling may be expected to yield less reliable results than simple random sampling for samples of the same size. The efficiency of a sampling scheme shall be judged in relation to both reliability and unit cost. Thus, a geographic cluster sample consisting of, say, 20 locations in a larger state may have a lower cost per sampling unit than a random sample of the same size with the units scattered here and there about the state. The difference in unit cost may be so large that the cluster sample may be made enough larger than the random sample so that the former will yield more reliable results than the latter for the same expenditure. However, no general statement can be made to the effect that more reliable or less reliable results may be had from a systematic sample than from a random sample of the same size. The conditions under which systematic selection is to be preferred to simple random sampling, or vice versa, are too to be considered.

2.4 PRIMARY AND SECONDARY DATA

Data can be primary or secondary. Data is termed *primary* when it is collected by the investigator himself or his men for the present purpose. If the data already available is taken from published reports or from other agencies for the present use of the investigator, then it is termed as *secondary* as far as the user is concerned. The meteorological department regularly collects data on different aspects of the weather and climate such as amount of rainfall, humidity, maximum and minimum temperature of a certain place. These constitute primary data to the meteorologists. If this data is used by somebody else for some other purpose then it is a secondary data to them.

Secondary data is economical as it is readily available. Extra care should be taken when using a secondary data as it may be collected by some other person for a different purpose at a different time period. It is always recommended to use the data collected by reliable agencies as secondary data. The main sources of secondary data are research publications, project reports, summarized census report, monthly abstracts of statistical organizations, various publications of UN etc.

2.5 COLLECTION OF PRIMARY DATA

Primary data is collected either through census or sample survey. There are four different stages in a survey - planning, execution, analysis and preparation of the report. At the planning stage the investigator must have a clear picture regarding the objective of the study, *sampling frame* (that is, a complete list of the sampling units in the population), method of sampling, way in which data to be collected. At this stage a *pilot survey* is sometimes conducted on a small number of units before the actual commencement of the original survey in order to understand and rectify the lapses if any.

The usual procedure of collecting information is through a *questionnaire* or a *schedule*. The relevant aspects to be collected are put in the form of questions in a questionnaire or schedule. A distinction is made between a questionnaire and a schedule. The answers to the questions in a questionnaire are entered by the informant or respondent himself, whereas in a schedule the answers are recorded by the investigator or an enumerator on behalf of the respondent.

There are direct and indirect ways of collecting data. Direct personal interview or observation is the most common way of data collection. With the prevalence of electronic media, data collection through email and SMS is also very common. Each method has its own merits and demerits.

The conclusions based on an inadequate data will be misleading. Therefore it is necessary to ensure that all the relevant aspects related to the phenomenon under investigation are asked in the questionnaire or schedule. The questions must be unambiguous and they are to be put in a logical order. A self-explanatory title should be given to a questionnaire or schedule. The instructions and definitions should be concise. The enumerator and informant should never be in doubt as to what information is desired and what terms or units are to be used. Statisticians having much experience and wisdom are needed for drafting a good questionnaire or schedule.

Completed questionnaires or schedules are edited by deleting data that are obviously erroneous. Further steps are data processing, analysis of the data, interpretation of results and report writing.

Example 2.5.1. Self-help groups are active among women in rural areas of Kerala. Prepare a questionnaire for assessing the savings pattern of rural women in Kerala.

Questionnaire on Pattern of Savings of Rural Women in Kerala

1.
   (a) District . . . . . (e) Date of filling the form . . . . . .
   (b) Panchayat . . . . . (f) Caste. . . . . .
   (c) Ward No. . . . . (g) Name of the respondent. . . . . .
   (d) House No. . . . . (h) Age of the respondent. . . . . .

2. Source of Income of the respondent
   Wages ☐ Salary ☐ Agriculture ☐ Industrial Unit ☐
   Remittance from abroad ☐ From other income earning members in the family☐

3. Do you hold any

   (a) deposits in Postal Savings Bank? Yes/No.

   (b) national Savings Certificates or Treasury Bonds? Yes/No.

   (c) insurance policies? Yes/No.

   (d) deposits with a private banker or shop-keeper? Yes/No.

   (e) membership in Group Credit Deposit Scheme? Yes/No.

4. Total amount of investments in an year:

5. If you do not,

   (a) is it because you have no margin for saving? Yes/No;

   (b) or because you prefer to hold them in cash? Yes/No;

   (c) are you saving the money to buy land? Yes/No;

   (d) are you saving the money to purchase or build a house? Yes/No;

   (e) do you prefer to purchase gold and jewelry? Yes/No;
   If yes, is it your customary practice to purchase gold every year? Yes/No.

   (f) do you prefer to lend money? Yes/No;
   If yes, what rate of interest do you get? .....

6. (a) Do you know that there is a Postal Savings Bank? Yes/No.

   (b) Do you know the rate of interest given by it? Yes/No.

7. If you do not hold deposits in a Postal Savings Bank, is it because

   (a) there are no local facilities? Yes/No;
   If yes, will you use facilities, if made available? Yes/No.

   (b) of low rate of interest? Yes/No;
   If yes, what rate of interest do you expect? .....

8. If you do not hold any National Savings Certificates, Treasury Bonds etc., is it because

   (a) there are no local facilities? Yes/No;
   If yes, will you use facilities, if made available? Yes/No.

   (b) you think they are inconvenient for holding? Yes/No;

   (c) there are difficulties of encashing? Yes/No;

   (d) of low rate of interest? Yes/No;
   If yes, what rate of interest do you expect? .....

9. If you do not hold deposits in Cooperative Banks, is it because

   (a) there are no local facilities? Yes/No;
   If yes, will you use facilities, if made available? Yes/No.

   (b) you do not trust them? Yes/No;

(c) of low rate of interest? Yes/No;
If yes, what rate of interest do you expect? .....

10. If you have no insurance policies, is it because

(a) of superstition? Yes/No;

(b) there are no local facilities? Yes/No;

(c) it is too complicated? Yes/No;

(d) there are difficulties of paying premia regularly? Yes/No;

(e) the money cannot be readily realized? Yes/No.

2.6 TYPES OF VARIABLES

In a statistical survey we gather information on several characteristics. The values of these characteristics vary from one unit to another. We call these characteristics as *variables*. There are two kinds of variables - qualitative and quantitative.

2.6.1 Qualitative Variables

When the characteristic being studied is nonnumeric, it is called a *qualitative variable* or an *attribute*. Examples of qualitative variables are gender, religious affiliation, eye colour etc. Although a qualitative variable has no numerical value, it is possible to assign numerical values to a qualitative variable by giving values to each quality. For example, one's preference about how pleasant is a perfume to her can be assigned a number from 0 to 10 with 0 being unpleasant and 10 being highly pleasant; tolerance to noise pollution can be measured by assigning a number, say from 0 to 100, satisfaction level can be assigned a number, colours can be assigned various numbers as labels etc. Qualitative variables are sometimes termed *categorical variables* or *descriptive variables*. A categorical variable which can assume only two values (such as the variable 'gender') is known as a *dichotomous variable*.

There are two sub-groups of categorical variables: nominal and ordinal variables. *Nominal variables* are categorical variables which can only be classified and counted. Examples include gender, religious affiliation, marital status, identification code etc. A numeric code may be used as a label for representing the observations of a nominal variable so as to facilitate recording and computer processing of the data. For example, the data for the type of school can be 1 for boys, 2 for girls and 3 for coed. However, it is important to remember that the numeric values 1, 2 and 3 are simply labels used to identify the type of school.

Ordinal variables are categorical variables which have an inherent or explicit ordering. For example, the rating of a statistics professor's lecture by the students can be obtained under the heads: strongly like, like, indifferent, dislike, strongly dislike. However we are not able to distinguish the magnitude of the differences between groups. If we assign 5 for strongly like, and 1 for strongly dislike we cannot conclude that strongly like is necessarily 5 times as high as strongly dislike. We can conclude only that a rating of strongly like is better than a rating of strongly dislike. Ordinal measurements describe

order, but not relative size or degree of difference between the items measured. Unlike nominal data here the data can be ordered and assign ranks.

Even if the data on qualitative variables are numeric, arithmetic operations such as addition, subtraction, multiplication and division do not make sense and are inappropriate.

2.6.2 Quantitative Variables

A variable is *quantitative* if it can be reported on a numerical scale where all the basic operations such as addition, multiplication etc. can be meaningfully interpreted. It is also called *numerical variable*. Height of students (in inches), waiting time at the bus stop (in minutes), family size, the number of accidents in a day are examples of numerical variables.

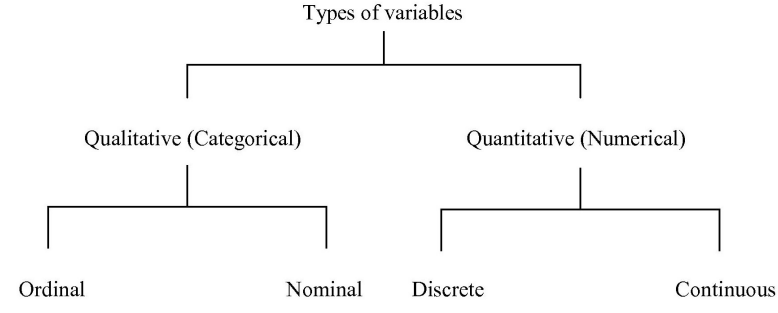
We can classify numerical variables further as continuous or discrete. If the variable can assume any numerical value over an interval or different intervals of the real line or over a continuum of points then it is said to be *continuous*. Weight, area, time etc. are examples of continuous variables. The observations on a continuous variable can be measured to any degree of accuracy on a numerical scale.

A *discrete variable* can assume either a finite or a countable number of values. Number of bedrooms in a house, family size and number of telephone calls received in an hour, score of a multiple choice objective type examination etc. are examples of discrete variables.

For certain types of data on numeric variables, the ratio of two data values is not meaningful however their difference is meaningful. Temperature measured in degrees Fahrenheit or Celsius constitutes an example. We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees. Ratios between the values are not meaningful in this case and hence operations such as multiplication and division cannot be carried out directly. But ratios of differences can be expressed; for example, one difference can be twice another. If the ratios of data values of a certain numeric variable is not meaningful however the difference or interval between data values may be meaningful, then the measurement scale is termed as *interval scale*. In addition to the difference of data values, if the ratio of values is also meaningful then the scale of measurement is known as *ratio scale*. For example, height, weight, distance, area, volume and time are ratio scale measurements. A person who weighs 80 kilograms is twice as heavy as a person who weighs 40 kilograms, unless some origin other than zero is used. Temperature measured in Kelvin scale is a ratio scale. An important feature of the ratio scale of measurements is that it has an absolute zero point.

Remark 2.6.1. The four measurement scales- nominal, ordinal, interval and ratio- have been mentioned in the increasing order of refinement. Note that any measurement in one scale can be transformed to a measurement in the lower scale. For example, interval data can be converted to ordinal (or even nominal) data. However the reverse process is not possible. Hence it is important that measurement be made at a level not lower than that is needed or required for analysis. On the other hand, recording data at

a scale higher than that needed for analysis is a waste of time, energy and resources. Thus the investigator must take care of these aspects in the design of the experiment or study.



The preliminary steps in a statistical study which involves the collection of data may be designated as follows:

(i) Planning the study.

(ii) Devising the questions and making the schedule.

(iii) Deciding the type of sample, if the enumeration is not to be a complete one.

(iv) Determining the level of measurement of each of the variables.

2.7 EXERCISES

1. Which of the following measures involve nominal scale?
   (a) The test score on an exam; (b)The class rank of a student;
   (c) The number on an automobile license plate; (d) The weight of a fish.

2. Which of the following measures involve ordinal scale?
   (a) The place a person finishes in a competition; (b)The height of a flagpole;
   (c) A football player's uniform number; (d) The family size of a student.

3. As part of an environmental impact study, fish were captured, tagged, and released. The following information was recorded for each fish:
   sex (0=female, 1=male), length (cm), maturation (0=young, 1=adult), weight (gm).
   The scale of these variables is:
   (a) nominal, ratio, nominal, ratio; (b) nominal, interval, ordinal, ratio;
   (c) nominal, ratio, ordinal, ratio; (d) ordinal, ratio, nominal, ratio.

4. Suppose we subdivide the population into at least two sub groups (such as by marital status) and then draw a random sample from each of the groups. This type of sampling scheme is called
   (a) cluster sampling; (b) stratified sampling;
   (c) systematic sampling; (d) none of these.

5. A random sample of 1500 households in New Delhi was selected and several variables were recorded for each household. Which of the following is not correct?

(a) Household total income is a ratio scaled variable;

(b) Household income rounded to the nearest 100 can be treated as a continuous variable even though it is discrete;

(c) Socio-economic status was coded as 1=low income, 2=middle income, 3=high income and is an interval scaled variable;

(d) The primary language used at home is a nominal scaled variable.

6. Judge whether the following statements are true or false.

(a) A population is the set of all elements of interest in a particular study.

(b) A sample may be larger than the population.

(c) Sample survey is free of non-sampling errors.

(d) Sampling error is present in both census and sample surveys.

(e) Non-sampling error is comparatively low in sample surveys.

(f) The speed of a train measured in miles per hour is ordinal.

(g) Scores of a cricket batsman during the last one year is an example of a discrete variable.

(h) Quantitative variables must be measurable either in interval scale or in ratio scale.

(i) Nominal and ordinal data must be numeric data.

(j) The data on cause of death of persons is quantitative.

7. Classify the following variables and indicate the measurement scale that is appropriate for each.

(a) Amount of income tax paid every year for 30 years.

(b) Number of siblings (brothers and sisters) in households in Kerala in 2011.

(c) The time it takes an employee to drive to office.

(d) The occupation type of adult males in Kochi Corporation in 2011.

(e) The annual rainfall in Kerala for 1950 to 2012 period.

(f) The breaking-strength of a certain type of cable in 10 test runs.

(g) The number of rubber trees on an acre of land in a sample of 10 such acres.

(h) Brand of TV in the households in New Delhi.

(i) A person's nationality enumerated among tourists visiting Kumarakom.

(j) Possible modes of travel for travelling from Kochi to Thiruvananthapuram.

(k) Mode of payment of insurance premium by policy holders (cash, cheque, credit card).

(l) Annual sales of automobiles by different companies.

(m) Opinion about performance of present Central Government taken from among the people of Kozhikode Corporation on a particular day.

8. A Gallup poll investigated whether adults in Mumbai preferred staying at home or going out as their favourite way of spending time in the evening. Out of the 500 adults the majority of adults (70%) indicated that staying at home was their favourite evening activity.

   (a) What is the population of this study?

   (b) What is the variable being used?

   (c) Is the variable being studied qualitative or quantitative?

   (d) What was the size of the sample used?

9. Students are asked to fill-up a course evaluation questionnaire upon completion of their course. They have to put numeric codes 1, 2 ,3, 4, 5 against each question where 1=Poor, 2=Fair, 3=Good, 4=Very Good and 5=Excellent. Comment on the nature of variables and scale of measurement used in this study.

10. Draft a suitable questionnaire to study the spending pattern of students studying in your college.

11. It is required to determine whether extracurricular activities adversely affect the academic performance of students. Draft a suitable questionnaire for this study.

12. Draft a questionnaire to study the mobile phone and internet usage of students in your college.

13. Street begging is a disgrace to our country and requires immediate solution for the problem. Prepare a questionnaire for conducting a socio-economic survey for investigating the forces which drive people begging and the possible remedies to abolish begging.

14. Design a survey to compare the yield of different varieties of rubber trees.

15. A manufacturing company has come up with a new motorcycle. The company claims that it will capture a large share of the young adult market.

   (a) What data would the company must see before deciding to invest substantial funds to introduce new product into the market?

   (b) How would you expect the data mentioned in part (a) to be obtained?

16. Suggest a possible source of bias in the following samples:

   (a) A basket of lemon is sampled by taking a handful from the top.

   (b) A sample of household survey which includes houses having road access.

17. Describe the principal steps in a survey.

Chapter 3

PRESENTATION OF DATA

"The purpose of statistical methods is to simplify great bodies of numerical data"- A.E.Waugh.

3.1 INTRODUCTION

The process of collection of data was described earlier. The data so collected is known as *raw data*. The information contained in the raw data spreads over several sheets. Nowadays handling of massive raw data is not a problem due to the easy availability of computers and packages like MS Excel, SPSS, Minitab etc. A reliable data must be complete, consistent, accurate and homogeneous in respect of unit of measurement. This has to be ensured while editing the data and before entering the data into the computer. Tables, graphs and diagrams are useful in the presentation of data. They give better grasp of the information at a glance.

3.2 TABULAR PRESENTATION

The relative importance of all the variables is not the same and hence it is necessary to break down the information according to certain salient features or characteristics. Such a process is called *classification and tabulation*. The characteristics which form the bases of classification should be first determined. In respect of data on population census, the main factors are sex, age-groups, marital status, geographical locations, literacy etc. Workers in a factory may be classified as male and female, skilled and unskilled etc. Obviously it would not be possible to include all variables in a single table. Tabulation is called single, double or manifold according as one, two or many variables are used for classifying the information.

A good table should have -

title and footnotes;
captions (column heading) and stubs (row heading);
neatness and accuracy;
logical order.

The main objective of tabulation is to present the data in a compact and concise form which facilitate comparisons and study of relationships between different variables.

Table 3.1: Size of households in Kerala

Size of households	Total	%	Rural	%	Urban	%
One member	187,102	2.8	145,100	2.9	4 2,002	2.5
Two members	457,651	6.9	3 44,194	7.0	113,457	6.9
Three members	8 64,207	13.1	6 38,909	12.9	2 25,298	13.6
Four members	1,934,787	29.3	1,439,661	29.1	4 95,126	30.0
Five members	1,389,884	21.1	1,056,917	21.4	3 32,967	20.1
Six to eight members	1,340,833	20.3	1,018,676	20.6	3 22,157	19.5
Nine members and above	4 20,742	6.4	299,093	6.1	121,649	7.4
Total number of households	6,595,206	100	4,942,550	100	1,652,656	100

Source: Table H-5 India : Census of India 2001

Table 3.1 is an example of a table which shows the size of households in Kerala. Most of the studies contain several numerical variables and tabulation of data naturally leads to a frequency table.

3.2.1 Frequency Table

Consider a discrete variable that assumes only a few distinct values such as number of births/deaths in a city in a day and suppose that it is recorded for a large number of days. Obviously, each value of the variable will repeat a certain number of times. If the order in which the observations occurred is immaterial then the data can be summarized in a table that shows the number of times each value of the variable occurred. Such a table is called a frequency table and Table 3.1 is an example of such a table.

When there is a relatively large number of observations on a continuous variable then partial summarization of data without losing any information of interest is possible. From a look at the raw data not much inference can be made. For example, suppose we have a data on cholesterol level (mg/dL) of 250 patients ranging from a low value of 141 to a high value of 296. We therefore condense the data by forming a frequency table. This can be done in different stages. Firstly, a set of non-overlapping consecutive intervals is set up so as to include all the values within its range. The intervals are called '*class intervals*'. The class intervals can be defined arbitrarily by the user. There is no hard-and-fast rule to determine the number of class intervals. As a general guideline it is recommended to use 5 to 10 class intervals. For large data sets it may require more classes. The greater the number of observations, the more classes we may have. After deciding the number of classes the next step is to determine the class width. The relationship between the number of classes and the class width can be written as follows:

$$\text{Class width} = \frac{\text{Largest data value - Smallest data value}}{\text{Number of classes}}.$$

Assume that we want eight class intervals for the above mentioned data on cholesterol level, then

$$\text{class width} = \frac{296-141}{8} = \frac{155}{8} = 19.375.$$

For the sake of convenience, we may take 20 as the class width and the class intervals as [140, 160), [160, 180),...,[280, 300). Here the notation [140, 160) means that a value of 140 is included and a value of 160 is excluded from that interval. The *class limits*

determine the range of values that are grouped into each class. For the class interval 140-160, the lower limit is 140 and the upper limit is 160. The midpoints of the class intervals are called *class marks*. The class mark of the class interval 140-160 is 150. Secondly, a count is made of the number of values which fall within each of the intervals. The counts are called *frequencies*. For example, a value of cholesterol level 183 is recorded by a tally mark corresponding to the class interval 180-200. Complete the process of recording a tally mark for each value in the data against the class interval in which it falls. The number of tally marks corresponding to each of the class intervals is then counted to obtain frequencies. The resulting table is now called a *frequency table*.

If the frequencies are replaced by the proportion of values (obtained by dividing class frequency by total frequency) falling within each class interval, then a *relative frequency table* is obtained. Relative frequencies are always between 0 and 1, and their sum is always equal to unity.

Example 3.2.1. The following observations give the yield of paddy in kilograms from 50 experimental plots in a research station:

58.3	56.2	62.3	44.2	46.3	47.2	52.5	53.7	51.9	54.7
40.4	37.4	45.5	44.8	35.1	45.1	51.7	48.6	50.8	47.4
39.8	46.3	50.9	43.8	49.8	56.4	61.1	57.8	40.5	46.2
45.5	46.7	41.3	52.7	50.8	59.5	46.1	54.3	56.7	50.1
57.4	58.3	55.9	47.2	36.8	51.3	54.1	50.6	44.4	42.7

If we choose a class width of 5 and take the class intervals as [35, 40), [40, 45), [45, 50), [50, 55), [55, 60), [60, 65), then the frequency table obtained is given in Table 3.2.

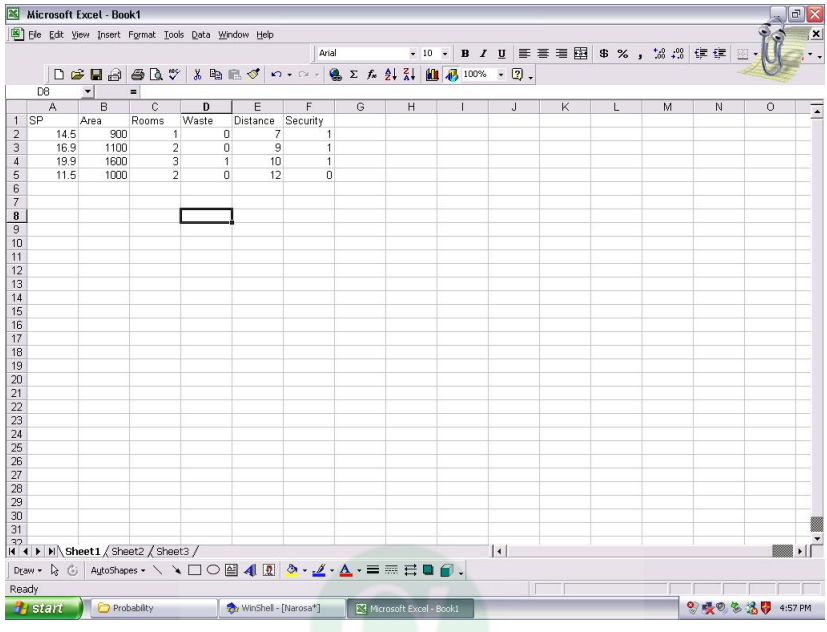
Table 3.2: Frequency table

Class	Tally marks	Frequency	Relative frequency
$[35,40)$	𝍸	4	$\frac{4}{50} = 0.08$
$[40,45)$	𝍸 𝍸 | | |	8	$\frac{8}{50} = 0.16$
$[45,50)$	𝍸 𝍸 | | |	13	$\frac{13}{50} = 0.26$
$[50,55)$	𝍸 𝍸 | | | |	14	$\frac{14}{50} = 0.28$
$[55,60)$	𝍸 | | | |	9	$\frac{9}{50} = 0.18$
$[60,65)$	| |	2	$\frac{2}{50} = 0.04$

In grouping the individual observations into different class intervals, the identity of individual values and the order in which the observations occur will become irrelevant. In constructing frequency tables, we may lose the accuracy of data. To see how this accuracy can be lost, consider the first class of 35-40 in Table 3.2. The table indicates that there are four values in that class, but there is no way to determine from the table exactly what those original values are. We cannot regain the original list of 50 values from Table 3.2.

Care must be taken to ensure that all values fall into one and only one class. For example, suppose we have an integer data and are interested in taking class lim-

Figure 3.1: MS Excel Spreadsheet



its as 0-10, 10-20, 20-30 and so on. The data values 10 and 20 appear to belong to two classes. So in this case it is better to write the class limits as 0-9, 10-19, 20-29 because with integer data there are no values in the intervals 9-10 and 19-20. If the variable under study is a continuous one then there may be values in the intervals 9-10 and 19-20. To avoid confusion, it is better to write the above class intervals as 0-under 10, 10-under 20, 20-under 30 and so on. If in a frequency table the upper limit of a class interval is the lower limit of the next class interval then it is called a *continuous frequency table*. In the continuous frequency table a data value equal to the lower limit of the class interval is included in that interval and a data value equal to the upper limit is excluded from that interval. This type of class intervals are called *exclusive class intervals*. In *inclusive class intervals* the upper limit of a class is not the lower limit of the next class.

Remark 3.2.1. Summarizing the data generally involves a compromise between accuracy and simplicity. A frequency table with too few classes is simple but not accurate. A frequency table with too many classes is more accurate but not easy to understand. The rule of equal class intervals is inconvenient when data are spread over a wide range but are highly concentrated in a small part of the range with relatively few numbers elsewhere. Using smaller intervals where the data are highly concentrated and larger intervals where the data are sparse help to reduce the loss of information due to grouping. Tabulations of income, population and some such characteristics in official reports are often made with unequal class intervals. It is sometimes impossible to avoid open-ended intervals such as “180cms or taller”.

Once the raw data is entered in a computer spreadsheet it is easy to form a frequency table with the help of packages and there is no need of tally marks. We can make use of **Histogram** tool in MS Excel to form the frequency table. We shall study this tool in the next section. Construction of manifold tables can be very easily done with the help of SPSS, Minitab etc. Nowadays computer packages will carry out the analysis of the data irrespective of its size and there is no need of forming frequency table for the purpose of analysis. Hence frequency tables are now used only for presentation of data in a summarized form. Figure 3.1 depicts the way in which a data has to be entered in a spreadsheet of MS Excel. The data entered in the spreadsheet reports information on flats sold in Kochi city last year. The variables considered are,

SP	:	Selling price (in lakhs of rupees)
Area	:	Area in square feet
Rooms	:	Number of bedrooms
Waste	:	Facility for waste disposal (Yes=1, No=0)
Distance	:	Distance from the center of the city (in kilometers)
Security	:	Round the clock security (Yes=1, No=0).

3.3 GRAPHIC PRESENTATIONS OF DATA

The graphic display of data and information has a very long history, but the age of modern statistical graphs only began around the beginning of 19th century. Graphs are important as they convey a large amount of information in a compact and attractive way. Pictorial presentation has the advantage that even an uneducated person can assimilate the information by looking at a chart or graph. Pictures can create a stronger visual impact on the viewers. A person who is imaginative enough can present the information pictorially in several ways. Graphical excellence is achieved when a viewer can get the most accurate and comprehensive picture of the underlying information in the data in the shortest possible time. Graphs can reveal patterns, trends or anomalies, constancy or variation. In comparison to tables and textual forms, charts and graphs are easily understood by a layman. An old adage says that “one picture is worth one thousand words”.

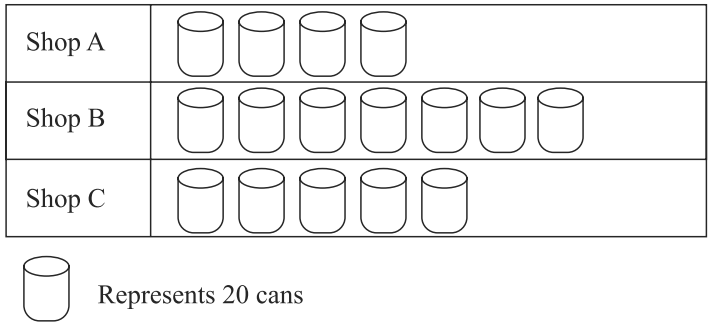
Each chart, like each table, should have a title, which should state clearly what the chart intends to show. A chart must show source reference, scale labels or values and legend. All lettering on a chart should be legible. When two or more graphs appear on a chart, each should be clearly identified.

In statistical graphics, William Playfair (1759-1823) introduced the bar chart and the pie chart. He was a Scottish engineer, who is considered the founder of graphical methods of statistics. Important graphical techniques are described here very briefly.

Pictographs

A graph in which the size of the object in the picture indicates the relative size of the variable the object represents. In pictograph, each picture or symbol represents a fixed quantity of the variable. For example, the population size of a country is shown by man, milk production by milk cans etc. Pictographs are inaccurate as they can represent only approximate values. The sale of milk in three shops is shown using a pictograph in Figure 3.2.

Figure 3.2: Pictograph



Cartograms

Statistical data classified according to different geographical regions can be represented with the help of a suitable map. Such a representation is known as a cartogram. For example, the data relating to the seats won by two political parties during the last parliament election of a particular country can be shown in a map of the country by shading the constituencies won by a particular party with one colour and those of the other party with a different colour.

Bar Charts

Bar diagrams are the most commonly used pictorial presentation of data involving categorical variables and discrete quantitative variables. Bar diagrams are of the following types:

1. Simple bar diagram;
2. Multiple bar diagram;
3. Component or subdivided bar diagram.

A bar chart consists of either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Bars are of equal width and their heights are proportional to the frequencies or quantities of the variables. Simple bar charts are used to exhibit changes in magnitudes of the variable over time (chronological) or region (geographical).

A bar chart arranged from highest to lowest incidence is called a *Pareto chart*.

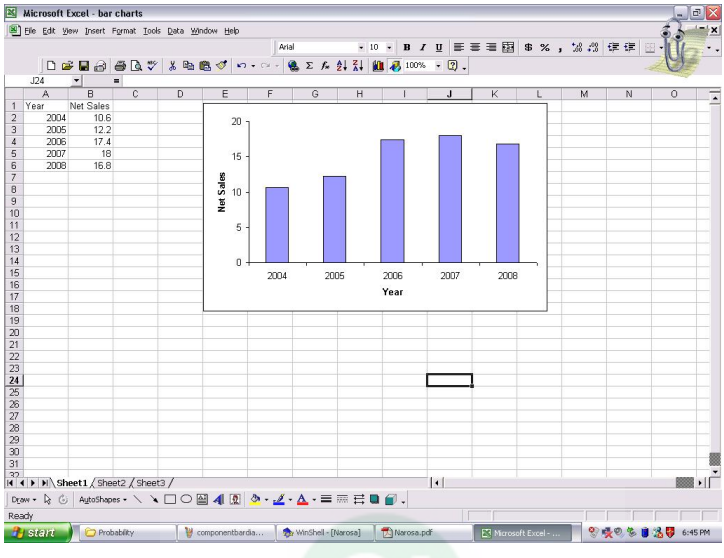
Example 3.3.1. A particular trading company sells fashion apparel for men and women. Listed below are the net sales of the company from 2004 to 2008.

Year	:	2004	2005	2006	2007	2008
Net sales	:	10.6	12.2	17.4	18	16.8

The bar chart depicting the net sales over the time period is shown in Figure 3.3.

To draw the simple bar diagram using MS Excel the menu commands are as follows:

Figure 3.3: Simple bar diagram



Enter the data in the spreadsheet of MS Excel as shown in Figure 3.3. Then click **Insert**→ **Chart** → Select appropriate chart from **Chart type** and **Chart sub-type** → **Next** → Click in **Data Range** and then select the cells (drag) from A1 to B6 → Click at Series in **Columns** → **Series**→ Remove 'Year' from Series → Click in **Name** and select cell B1 to enter the name of the series as 'Net sales'→ Click in **Values** and drag cells from B2 to B6 → Click in **Category (X) axis labels** and drag cells from A2 to A6 → **Next**. Now different chart options will appear and you can include the listed options as you like. Finally click **Finish**.

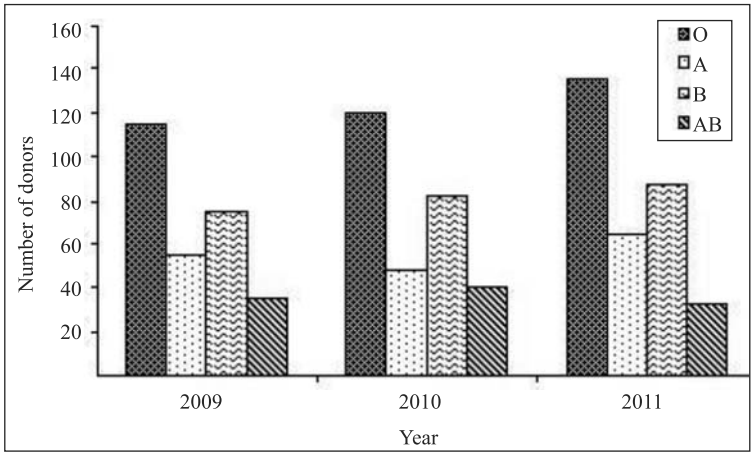
Multiple bar diagrams are used for multidimensional comparison. For comparison of magnitudes of one variable in multiple aspects, or for comparison of magnitudes of several variables, a group of bars placed side by side is used. The bars are to be distinguished by shading or colouring to show the aspects or variables represented. If you are preparing a multiple bar diagram, remember to present the information in the same order in each grouping. Figure 3.4 is a multiple bar diagram.

Example 3.3.2. The data on donation of various types of blood by the students in a college during three consecutive years is shown in Table 3.3. The data represented using a multiple bar diagram is shown in Figure 3.4.

Table 3.3: Number of blood donors

Year	O	A	B	AB
2009	115	55	75	35
2010	120	48	82	40
2011	135	65	87	33

Figure 3.4: Multiple bar chart



Sometimes it is necessary to show the break up of one variable in several components so that each bar is subdivided into several components. The heights of bars represent the aggregate magnitude of the variable. The resulting bar diagram is called component bar diagram.

Example 3.3.3. The data on the number of tourists from America, Britain, France, and other countries who visited various tourist centres in India during the last New Year day is shown in Table 3.4. The data given in Table 3.4 represented using component bar diagram is shown in Figure 3.5.

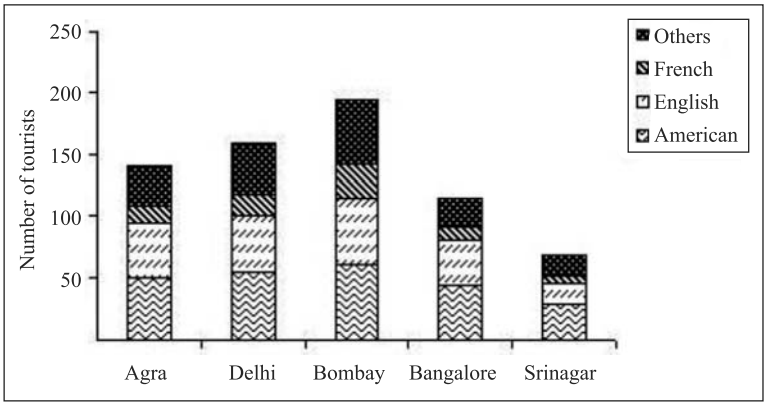
Table 3.4: Number of tourists

	American	English	French	Others
Agra	50	45	15	30
Delhi	55	47	17	41
Bombay	62	53	27	53
Bangalore	45	36	12	22
Srinagar	30	16	7	17

Instead of taking the magnitudes of the variable and components, the magnitudes expressed in percentages can be considered and a component bar chart can be drawn. In this case all bars are drawn with equal heights each representing a total of 100.

The MS Excel commands for drawing multiple and component bar charts are similar to that of simple bar chart except a few clicks which do not require any explanation. To fill the different bars or component of bars of the diagram as shown in Figure 3.4 and Figure 3.5 one may proceed as follows:

Figure 3.5: Component bar diagram



Right click the mouse after placing the cursor at one of the bars of the diagram → Click **Format Data Series**→ Select the desired colour from **Area**→ Click **Fill effects**→ **Pattern**.

Now select the desired pattern from the listed patterns in the window and click **OK**. Repeat this until all the bars or components are filled with different patterns.

Pie Charts

Pie charts are especially useful for illustrating nominal data. Here the area inside a circle is partitioned into several sectors such that each sector has an area proportional to the percentage of the total quantity. Suppose a category accounts for 5% of the total, then the angle of the sector representing this category is 360 × 0.05=18 degrees. The calculation of angles is shown in the following example.

Example 3.3.4. Data on responses of 150 viewers of a particular TV program is given in the first two columns of Table 3.5.

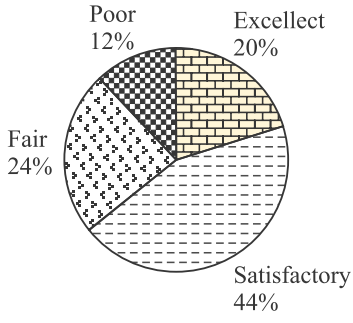
The calculation of angles is shown in the last two columns of Table 3.5 and the corresponding pie diagram is displayed in Figure 3.6..

Table 3.5: Responses of viewers of a TV program

Response	Frequency	Relative frequency	Angle in degrees
Excellent	30	0.20	360×0.20=72
Satisfactory	66	0.44	360×0.44=158.4
Fair	36	0.24	360×0.24=86.4
Poor	18	0.12	360×0.12=43.2
Total	150	1	360

If more than one set of data have to be represented simultaneously using pie charts then the area of each chart must be proportional to the total quantity of the cor-

Figure 3.6: Pie chart showing responses of a TV programme

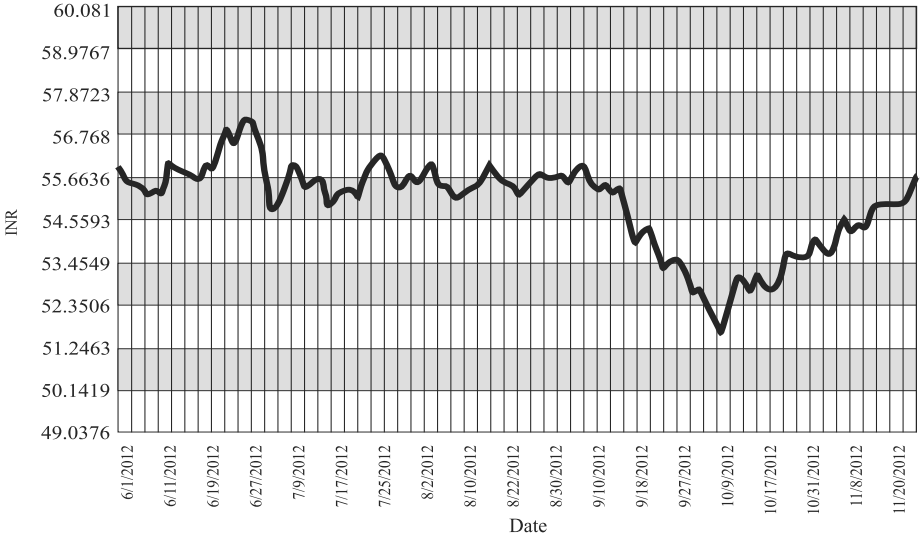


The procedure for drawing pie chart in MS Excel is similar to that of bar chart.

Time Series Charts

The value of certain variables such as price, sales, production of wheat, temperature, population size and blood pressure which vary with time and it is required to display their changes by years, months, weeks, days or seconds. A set of values of a variable over a period of time arranged chronologically is called a *time series*. Whenever a chronological data is depicted by means of a curve, the time periods or time points are shown on the horizontal axis; the other series, the value of the variable which varies with time, is placed on the vertical axis. A time series chart showing the exchange rate of Indian rupee against US dollar is shown in Figure 3.7.

Figure 3.7: Time series chart showing exchange rate of Indian rupee against US dollar



The steps for drawing a time series graph is similar to that of a bar chart except you must select **XY(Scatter)** from Chart-type dialogue box.

3.3.1 Stem and Leaf Plot

Stem and leaf plot is a way of representing a data corresponding to a quantitative variable. It is a mixture of a graph and a table. In this method a numerical value in the data set is separated into two parts, namely, the stem and the leaf. The stem consists of the leftmost digits and the leaves consist of the remaining digits. A stem and leaf diagram is constructed as a series of horizontal rows of numbers. The first number of each row is the label of that row and called the *stem*. The remaining numbers in a row following the stem are called the *leaves*. Stem and leaf plot reveals whether there is any tendency for the leaves to cluster around a particular stem or stems. This plot is similar to a bar chart or histogram, but contains more information. It is illustrated with the help of the following example.

Example 3.3.5. The following data give the amounts (in rupees) spent on groceries by 42 housewives during a week:

22	42	45	18	33	32	40	33	18	45	48	16	32	15
37	30	46	22	35	34	58	25	51	66	37	29	44	55
25	33	62	41	39	28	35	40	25	23	64	26	58	60

The stem and leaf plot of the above data is shown below:

1	5688
2	223555689
3	022333455779
4	001245568
5	1588
6	0246

We note that the values in the data have first digits of 1,2,3,4,5,6 and we let those values become the stem. In the above plot the column of numbers to the left of the vertical line is the 'stem' while to the right of the line are 'leaves'. The first row corresponds to the values 15, 16, 18, 18 and similarly the other rows.The main advantages of using a stem-and-leaf plot are that it shows the general shape of the data (like a bar chart or histogram), and that all the values in the data can be recovered (to the nearest leaf unit). For example, we can see from the plot that there is only one value of 30, and three values of 25.

Example 3.3.6. The purity of 16 gold coins were examined and the readings obtained in carat are given below. Construct a stem and leaf plot.

21.8 20.0 22.4 22.8 20.3 22.6 22.3 19.2

20.1 21.7 23.2 21.9 23.2 21.9 23.5 21.8.

The stem and leaf plot of the above data is shown below:

19.	2
20.	013
21.	78899
22.	3468
23.	225

Occasionally one or two scores may be far distant from the rest of the data, in which case it is not realistic to continue the stems all the way down to those values. These extreme values are shown by listing them on the high side or low side of the data.

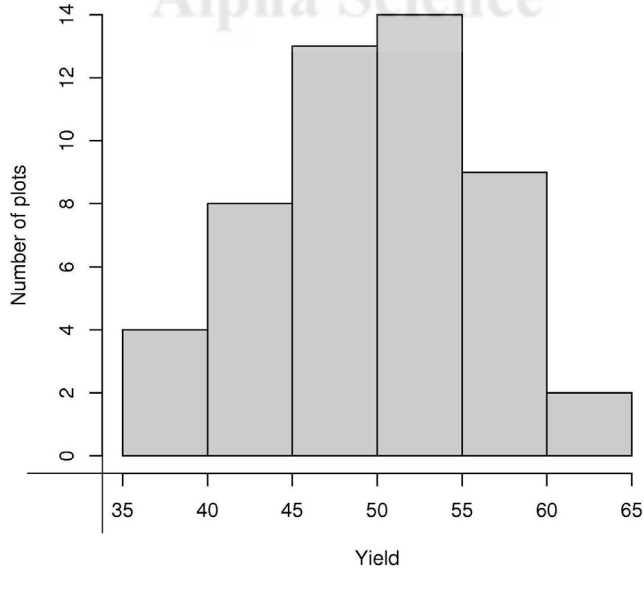
3.3.2 Graphic Presentation of Grouped Frequencies

Histogram

A graphic representation of grouped frequencies is called a histogram. It is a graph in which the class intervals are marked on the horizontal axis and the frequencies on the vertical axis. The class frequencies are represented by the heights of bars, and the bars are drawn adjacent to each other. If we replace frequencies by relative frequencies, then the height of the bar represents the proportion of observations in each class.

Example 3.3.7. The frequency table formed in Example 3.2.1 can be represented by the histogram in Figure 3.8.

Figure 3.8: Histogram



How to use the Histogram tool in Excel?

You can draw a histogram by using the Histogram tool of the **Analysis ToolPak** in Microsoft Office Excel. To draw a histogram, you must enter the data into a column of the Excel worksheet. Then you must give bin numbers in the next column. *Bin numbers*

are the numbers that represent the intervals that you want the Histogram tool to use for measuring the input data in the data analysis.

When you use the Histogram tool, Excel counts the number of data points in each data bin. A data point is included in a particular data bin if the number is greater than the lowest bound and equal to or less than the largest bound for the data bin. If you omit the bin range, Excel creates a set of evenly distributed bins between the minimum and maximum values of the input data.

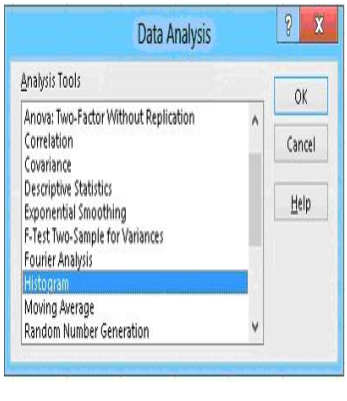
Before you try the Histogram tool, you have to make sure that the Analysis ToolPak is installed in your computer. To make sure that the Analysis ToolPak Add-ins is installed, follow these steps:

1. Click Excel **Add-Ins**.
2. In the Add-Ins dialog box, make sure that the **Analysis ToolPak** check box under **Add-Ins available** is selected, and then click **OK**.

For the Analysis ToolPak Add-ins to be displayed in the Add-Ins dialog box, it must be installed on your computer. If you do not see Analysis ToolPak in the Add-Ins dialog box, run Microsoft Excel Setup and add this component to the list of installed items.

To create a frequency table and corresponding histogram, follow these steps:

1. Type in the data into a column, say $A1$ to $A100$, of a new worksheet.
2. Type in the bin numbers in column B, say $B1$ to $B7$.
3. Do one of the following:
   In Excel 2007 and recent versions, on the **Data** tab, click **Data Analysis** in the **Analysis** group.
   For Excel 2003 and earlier versions, click **Data Analysis** on the **Tools** menu.



4. In the **Data Analysis** dialog box, click **Histogram**, and then click **OK**.
5. In the **Input Range** box, type $A1 : A100$.
6. In the **Bin Range** box, type $B1 : B7$.

7. Under Output Options, click New Workbook, select the Chart Output check box, and then click OK.Excel generates a new workbook with a frequency table and an embedded chart.

If Analysis ToolPak is not installed in your computer and you are unable to run Microsoft Excel Setup then also you can draw histogram having bars of equal width by following the procedure for drawing simple bar chart and making use of the chart option for setting gaps between the bars to be 0. The only disadvantage of this method is that we are unable to mark the class limits; but we can display the class intervals or the mid values of the class intervals as labels. The steps to be followed are given below:
Firstly, prepare the frequency table having class intervals of equal width. For this you can make use of **COUNT** function or Sort option of Excel. Please note that Excel cannot generate frequency table automatically without Analysis ToolPak option. Enter the class marks or class intervals in one column and corresponding frequencies in the next column of the MS Excel worksheet. Now follow the steps to draw a simple bar diagram and obtain the same. Then place the cursor on one of the bars and right click on the mouse. Then, click **Format Data Series** $\rightarrow$ **Options**$\rightarrow$ type 0 in **Gap width**. Finally click **OK** to obtain the required histogram.
The histogram graphically shows the following:

- the center of the data,
- the spread or scatter of the data,
- the skewness in the data.

We shall study the above characteristics of a data in coming chapters.

Frequency Polygon and Frequency Curve
A frequency polygon is another graphical presentation of a grouped frequency table. It is a line graph where we mark midpoints of class intervals (class marks) along the horizontal axis and the corresponding frequencies along the vertical axis. Thus we get one point each in each class interval. Then these points are joined by pieces of straight lines and the resulting graph is called a *frequency polygon*. If the points are joined by a smooth curve then it is called a *frequency curve*. Figure 3.9 show the frequency polygon corresponding to the frequency table obtained in Example 3.2.1.

To draw the frequency polygon in MS Excel, enter the class marks in one column and the corresponding frequencies in the next column. Then, make the following sequence of clicks. **Insert**$\rightarrow$ **Chart** $\rightarrow$ select chart type **XY (Scatter)** $\rightarrow$ select chart sub-type 'Scatter with data points connected by lines' $\rightarrow$ **Next** $\rightarrow$ fill **Data range** by dragging the cells containing the co-ordinate values$\rightarrow$ Series in **Columns** $\rightarrow$ **Next**$\rightarrow$ **Finish**. Now the procedure for drawing frequency curve using MS Excel is obvious.

Frequency curve and histogram of a data drawn as a single graph is shown in Figure 3.10. From the figure notice that the area enclosed by the frequency curve and the histogram are equal. One of the primary purposes of a histogram or frequency polygon is to exhibit the symmetry, or its absence, in a data representation. Note that frequency polygon can be obtained from the histogram by linking the midpoints of the tops of the rectangles of the histogram.

Figure 3.9: Frequency polygon

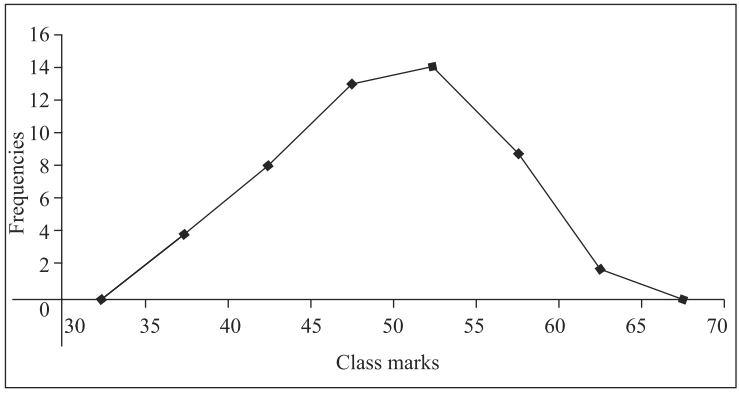
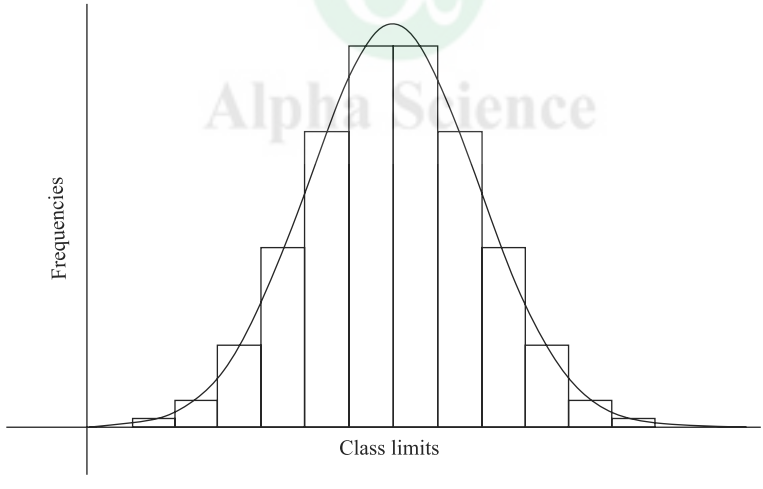


Figure 3.10: Frequency curve superimposed on a histogram



If the class intervals are of unequal width it is recommended to divide the class frequencies by the corresponding class widths. This gives the proportional frequencies of the respective classes. When we draw a histogram of with unequal class widths it is better to consider proportional frequencies instead of frequencies.

If we divide the relative frequencies by the corresponding class width we obtain the *relative frequency densities*. When we draw a histogram with the heights of bars erected at class intervals correspond to the relative frequency densities then the histogram so obtained is a relative frequency density histogram. Thus, in geometric terms, the area under a relative frequency histogram is always equal to 1.

In some situations the maximum frequency is at one end of the range of observations. It may also happen that the frequencies may increase on either side of the point of minimum frequency and the corresponding frequency curve is U-shaped.

Ogives

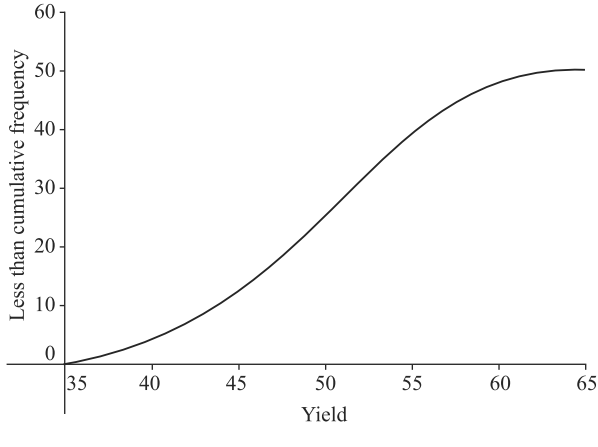
To draw an ogive one has to obtain the cumulative frequencies first. The *cumulative frequencies* are obtained by adding actual frequency of the particular class to the previous cumulative frequency and it is explained in the following example. Since the underlying variable on which observations are made is continuous and assuming continuity in the data points, a smooth curve can be drawn by representing the upper class boundaries along the horizontal axis and the corresponding cumulative frequencies along the vertical axis. The resulting graph is called an *ogive*.

Example 3.3.8. The cumulative frequencies of the frequency table in Example 3.2.1 is shown in Table 3.6 and the corresponding ogive is displayed in Figure 3.11.

Table 3.6: Less than cumulative frequency

Yield	Cumulative frequency	Found by
less than 35	0	
less than 40	4	
less than 45	12	$\leftarrow$ 4+8
less than 50	25	$\leftarrow$ 12+13
less than 55	39	$\leftarrow$ 25+14
less than 60	48	$\leftarrow$ 39+9
less than 65	50	$\leftarrow$ 48+2

Figure 3.11: Less than ogive

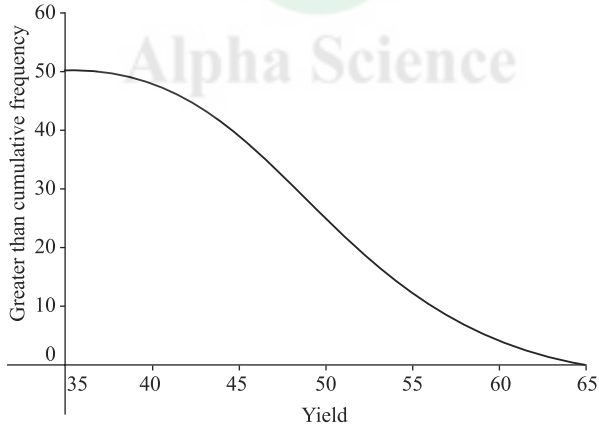


Remark 3.3.1. In the above we have considered '*less than*' cumulative frequencies and ogives. It is also possible to form '*greater than*' or '*more than*' cumulative frequencies and ogives, which show the number of measurements that exceed particular values. In this case the lower limits of class intervals are used as abcissae and the greater than cumulative frequencies as ordinates. The greater than cumulative frequencies based on the frequency table in Example 3.2.1 is given in Table 3.7 and the corresponding ogive is shown in Figure 3.12.

Table 3.7: Greater than cumulative frequency

Yield	Cumulative frequency
greater than 35	50
greater than 40	46
greater than 45	38
greater than 50	25
greater than 55	11
greater than 60	2
greater than 65	0

Figure 3.12: Greater than ogive



When we draw the less than ogive and the greater than ogive as a single graph the two ogives will intersect at a point whose ordinate corresponds to half of the total frequency.

If the frequency table formed is based on a discrete variable, the graph drawn corresponding to cumulative frequencies will look like the graph of a step-function. This is illustrated in the following example.

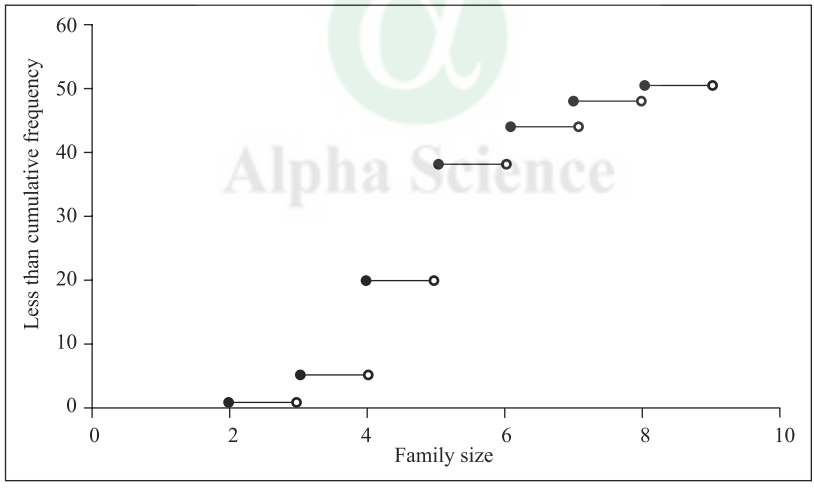
Example 3.3.9. The total number of members in the family (family size) of 50 students in a class is noted and the data is summarized in the first two columns of Table 3.8.

Table 3.8: Cumulative frequencies

Family size	Number of students	Family size	Cumulative frequency
2	1	less than 2	0
3	4	less than 3	1
4	15	less than 4	5
5	18	less than 5	20
6	6	less than 6	38
7	4	less than 7	44
8	2	less than 8	48
Total	50	less than 9	50

The less than cumulative frequencies corresponding to family sizes is shown in the last column of Table 3.8 and the corresponding graph is shown in Figure 3.13.

Figure 3.13: Graph of a step function



The main factors to be taken into account in the choice of a chart or graph are:

(i) Aim of the display of data.

(ii) Type of the variable or data.

(iii) Audience for whom a chart or graph is prepared.

(iv) Level of accuracy of facts to be presented.

Good data presentation skills are to data-based analysis what good writing is to literature. The development of spreadsheet and other statistical packages have greatly simplified the tasks of tabular and graphical data presentation. The Web site *http://phoenix.phys.clemson.edu/tutorials/excel/* has a discussion of many of the Excel functions for mathematics and statistics, in particular graphing data. The following sites

http://lilt.ilstu.edu/gmklass/pos138/datadisplay/sections/ goodcharts.htm and *http://lilt.ilstu.edu/gmklass/pos138/ datadisplay/badchart.htm* show how to create meaningful and readable graphs. Both sites also give several examples of poorly constructed graphs.

3.4 EXERCISES

1. Classification is applicable in case of:
   (a) quantitative variables; (b) ordinal variables;
   (c) nominal variables; (d) all of these.

2. A series showing the set of all distinct values individually with their frequencies is known as:
   (a) grouped frequency data; (b) simple frequency data;
   (c) cumulative frequency data; (d) none of these.

3. Choice of a particular chart depends on :
   (a) the purpose of the study; (b) the nature of the data;
   (c) the type of audience; (d) all of these.

4. Graphs and charts facilitate:
   (a) comparison of values; (b) to know the trend;
   (c) to know relationship; (d) all of these.

5. Which is best for categorical variables?
   (a) bar chart; (b) ogive; (c) stem and leaf plot; (d) histogram.

6. Which is best for numerical variables?
   (a) bar chart; (b) pie chart; (c) stem and leaf plot; (d) pictograph.

7. Pie chart represents the components of a factor by:
   (a) angles; (b) percentages; (c) sectors; (d) circles.

8. In an ogive the points are plotted for:
   (a) class limits and frequencies;
   (b)class limits and cumulative frequencies;
   (c) frequencies and cumulative frequencies;
   (d) class marks and cumulative frequencies.

9. Collect the data on voter turnout rates by State during the last parliament election and prepare a stem and leaf plots of the data.

10. Decide which type of table and graphical method you would use on the following univariate(one variable case) data sets:

    (a) The number of seats won by major political parties during the last parliament election.

    (b) The number of deaths due to lightning on each type of land. The land types are flat, hilly and mountainous.

    (c) The fluoride content of the public water supply for 5 cities.

(d) There are 965 male students and 524 female students in a certain college. Among the males, 785 come from the rural areas, the rest from urban areas. Among the females, 302 come from rural areas, the rest from urban areas. Present this data in a suitable table and also draw a suitable chart to represent it.

11. A recent study reported that the time spent in hours of personal computer usage per week for a sample of 60 students is given below.

9.3	5.3	6.3	8.8	6.5	0.6	5.2	6.6	4.3	6.3	2.1	2.7
0.4	3.7	3.3	1.1	6.7	6.5	4.3	9.7	7.7	5.2	1.7	8.5
4.2	5.1	5.6	5.4	4.8	2.1	10.1	1.3	5.6	2.4	4.7	1.7
2.0	6.7	1.1	6.7	2.2	2.6	9.8	4.9	5.2	4.5	9.3	7.9
4.6	4.3	4.5	9.2	6.0	8.1	6.2	9.0	3.4	5.5	5.6	7.5

(a) Make a stem and leaf plot of the data.
(b) What type of graph does a stem and leaf plot represent when turned vertically?
(c) For what time interval did most students use computer?

12. Taking the poverty line in Kerala as an annual income of ₹ 24000, below poverty line (BPL) as negative, and above poverty line as positive, for example, an annual income of ₹35000 is 11000 above BPL which means + 11000 and an annual income of ₹14000 will be 10,000 below poverty line and will be recorded as -10,000. If a person is in debt by 5000 then the reading will be -24000 - 5000 = -29000. The following are the income in thousands in a random sample of 99 households.

35	67	375	-10	7	23	850	450	0	94	-3
255	170	3	73	160	745	-2	365	180	100	115
-22	135	20	15	556	230	45	1	6	33	-1
1520	930	75	10	-45	23	80	123	135	250	740
925	330	8	-7	-17	0	18	17	37	246	650
12	47	64	4	523	1	34	169	295	0	4
250	-1	0	65	186	284	-21	26	41	46	91
125	476	276	436	26	1	7	51	101	6	176
-16	0	376	11	158	260	87	5	10	70	-11.

Form a frequency table and draw the histogram.

13. In Simla, the temperature can vary from -20^0C to 40^0C. Noon temperature reading is done for 100 days. The following is the data:
-16, -18, -19, -20, -16, -15, -14, 0, -5, 5, -1, -3, -8, 0, 12, 14, 16, 18, 21, 15, 14, 10, 11, 12, 10, 8, 5, 0, -1, -3, -7, -10, -5, 0, 5, -6, -8, -12, 0, 1, -2, -8, -3, 0, -2, 1, 2, 6, 8, 12, 1, 14, 15, 18, 19, 20, 21, 23, 24, 20, 22, 21, 25, 30, 28, 29, 25, 24, 20, 25, 28, 30, 24, 26, 31, 32, 35, 38, 40, 38, 39, 40, 35, 34, 36, 31, 30, 32, 36, 38, 35, 35, 30, 30, 38, 38, 32.
Taking the classes as [-20, -10), [-10, 0), [0, 10), [10, 20), [20, 30), [30, 40] (a) form the frequency table, (b) draw the histogram, (c) draw the frequency polygon.

14. Draw an ogive based on the frequency table constructed in Exercise 12. Estimate the percentage of people who uses computer for (a) more than 5 hours; (b) less than 2 hours.

15. The following data gives the area of land in a certain country. Describe the data with a pie chart.

Forest land	Farm land	Urban	Other land
30%	40%	10%	20%

16. Of the 500 students in a college, 160 were of blood type O, 180 of type A, 135 of type B and 25 of type AB. Construct a pie chart to depict the data.

17. The following table shows the revenue collection of India in crores during the current year. Represent the data using suitable diagrams.

Direct taxes during 2008-2009 (Rupees in crores)			
Corporate tax	Income tax	Other direct taxes	Total
2,08,991	1,17,740	416	3,27,147
Indirect taxes during 2008-2009 (Rupees in crores)			
Customs	Central Excise	Service tax	Total
93,893	95,258	49,479	2,38,630

18. A sample data regarding automobiles were collected from 45 persons in a city. The data is summarized with the following letter codes:

C: Chevrolet F: Ford H: Hyundai M: Mahindra S: Suzuki T: Toyota.

The data set is:

S	M	S	H	C	M	S	T	H	T	F	S	S	T	H
T	M	S	S	F	H	H	T	S	M	S	F	S	C	S
C	S	H	S	M	T	F	H	H	F	S	C	S	M	T

(a) Prepare a frequency table for the above data. Also compute the relative frequencies.

(b) Represent the data using a suitable chart.

19. What is the difference between a histogram and a bar diagram?

20. Prepare a blank table for presenting the data of students according to religion, year of study (First/ Second/ Third) and place of residence (home/ hostel/ others).

21. Given the following cumulative frequency table, fill in the missing frequencies:

Class interval	Frequency	Cumulative frequency
0-10	7	7
10–20	12	19
20-30	24	43
30-40	-	-
40-50	14	-
50-60	5	80

22. A partial relative frequency table is given below.

Class	A	B	C	D
Relative frequency	0.26	0.40	0.24	-

(a) What is the relative frequency of class D?

(b) The total sample size was 150. Write down the frequencies of each class.

23. There are 1312 male students and 1024 female students in a college in Kerala. Among the males, 685 come from the southern Kerala, the rest from northern Kerala. Among the females, 407 come from the southern Kerala, the rest from the northern Kerala.

    (a) Draw a bar chart showing the number of students by sex.

    (b) Draw a bar chart showing the number of students by region.

24. Table 3.9 gives the data of annual income of households in a certain municipality of Kerala. Prepare a frequency table from this data. Also draw an appropriate graph to represent this data.

Table 3.9: Annual income of households

Income (in Rupees)	Cumulative frequency
20,000 or more	35,251
50,000 or more	22,075
1,00,000 or more	13,766
1,50,000 or more	7,341
2,00,000 or more	3,851
3,00,000 or more	1,885
5,00,000 or more	967

25. A fresh fish sales lady who carries fish in a basket and walk to the interior parts of Kerala has the following situations. On certain days she has profits (total income from sales - total expenses), on certain days there are loses due to non-sale of fish and on certain days she breaks even. The following data indicates her 'profit or loss' (loss taken as negative profit) she incurred in two months period. Taking the class intervals as [-200, -100), [-100, 0), [0, 100), [100, 200), [200, 300), [300, 400] (a) form the frequency table, (b) draw histogram and (c) draw frequency polygon for the data.
    390, -50, -150, 45, 86, 269, 340, -160, -180, 270, -20, 360, 290, -70, 350, 265, -50, -90, 275, -25, 320, 340, 180, -120, 310, -200, 100, -90, 45, 300, 220, 135, 180, 250, -20, 125, -110, 75, 60, 275, 230, 175, 35, 145, 190, 0, 125, 130, 170, 0, 10, -5, 50, 245, 350, 275, 310, 240, 320, 140.

26. For the data given in the above exercise and the same class intervals, (a) form the table of less than cumulative frequencies, (b) form the table of greater than cumulative frequencies, (c) draw the graph in (a) as a step function, (d) draw the graph in (b) as a step function, (e) assuming continuity in the points draw a smooth curve to obtain less than cumulative curve, (f) assuming continuity in the points draw a greater than cumulative curve, (g) draw the curve in (e) and (f) in a single figure and make comments on the findings.

Chapter 4

MEASURES OF CENTRAL TENDENCY

"The inherent inability of human to grasp in its entirely a large body of numerical data compels us to seek relatively few constants that will adequately describe the data"- Sir R.A.Fisher.

4.1 INTRODUCTION

In the last chapter we have studied how to summarize a mass of raw data into a meaningful form and present it through tables and graphs. This chapter is concerned with how to represent a given set of data in terms of a single value. A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. Measures of central tendency are also known as *averages*.

Average is a word used frequently in everyday conversation. One talks about the batting average of a cricketer, average family, average annual rainfall, average student, the grade point average and so on. There are different measures that are used for this purpose. It seems reasonable that the representative value should be a number which on a numerical scale is somewhere at the center of the spread of the data. That is why an average is referred to as a measure of central tendency. The commonly used averages are the arithmetic mean, the median, the mode, the geometric mean and the harmonic mean. An average summarizes the data and hence it is sometimes termed as summary measure.

4.2 THE ARITHMETIC MEAN

The arithmetic mean or simply the mean, is the most commonly used measure of central tendency of a given data. In everyday usage this term is erroneously regarded as a synonym for average. The mean is only one of the measures of average or central tendency.

Before we define arithmetic mean we must introduce a notation called sigma notation or summation notation. Suppose we record values of a quantitative variable, say the heights of a sample of n students. Let us denote the variable height by x. Let x_1 be the height of the first student chosen, x_2 be the height of the second student, and so on. Then $x_1, \ldots, x_n$ constitute n observations on x. The symbol Σ, upper-case Greek letter sigma, is used as a summation notation. Thus

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Some properties of Σ notation

(i) When a is a constant and if it is summed up n times then

$$\begin{aligned} \sum_{i=1}^{n} a &= a + a + \cdots + a = na; \quad \sum_{i=1}^{5} 7 = 7+7+7+7+7 = 35 = 5 \times 7; \\ \sum_{i=1}^{3} (-1) &= (-1) + (-1) + (-1) = -3 = 3(-1). \end{aligned}$$

(ii) When a is a constant then

$$\begin{aligned} \sum_{i=1}^{n} ax_i &= ax_1 + ax_2 + \cdots + ax_n = a(x_1 + \cdots + x_n) = a\sum_{i=1}^{n} x_i; \\ \sum_{i=1}^{3} 4x_i &= 4\sum_{i=1}^{3} x_i = 4(x_1 + x_2 + x_3). \end{aligned}$$

Combining (i) and (ii) we have

$$\sum_{i=1}^{n} (ax_i + b) = a\sum_{i=1}^{n} x_i + nb.$$

(iii)

$$\begin{aligned} \left(\sum_{i=1}^{n} a_i\right)^2 = (a_1 + \cdots + a_n)^2 &= \sum_{i=1}^{n} a_i^2 + 2\sum_{i<j} a_i a_j = \sum_{i=1}^{n} a_i^2 + 2\sum_{i>j} a_i a_j \\ &= \sum_{i=1}^{n} a_i^2 + \sum_{i \neq j} a_i a_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j. \end{aligned}$$

$$(a_1 + a_2 + a_3)^2 = (a_1^2 + a_2^2 + a_3^2) + 2(a_1a_2 + a_1a_3 + a_2a_3) = \sum_{i=1}^{3} a_i^2 + 2\sum_{i<j} a_i a_j.$$

$$(a_1 + a_2 + a_3)^2 = (a_1^2 + a_2^2 + a_3^2) + 2(a_2a_1 + a_3a_1 + a_3a_2) = \sum_{i=1}^{3} a_i^2 + 2\sum_{i>j} a_i a_j.$$

$$\begin{aligned} (a_1 + a_2 + a_3)^2 &= (a_1^2 + a_2^2 + a_3^2) + (a_1a_2 + a_2a_1 + a_1a_3 + a_3a_1 + a_2a_3 + a_3a_2) \\ &= \sum_{i=1}^{3} a_i^2 + \sum_{i \neq j} a_i a_j. \end{aligned}$$

$$\begin{aligned} (a_1 + a_2 + a_3)^2 &= (a_1a_1 + a_2a_2 + a_3a_3 + a_1a_2 + a_2a_1 + a_1a_3 + a_3a_1 + a_2a_3 + a_3a_2) \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} a_i a_j. \end{aligned}$$

(iv) $\left(\sum_{i=1}^{m} a_i\right)\left(\sum_{j=1}^{n} b_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j.$

(v) $\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}.$

If a variable is denoted by x, where x takes the values 0, 5 and 8 then $\Sigma x = 0+5+8 = 13$.

The arithmetic mean, denoted by $\bar{x}$ and read as x bar, of a set of n observations $x_1,\ldots,x_n$ is given by

$$\bar{x} = \frac{x_1+\cdots+x_n}{n} = \frac{\text{Sum of all the values}}{\text{Number of values}}.$$

If the data consists of only k distinct values $x_1,\ldots,x_k$ and x_i is repeated f_i times $i = 1,\ldots,k$ then the arithmetic mean will become the following, or is obtained by the formula

$$\bar{x} = \frac{f_1x_1+\cdots+f_kx_k}{f_1+\cdots+f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}. \tag{4.2.1}$$

A data is said to be grouped when it is grouped in class intervals and presented in the form of a frequency table. When we form a grouped frequency table the identity of the individual observations is lost. The formula (4.2.1) can be used for computing arithmetic mean of a grouped data by replacing x_i's with the corresponding class marks. The idea is to take the class marks as representatives of the respective classes and to treat them as though they were the actual observed values.

Example 4.2.1. A hospital wants to determine the mean age(average age) of its doctors where the ages of doctors are:

36 59 42 50 57 29 32 32 36 41 45.

Find the mean age.

$$\text{Mean age } = \frac{36+59+\cdots+45}{10} = 42.7.$$

Example 4.2.2. The numbers of hours taught by 20 professors during a semester are given in the first two columns of Table 4.1. Find the mean number of hours taught by the professors in the semester.

Let x denotes the number of hours taught by a professor. Mean number of hours taught by the professors is given by

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{1588}{20} = 79.4 \text{ hours}.$$

Table 4.1: Number of hours taught by 20 professors

Hours (x_i)	Frequency (f_i)	$f_i x_i$
74	3	222
78	6	468
80	8	640
85	2	170
88	1	88
Total	20	1588

Example 4.2.3. The lives (in hours) of 400 electric bulbs are given in the first two columns of Table 4.2. Find the mean life of bulbs.

Let x denote the life length of an electric bulb. To calculate the arithmetic mean of the above data we form the last two columns of Table 4.2.

Mean life length of bulbs is given by

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{330300}{400} = 825.75 \text{ hours} .$$

Table 4.2: Life lengths of electric bulbs

Life of bulb (in hours)	Number of bulbs (f_i)	Class mark (x_i)	$f_i x_i$
$[600, 700)$	85	650	55250
$[700, 800)$	77	750	57750
$[800, 900)$	124	850	105400
$[900, 1000)$	78	950	74100
$[1000, 1100)$	36	1050	37800
Total	400		330300

Computation Using MS Excel

Once the values are entered in the worksheet of MS Excel the arithmetic mean can be computed as follows:

Select a blank cell where you wish to display the answer→ **Insert** → **Function**→ select Function category 'Statistical'→ click on Function name **AVERAGE**→ **OK**→ drag to the cells in which the values are entered→ **OK**. In an alternative way one can obtain the answer very quickly as follows:

Suppose the values are entered in the cells from A1 to A20 and we want to get the answer in cell C2. Type an equal sign $(=)$ in cell C2. An equal sign tells Excel you

are going to enter a formula. The equal sign will appear on the formula bar. Then type **AVERAGE(A1:A20)** and press Enter key to obtain the answer.

If we make any change in the data values then the answer also will change accordingly. If you wish to make a change in the formula after entering it in a cell then you can do it only in the formula bar and not in the cell. The steps for computing the arithmetic mean of the data in Table 4.2 are the following:

1. Enter the class marks in the cells from A1 to A5 and the corresponding frequencies in the cells from B1 to B5.

2. Now select cell B6, click Σ in the menu bar (or type in **=SUM(B1:B5)**) and press Enter to obtain the sum of the values in column B.

3. Then, select cell C1 and type the formula **=A1*B1** and press enter to obtain the product of values in the cells A1 and B1. Note that asterik (*) is the symbol for multiplication in computer formulae.

4. You need not type again the formula to obtain the product of values in A2 and B2, A3 and B3, A4 and B4 and A5 and B5. Instead, you select cell C1 and place the cursor at the right bottom corner of the selection box to show a thin '+' symbol. Now drag to cells B2, B3, B4 and B5. Excel will execute the copied formulae instantly.

5. Next, obtain the sum of the values in column C in cell C6.

6. Finally, select a blank cell and type the formula **=C6/B6** to obtain the arithmetic mean. Note that slash (/) is the symbol for division in computer formulae.

Weighted Mean

In some cases the values should not be weighted equally. For example, suppose that we have voting percentages of three States during the last Parliament election as 50%, 55% and 60%. The number of voters in these States were respectively 10 millions, 6 millions and 4 millions respectively. Therefore, in determining the mean voting percentage when the three states taken together should be weighted according to its total number of voters.

The weighted mean of a set of values $x_1, \ldots, x_k$ with the corresponding weights being $w_1, \ldots, w_k$ is computed by:

$$\bar{x}_w = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}.$$

Note that any measure of importance could be used as a weight. In computing arithmetic mean of a grouped data we used the frequencies as weights.

Example 4.2.4. A certain company pays its hourly employees ₹50, ₹60 and ₹ 75 per hour. There are 36 hourly employees, 20 are paid at ₹50 per hour, 12 at ₹ 60 per hour,

and 4 at ₹75 per hour. What is the mean hourly rate paid to the 36 employees?

From the above formula

$$\bar{x}_w = \frac{(20\times 50)+(12\times 60)+(4\times 75)}{36} = ₹56.11.$$

The weighted mean hourly wage is ₹56.11.

Combined Mean

Let there be k groups of observations of sizes $n_1,\ldots,n_k$ such that $n_1+\cdots+n_k=n$ and $\bar{x}_1,\ldots,\bar{x}_k$ be the respective arithmetic means of the k groups. Then $n_i\bar{x}_i$ is the sum of the values in group i and $\sum_{i=1}^{k} n_i\bar{x}_i$ is the sum of the values in all the k groups. Hence the arithmetic mean of $n_1+\cdots+n_k=n$ values is given by:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}.$$

Example 4.2.5. The mean weight of 50 students in a certain class is 60.25 kilograms. The mean weight of boys in the class is 64.3 kilograms and that of girls is 55.1 kilograms. Find the number of boys and girls in the class.

Let n_1 and n_2 be the number of boys and girls respectively. Then

$$\begin{aligned} 60.25 &= \frac{(n_1\times 64.3)+(n_2\times 55.1)}{50}. \\ 64.3n_1+55.1(50-n_1) &= 60.25\times 50. \end{aligned}$$

Therefore,

$$\begin{aligned} 9.2n_1 &= 3012.5-2755=257.5 \\ n_1 &= \frac{257.5}{9.2}=28. \end{aligned}$$

That is,

$$n_2 = 50-28=22.$$

Properties of the Arithmetic Mean

1. The sum of the deviations of the observations from their arithmetic mean is zero.
   Proof: Let

$$d_i = x_i-\bar{x}, i=1,\ldots,n.$$

Then

$$\begin{aligned} \sum_{i=1}^{n} d_i &= \sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x} \\ &= \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x}-n\bar{x}=0. \end{aligned}$$

The arithmetic mean can be regarded as the *center of gravity* of the data.

2. Effect of change of origin and scale on arithmetic mean.

   When we add or subtract a constant value from all the observations, we say that origin of measurements has been shifted or the data have been relocated. If we multiply or divide all the observations by a given constant then we say that the scale of measurements is changed even if the constant is a negative number.

   Let $x_1,\ldots,x_n$ be a set of n observations and $y_1,\ldots,y_n$ be the new set of observations after changing the origin and scale. Define

   $$y_i = a + bx_i, \quad i = 1,\ldots n \tag{4.2.2}$$

   where 'a' and 'b' are given constants. Hence

   $$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + b\sum_{i=1}^{n} x_i = na + b\sum_{i=1}^{n} x_i.$$

   Now dividing both sides by n, we obtain

   $$\bar{y} = a + b\bar{x}.$$

   Sometimes this property is used in the manual computation of arithmetic mean. The mean $\bar{y}$ of the transformed values may be found out very easily and hence we can determine the mean of the original set of observations by using the equation:

   $$\bar{x} = \frac{\bar{y} - a}{b}.$$

3. The sum of squares of deviations of a set of observations is minimum when it is taken about the arithmetic mean.

Proof: Define

$$\begin{aligned} S &= \sum_{i=1}^{n} (x_i - A)^2 \\ &= \sum_{i=1}^{n} [(x_i - \bar{x}) + (\bar{x} - A)]^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\bar{x} - A)^2 + 2\sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - A). \end{aligned}$$

Since $(\bar{x} - A)$ is free of i, it acts as a constant when summing up. Therefore,

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - A) &= (\bar{x} - A)\sum_{i=1}^{n} (x_i - \bar{x}) \\ &= (\bar{x} - A)\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}\right) \\ &= (\bar{x} - A)(n\bar{x} - n\bar{x}) \\ &= (\bar{x} - A) \times 0 \\ &= 0. \\ \sum_{i=1}^{n} (\bar{x} - A)^2 &= n(\bar{x} - A)^2. \end{aligned}$$

Hence,

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

where $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is free of A. But $(\bar{x} - A)^2 \geq 0$ being a square of real numbers. Hence the minimum for S is attained when $n(\bar{x} - A)^2 = 0$ which means $\bar{x} - A = 0$ since the square is non-negative. That is, $\bar{x} = A$.

Alternative proof using calculus method

From calculus, the conditions for minimum are:

$$\begin{aligned} \frac{\partial S}{\partial A} &= 0 \quad \text{and} \quad \frac{\partial^2 S}{\partial A^2} > 0. \\ \frac{\partial S}{\partial A} &= -2\sum_{i=1}^{n}(x_i - A) = 0 \Rightarrow \sum_{i=1}^{n}(x_i - A) = 0. \end{aligned}$$

That is

$$n\bar{x} - nA = 0 \text{ or } A = \bar{x}.$$

Again

$$\frac{\partial^2 S}{\partial A^2} = 2 > 0.$$

Hence S is minimum when $A = \bar{x}$.

Example 4.2.6. Let us consider the set of values 0.02, -0.05, 0.06, 0.05, -0.03. Note that the mean ($\bar{x}$) of this set of values is 0.01.
On multiplying each observation by 100 and adding 5 to each value we get the set of transformed values as 0,7,11, 2, 10. On comparing with (4.2.2) we have taken $a = 5$ and $b = 100$. The mean ($\bar{y}$) of the transformed set of values is 6. Note that

$$6 = \bar{y} = b\bar{x} + a = (100 \times 0.01) + 5.$$

Example 4.2.7. Suppose we want to find the mean of the values 10006, 10003, 10009, 10005 and 10002. On subtraction 10,000 from each of these values we obtain the new set of values as 6, 3, 9, 5 and 2. Here we have taken $a = -10,000$ and $b = 1$ in (4.2.2). Now compare the convenience in the manual computation of the mean of this new set of values with that of the original set of values. The mean of the transformed values is 5 and hence the mean of the original set is 10,000+5=10005.

Remark 4.2.1. In the case of a grouped frequency table having class intervals of equal width c the calculation of arithmetic mean by changing origin and scale is more convenient. Let

$$y_i = \frac{x_i - a}{c}, \quad i = 1, \ldots n. \tag{4.2.3}$$

That is,

$$x_i = cy_i + a$$

and hence

$$\bar{x} = a + c\bar{y}$$

where

$$\bar{y} = \frac{\Sigma f_i y_i}{\Sigma f_i}.$$

We shall illustrate it with the help of the following example.

Example 4.2.8. The amounts donated by 200 families towards social action fund is shown in the first two columns in Table 4.3. Calculate the mean amount of donation of a family.

Now considering the transformation in (4.2.3), we take $a = 1750$ and $c = 500$.

$$\sum f_i = 200, \quad \sum f_i y_i = -15 \text{ and } \bar{y} = \frac{\Sigma f_i y_i}{\Sigma f_i} = -\frac{15}{200}.$$

Thus

$$\bar{x} = a + c\bar{y} = 1750 + 500 \times \frac{(-15)}{200} = ₹1712.50.$$

Table 4.3: Computation of mean using transformation

Amount contributed	Number of families (f_i)	Class mark (x_i)	$y_i = \frac{x_i - 1750}{500}$	$f_i y_i$
0-500	10	250	-3	-30
500-1000	15	750	-2	-30
1000-1500	45	1250	-1	-45
1500-2000	70	1750	0	0
2000-2500	35	2250	1	35
2500-3000	20	2750	2	40
3000-3500	5	3250	3	15
Total	200			-15

Note that the value a is taken as the class mark of a suitably chosen class interval. It is convenient to take a as that class mark having largest frequency.

4.3 THE MEDIAN

The median of a set of numbers is the middlemost value when the numbers are arranged in the ascending or descending order of magnitude. If there is an odd number of numbers, the median always turns out to be one of the numbers in the data.

Example 4.3.1. The amount of pocket money in rupees with 7 students in a class are 15, 10, 18, 27, 5, 32, and 4. Compute the median of the data.

Arranging the data in a ascending order of magnitude, we get

$$4 \quad 5 \quad 10 \quad \boxed{15} \quad 18 \quad 27 \quad 32.$$

Since 15 is the middle number, the median is 15.

We have seen that if there are 7 observations, the median is the value of the $\left(\frac{7+1}{2}\right)^{\text{th}}$ largest observation. In general, if there are n observations and if n is odd, then the median is the value of the $\left(\frac{n+1}{2}\right)^{\text{th}}$ largest observation.

How is the median or middle value determined for an even number of observations? As before the observations are ordered. There will be two observations at the middle. Then by convention to obtain a unique value we calculate the arithmetic mean of the two middle observations. In general, if there are n observations and if n is even, then the median is taken as the mean of the $\frac{n}{2}^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ largest observations.

Example 4.3.2. The family size of 50 students in a class is given below.

Family size	:	3	4	5	6	7
Number of students	:	3	23	17	5	2

Find the median family size.

Since the number of observations is 50, an even number, the median is the mean of 25^{th} and 26^{th} largest observations. On writing the cumulative frequencies we can see that 25^{th} and 26^{th} largest observations are equal to 4. Hence the median family size is 4.

To obtain median using Excel, use the **MEDIAN** function just similar to the way in which we use **AVERAGE** function.

Median from Grouped Data

When the data is grouped into classes and arranged in a frequency table, the value obtained for median is an approximate value. We shall illustrate the way of finding the median of a grouped data with the help of the following example.

Example 4.3.3. The frequency table of the number of years of schooling of 100 workers in a factory is given below:

Number of years	:	0-8	8-12	12-16	16 and over
Number of workers	:	1	12	49	38

To compute median we have to form the cumulative frequency table and it is given below:

	Class	Frequency	Cumulative frequency
	0-8	12	12
	8-12	22	34
Median class	12-16	28	62
	16 and over	38	100

If the actual data were available we would have taken the median as the mean of 50^{th} or 51^{th} largest observations. From the cumulative frequency column, we see that median lies in 12-16 class and it is called the *median class*. Now, the total frequency upto the median class is 34, so that to get to the 50^{th} observation we need 16 more observations. There are 28 observations in the class 12-16 spread over a class width of 4 units. Therefore assuming that the frequencies are evenly spread, the 16^{th} largest observation

in this interval will be $\frac{4}{28}(50-34) = \frac{4}{28}(16)$ units above the lower boundary of the median class. Thus the median is taken as

$$12 + \frac{4}{28}(50-34) = 14.28 \text{ years.}$$

Based on the above illustration the formula for finding median of a grouped data can be given as follows:

$$\begin{aligned} \text{median} &= L_m + \frac{c}{f_m}\left(\frac{N}{2} - F\right) \qquad (4.3.1) \\ \text{where } L_m &= \text{the lower boundary of median class} \\ N &= \text{the total frequency} \\ F &= \text{the total frequency upto the lower limit of median class} \\ f_m &= \text{the frequency of the median class} \\ c &= \text{the length of the class interval.} \end{aligned}$$

The median can also be determined graphically by using ogive. The median is the abcissa of the point on the ogive whose ordinate corresponds to 50% of total frequency. When we draw the less than ogive and the greater than ogive as a single graph the two ogives will intersect at a point whose ordinate corresponds to half of the total frequency and abcissa corresponds to the median.

4.4 THE MODE

The mode is defined as the value that occurs most frequently in the data.

It may happen that there are some values occurring equally often, but more frequently than the rest of the values in the data. For example, if the data consists of the observations 5, 3, 3, 4, 2, 4, 4, 6, 3 and 7, then the values 3 and 4 occur equally often, namely, three times. Hence there are two modes, 3 and 4. It is therefore called bimodal. In general, if the data has more than one mode, it is said to be multimodal.

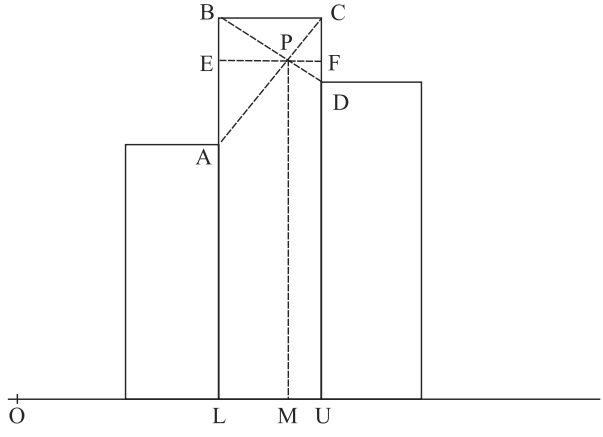
For many sets of data, there is no mode because no value appears more than once. For instance, the set of values 19, 21, 23, 20 and 18 has no mode.

The **MODE** function of MS Excel will display the value only if mode exists. If the data has more than one mode then also it will display only one value.

For grouped data, mode may be determined geometrically as follows: Draw the histogram of the frequency table and consider the three rectangles - the rectangle corresponding to the largest frequency and the two adjacent rectangles as shown in Figure 4.1. Then the mode is given by the abcissa of the point M in Figure 4.1.

Based on a frequency polygon or curve, it is easy to find the mode. Mode is the value along the horizontal axis where the frequency polygon or curve achieves its maximum vertical height.

Figure 4.1: Graphical determination of mode



Based on Figure 4.1 we can obtain the formula for finding mode in the case of a grouped data. Let L denote the lower limit of the modal class, c be the length of the modal class, f_0 be the frequency of the class preceding the modal class, f_1 be the frequency of the modal class and f_2 be the frequency of the class succeeding the modal class. Then from the Figure 4.1 we have

$$\begin{aligned} \frac{EP}{AB} &= \frac{PF}{CD} \\ \frac{M-L}{f_1-f_0} &= \frac{U-M}{f_0-f_2} = \frac{L+c-M}{f_0-f_2} \end{aligned}$$

On simplification we have

$$M = L + \left(\frac{f_1-f_0}{2f_1-f_0-f_2}\right)c.$$

For categorial data, we cannot calculate either the mean or median. On the other hand the mode is useful in describing nominal and ordinal levels of measurements also. For example, if an automobile company people were to ask the customers to indicate which of several colours they preferred, and if 100 customers preferred red, 80 preferred green and 20 preferred yellow, then the mode would be at the colour red.

Comparison of mean, median and mode

1. The arithmetic mean has some convenient mathematical properties.

2. Every observation is involved in the computation of the arithmetic mean whereas the median or the mode does not take into account the individual values in the data.

3. The mean is sensitive to the extreme values; that is, it is strongly affected by an extremely large value or an extremely small value. The median is influenced by their position only and not by the size of extreme values. But extreme values do

not affect the mode. For example, the median and the mode of 10, 12, 12, 13, 14 are equal to 12 and the median and the mode of -100, 10, 12, 12, 15 are also equal to 12. The data contain just a few observations and extreme values (outliers) are present then mean gives a distorted picture of the data. The median or mode are to be preferred in such situations.

4. Mean is least affected by sampling fluctuations.

5. Mean can be computed only for quantitative data. In addition to quantitative data, median can be determined for ordinal data which can be ranked from low to high. Mode can be obtained for both quantitative and qualitative data. The mode is generally used for nominal data, median for ordinal and mean for continuous data. Strictly speaking, mode cannot be called a measure of central tendency as it does not measure central value.

6. All three measures have the following property: If each value x_i is subjected to the linear or affine transformation which replaces x_i by $ax_i + b$, so are the mean, the median and the mode.

7. However, if there is an arbitrary monotonic transformation, only the median follows; for example, if x_i is replaced by $exp(x_i)$, the median, say m, changes from m to $exp(m)$ but the mean and mode will not.

8. In continuous unimodal data the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, mean-mode $\approx$ 3(mean - median). This rule, due to Karl Pearson, is however not always true and the three measures can appear in any order.

   More properties of mean, median and mode, as well as more rigorous definitions of these measures can be seen after studying the concepts of random variables, distribution, density, expected value etc.

4.5 THE GEOMETRIC MEAN

The geometric mean is useful in finding the average of percentages, ratios, indices or growth rates. The geometric mean of a set of n positive numbers is defined as the n^{th} root of the product of n values.

$$\text{Geometric mean } = \sqrt[n]{x_1 \cdots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

where the symbol $\prod$, upper-case Greek letter pi, is used as a product notation.

Example 4.5.1. Suppose that the employees in a company receive a 5% increase in salary this year and a 15% increase next year. The average annual percentage increase is 9.886 and not 10%. Why is this so?

Let the salary of an employee be ₹7000 which is 100% and he received two increases of 5% and 15%. Hence average increase = $\sqrt{1.05 \times 1.15} = 1.09886$.

$$\begin{aligned} \text{Raise } 1 = 7000 \times 0.05 &= ₹350 \\ \text{Raise } 2 = 7350 \times 0.15 &= \underline{₹1102.5} \\ \text{Total increase in salary} &= ₹1452.5. \end{aligned}$$

Therefore, total increase in salary is ₹1452.5. Since average annual percentage increase is 9.886, ₹1452.5 is equivalent to:

$$\begin{aligned} ₹7000 \times 0.09886 &= ₹692.02 \\ ₹7692.02 \times 0.09886 &= \underline{₹760.43} \\ \text{Total increase in salary} &= ₹1452.45. \end{aligned}$$

Another application of geometric mean is to find an average percentage increase over a period of time. The rate of increase is determined from the formula:

Average rate of increase over n time periods

$$= \sqrt[n]{\frac{\text{Value at the end of } n^{\text{th}} \text{ period}}{\text{Value at the start of first period}}} - 1. \tag{4.5.1}$$

Note that the formula in (4.5.1) is obtained from the compound interest formula

$$A = P\left(1 + \frac{r}{100}\right)^n.$$

Solving for r we obtain

$$r = \sqrt[n]{\frac{A}{P}} - 1.$$

Example 4.5.2. The rate of bonus given by LIC for a particular policy increased from ₹64 in 1992 to ₹100 in 2002. What is the average annual increase in bonus?

Now using (4.5.1),

$$\text{average rate of increase over 10 years} = \sqrt[10]{\frac{100}{64}} - 1 = 0.04564.$$

The value 0.04564 indicates that the average annual increase over the 10 years period was$0.04564 \times 100 = 4.564$ percent.

To do the above calculation in Excel, type the formula =(100/64)^ {10}-1 and press Enter key. Note that caret (^) is the symbol for exponentiation. For an array of positive numeric data the **GEOMEAN** function of Excel will compute geometric mean.

4.6 THE HARMONIC MEAN

The harmonic mean of a set of non-zero observations is the reciprocal of the arithmetic mean of the reciprocals of the values. Thus, the harmonic mean of n non-zero observations $x_1, \ldots, x_n$ is defined as follows:

$$\text{Harmonic mean} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}.$$

In Excel, the **HARMEAN** function gives the harmonic mean of a set of positive numbers.

The harmonic mean is rarely computed for data in a frequency table. We shall merely note the formula,

$$\text{harmonic mean} = \frac{1}{\frac{1}{N}\sum_{i=1}^{k}\left(\frac{f_i}{x_i}\right)} = \frac{N}{\sum_{i=1}^{k}\left(\frac{f_i}{x_i}\right)}; \quad N = \sum_{i=1}^{k} f_i.$$

It gives greater importance to small items and is useful only when small items have to be given a high weight.

Harmonic mean is useful in finding average speed when the speed for different parts of the journey is given in terms of distance covered per unit time. Harmonic mean is the suitable average for data sets which are customarily or conveniently given in terms of quantity purchased per rupee, questions answered per minute, and so forth.

Example 4.6.1. A car moves 5 km with speed 50 kmph, 10 kilometers with 60 kmph and 15 kilometers with speed 70 kmph. Find the average speed of the car.

The appropriate average in this case is the harmonic mean.

$$\text{Harmonic mean} = \frac{5+10+15}{\frac{5}{50}+\frac{10}{60}+\frac{15}{70}} = 62.376 \text{ kmph}.$$

Example 4.6.2. Given two values x_1 and x_2, prove that

$$\text{arithmetic mean} \geq \text{geometric mean} \geq \text{harmonic mean}.$$

Proof: We have,

$$\text{arithmetic mean} = \frac{x_1+x_2}{2}, \ \text{geometric mean} = \sqrt{x_1x_2}, \ \text{harmonic mean} = \frac{2x_1x_2}{x_1+x_2}.$$

Now,

$$(\sqrt{x_1}-\sqrt{x_2})^2 \geq 0$$

or

$$\begin{aligned} x_1+x_2-2\sqrt{x_1x_2} &\geq 0 \\ \frac{x_1+x_2}{2} &\geq \sqrt{x_1x_2} \\ \text{Arithmetic mean} &\geq \text{Geometric mean}. \end{aligned}$$

Again,

$$x_1 + x_2 \geq 2\sqrt{x_1 x_2}$$

or

$$\begin{aligned}(x_1+x_2)\sqrt{x_1x_2} &\geq 2(\sqrt{x_1x_2})^2\\ (x_1+x_2)\sqrt{x_1x_2} &\geq 2x_1x_2\\ \sqrt{x_1x_2} &\geq \frac{2x_1x_2}{x_1+x_2}\\ \text{Geometric mean} &\geq \text{Harmonic mean.}\end{aligned}$$

Hence,

$$\text{arithmetic mean} \geq \text{geometric mean} \geq \text{harmonic mean.}$$

4.7 PERCENTILES, DECILES AND QUARTILES

We have seen different measures of central tendency which can be used as a measure of some sort of central value of the data. However there are other ways of describing data by dividing the set of observations into equal parts. These measures include percentiles, deciles and quartiles.

The p-th *percentile* of a data set is a value such that at least p percent of the items take on this value or less and at least $(100-p)$ percent items take on this value or more. For example, the 20th percentile of a given data, which we shall write as P_{20}, is a value such that 20 percent of the observations are less than or equal to it and 80 percent of the observations are larger than or equal to it. Percentiles divide the data into hundred equal parts so that there are 99 percentiles. Note that median is 50th percentile. The determination of percentiles is exactly similar to that of median. The following are the steps for determining p-th percentile.

1. Arrange the observations in ascending order of magnitude.
2. Compute
$$i = \left(\frac{p}{100}\right) n$$
where p is the percentile of interest and n is the number of observations.
3. If i is not an integer, the next integer greater than i denotes the position of p-th percentile. If i is an integer, the p-th percentile is the arithmetic mean of the observations in positions i and $i+1$.

Example 4.7.1. Internet browsing time (in minutes) of a sample of 20 students in a college are 20, 7, 14, 6, 23, 9, 27, 12, 16, 15, 11, 7, 19, 5, 11, 7, 10, 18, 26, 21. Let us determine the 80th percentile.
On arranging the observations in ascending order we have

$$5,6,7,7,7,9,10,11,11,12,14,15,16,18,19,20,21,23,26,27.$$

$$i = \left(\frac{80}{100}\right) 20 = 16.$$

Since $i = 16$ is an integer, the 80-th percentile is the mean of 16th and 17th observations. Thus

$$P_{80} = \frac{20+21}{2} = 20.5.$$

Now comparing with the formula of median for grouped data given in (4.3.1) the formula for computing p-th percentile(P_p) from a grouped data is given by

$$P_p = L + \frac{c}{f}\left[\left(\frac{p}{100}\right)N - F\right],$$

where

L = the lower limit of the class containing P_p
p = the percentile of interest
N = the total frequency
f = the frequency of the class containing P_p
F = the total frequency upto the lower limit of the class containing P_p
c = the length of the class interval .

As percentiles, deciles and quartiles divide the data into several equal parts they are also called *partition values*. The graphical determination of the partition values is shown in the following example.

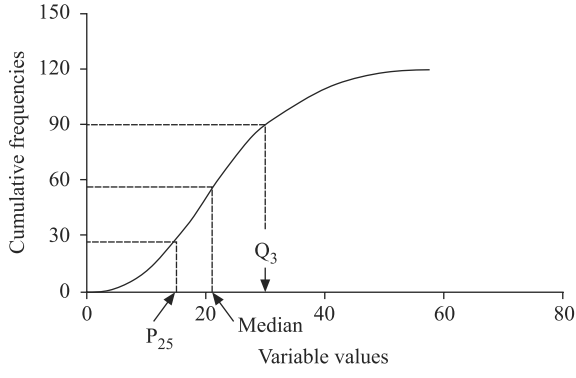
Example 4.7.2. The data on journey time of 120 students to reach college from their place of residence is given in the first two columns of the following table. Determine the partition values graphically.

Journey time (in minutes)	Number of students	Class mark	Cumulative frequency
0-5	0	2.5	0
5-10	5	7.5	5
10-15	14	12.5	19
15-20	19	17.5	38
20-25	24	22.5	62
25-30	21	27.5	83
30-35	13	32.5	96
35-40	10	37.5	106
40-45	7	42.5	113
45-50	4	47.5	117
50-55	2	52.5	119
55-60	1	57.5	120

The graphic determination of partition values is shown in Figure 4.2.

Percentile scores are frequently used to report results on such tests as the SAT, GMAT, GATE etc. Admission test scores for colleges and universities are frequently reported in terms of percentiles. For instance, suppose an applicant has a score of 430 of an admission test. It may not be readily apparent how this student performed relative to other students taking the same test. However if the score of 430 corresponds to the 78th percentile, then we know that 78% of the students had scores less than this

Figure 4.2: Graphic determination of partition values



student and 22% scored better. If you found that your GPA was in the 8th decile at your university, you could conclude that at least 80 percent of the students had a GPA lower than yours and at the most 20 percent had higher GPA.

Every 10th percentile, from 10 to 90, is called a *decile*. Deciles divide the data into 10 equal parts. *Quartiles* divide the data into 4 equal parts. Obviously, 25th percentile is called the lower quartile and is denoted by Q_1. The 75th percentile is called the upper quartile and is denoted by Q_3. The middle quartile, Q_2, which is the 50th percentile, is of course the median. The computation of deciles and quartiles are now apparent. The median and the other percentiles can also be obtained graphically from the ogive. For instance, Q_1 is the abcissae corresponding to the ordinate $\frac{N}{4}$.

The **PERCENTILE** function in Excel will give the specified k-th percentile of the given set of observations.

4.8 EXERCISES

1. What measures of central tendency do you suggest for:
   (a) Incomes of taxi drivers;
   (b) Number of coconuts in coconut trees in a plantation;
   (c) Shoe size of persons;
   (d) Weight of students in a class;
   (e) Family size of workers in a factory;
   (f) Mileage of vehicles per liter of petrol;
   (g) Number of petals of rose flowers in a garden;
   (h) Annual rate of interest when it is known that a sum doubles itself in 6 years;
   (i) Size of ready-made garments.
2. Which of the following statements are true?

(a) The mean, the median, and the mode represent frequencies of values.

(b) The mean, the median, and the mode represent values of the variable.

(c) Only the median and the mode represent frequencies.

(d) Only the arithmetic mean represent values of the variable.

3. The average which represents the value of a total when shared out equally is:
   (a) arithmetic mean; (b) median; (c) mode; (d) geometric mean.

4. A clothes store manager has sales data of trouser sizes for the last month's sales. Which measure of central tendency should the manager use, if the manager is interested in the most sellable size?
   (a) arithmetic mean; (b) median; (c) mode; (d) geometric mean.

5. The mean of five numbers is 20 and the three of the five numbers are 10, 18 and 22. The remaining two values share the same value, which is:
   (a) 50; (b) 25; (c) 15; (d) none of these.

6. The mean age of 50 students in a class is 20 years. When the age of the teacher is included, the mean age increased by one year. The age of the teacher is:
   (a) 71; (b) 55; (c) 51; (d) 50.

7. The mean of 12 numbers is 21 and their median is 19. Suppose the largest number is increased by 4 and the smallest number is reduced by 7, the mean and median of the modified set of numbers are, respectively:
   (a) 20.75, 19; (b) 21, 18.75; (c) 20.75, 18.75; (d) 21, 19.

8. Sum of the deviation about mean is:
   (a) zero; (b) minimum; (c) maximum; (d) none of these.

9. Sum of the absolute deviations about median is:
   (a) zero; (b) minimum; (c) maximum; (d) none of these.

10. The suitable measure of central tendency for qualitative data is:
    (a) mode; (b) arithmetic mean; (c) geometric mean; (d) median.

11. In a class test, 40 students out of 50 passed with mean marks 60 and the overall average of class marks was 55. The average marks of students who failed were:
    (a) 25; (b) 30; (c) 48; (d) 35.

12. The average marks of section A are 65 and that of section B are 70. The average of both the sections combined is 67. The ratio of number of students of section A to B is:
    (a) 1:3; (b) 2:3; (c) 3:1; (d) 3:2.

13. Harmonic mean give more weight to:
    (a) small values; (b) large values; (c) positive values; (d) negative values.

14. If we plot the more than type and the less than type cumulative frequencies of the same set of data, their graphs intersect at the point which is known as:
    (a) median; (b) mode; (c) mean; (d) none of these.

15. The average of $2n$ natural numbers from 1 to $2n$ is:
    (a) $\frac{n+1}{2}$; (b) $\frac{2n+1}{2}$; (c) $\frac{n(n+1)}{2}$; (d) $\frac{2n+1}{4}$.

16. There were 25 teachers in a school whose mean age was 40 years. A teacher retired at the age of 55 years and a new teacher was appointed in his place. The mean age of teachers in the school was reduced by one year. The age of the new teacher was:
    (a) 25 years; (b) 30 years; (c) 35 years; (d) 40 years.

17. A man goes from his house to his office at the speed of 20km/h and returns from his office to home at the speed of 30 km/h. His average speed in the to and fro journey is:
    (a) 24 km/h; (b) $10\sqrt{6}$ km/h; (c) 25 km/h; (d) none of these.

18. The value of the variable corresponding to the highest point of a frequency curve represents:
    (a) mean; (b) median; (c) mode; (d) none of these.

19. The average of proportions 0.16 and 0.01 is:
    (a) 0.40; (b) 0.085; (c) 0.04; (d) none of these.

20. For percentiles, the total number of partition values are:
    (a) 100; (b) 99; (c) 10; (d) none of these.

21. The second decile divides a data series in the ratio:
    (a) 1:1; (b) 1:2; (c) 1:4; (d) 2:5.

22. Which average has the same number of observations above it and below it?
    (a) mean; (b) median; (c) mode; (d) geometric mean.

23. Which measure of central tendency can have more than one value?
    (a) mean; (b) median; (c) mode; (d) harmonic mean.

24. The mean monthly salary of 12 workers and 3 managers in a factory was ₹ 26000. When one of the managers whose salary was ₹27500, was replaced with a new manager, the mean salary of the team went down to ₹25000. What is the salary of the new manager?

25. Suppose your grades on three mathematics examinations are 81, 92 and 94. What grade do you need on your next examination to have a 90 average on the four exams?

26. When a student weighing 45 kgms left a class, the mean weight of the remaining 59 students increased by 200 gm. What is the mean weight of the remaining 59 students?

27. A company has 20 junior executives, 6 senior executives and 2 managers. During one month the mean salary for junior executives was ₹18,000, that of senior executives was ₹24,000 and that of the managers was ₹ 35,000. Find the mean salary of all the 28 employees combined.

28. The mean age of a family of 5 members is 20 years. If the age of the youngest member is 10 years, what was the mean age of the family at the time of the birth of the youngest member?

29. The number of rooms occupied in a hotel during 7 days in a week are 71, 83, 68, 74, 82, 79 and 90. Find the average number of rooms occupied.

30. The arithmetic mean of 20 observations is -8.5. If the observation -11.5 is replaced by -1.5 then find the new arithmetic mean.

31. Give a set of data for which the mean, the median, and the mode coincide.

32. Measure the height of each student in your class to the nearest centimeter. Find the mean, median and mode and compare all three measures. Which value gives the best measure of central tendency? Why? Which organizations or companies would find such statistics useful?

33. Your younger brother comes home one day after taking a science test. He says that someone at school told him that "60% of the students in the class scored above the median test grade". What is wrong with this statement? What if he said "60% of the students scored below the mean?"

34. A psychologist assigned identical tasks to each of eight mentally challenged children. The following figures give the time (in minutes) that each child took to complete the task: 30, 34, 42, 54, 58, 62, 136 and 170. What measure of central tendency do you think is most appropriate? What is its value?

35. Compute the mean, median and mode from the following stem and leaf chart.

5	58
6	257
7	2347779
8	0148
9	001589

36. Give three different sets of data each with five observations having the same mean and median but different mode.

37. When is the median preferred to the arithmetic mean?

38. For an ordinal data, can you think of occasions where the mode would be of more use than the median or mean? Discuss it in your classroom.

39. Why do we need averages?

40. The mean weight of 10 apples in the first box was 210 grams, of 15 apples in the second box was 235 grams, and of 20 apples in the third box was 230 grams. Find the mean weight of all the apples in the three boxes.

41. The mean score of a student on the first three tests is 57 points. How many points must he score on the next two tests if his overall mean score has to be 60 points?

42. During the one hour period 11 am to 12 noon on a certain day a shopkeeper observed that customers arrived at the following hours: 11:02, 11:06, 11:08, 11:13, 11:25, 11:31, 11:42, 11:48, 11:53, 11:57 and 11:59. If the time between the consecutive arrivals is called inter-arrival time, find the mean inter-arrival time.

43. The number of vehicles passing a certain fixed location in a highway in 10 intervals of five minutes duration each were recorded as 9, 7, 10, 5, 8, 5, 10, 12, 10 and 11. Compute the median of the data.

44. The office of the Dean of the Faculty of Science at a certain university compiled the following information regarding the ages of 52 students who were awarded Ph.D. degrees during the last academic year.

Age	25-30	30-35	35-40	40-45	45-50	50-55
Number of students	6	17	14	8	4	3

Find the mean age of the students who were awarded Ph.D. degrees.

45. Three mathematics classes A, B and C take an algebra test. The average score in class A is 82. The average score in class B is 75. The average score in class C is 85. The average score of all students in classes A and B together is 79. The average score of all students in classes B and C together is 81. What is the average score for all the three classes, taken together?

46. The marks secured by a student for three subjects are as follows: Statistics-180/200, Physics-215/300, Mathematics-565/600. Find the average score.
If these subjects are considered to vary in their importance and are given weights 3, 2 and 15 respectively, find the average score in this case.

47. A student finds the average of 10 positive integers. Each integer contains two digits. By mistake, the boy interchanges the digits of one number say ba for ab. Due to this, the average becomes 2.7 less than the previous one. What was the difference of the two digits a and b?

48. Show that the weighted arithmetic mean of first n natural numbers whose weights are equal to the corresponding number is equal to $(2n+1)/3$.

49. In a township in Chennai, all plots are 10, 15, 20 and 30 cents. The frequency table of plot sizes for all residential property in this township is given below:

Plot size (in cents)	10	15	20	30
Number of plots	50	150	35	15

(a) Is the median bigger than the mode?

(b) If a real estate agent says that typical size of the plot is 16 cents, what sort of average is the agent using? How many plots are of this "typical" size?

50. The following table gives the data on the holdings of shares in a certain company.

Number of shares held	below 25	25-50	50-75	75-100	100-150	150-200	above 200
Number of persons	130	180	215	150	90	30	10

Calculate the median holdings per person.

51. A frequency table has six consecutive intervals of equal width. The midpoint of the second class interval is 22 and the right endpoint of the fourth class interval is 42. The respective frequencies are 6, 12, 19, 14, 6 and 3. Find the arithmetic mean and the mode.

52. A market survey questionnaire asked its respondents to mark their opinion about a particular design of a product on a scale from 0(strongly dislike) to 6(strongly like), where the score 3 implies neutral. The responses obtained from several persons are reported below:

Score	0	1	2	3	4	5	6
Number of respondents	9	15	22	7	27	33	17

Find the mode and the median of the opinion score.

53. In a frog race with ten frogs, three frogs strayed away. For the remaining frogs that completed the race, the following times (in seconds) were recorded: 345, 250, 360, 372, 450, and 600. Pick an appropriate measure of central tendency and find its value.

54. Data on the duration of advertisements per half-hour of prime-time television programs on major networks at 8.00pm are given below:

$$6.8, 5.7, 7.1, 6.3, 6.0, 7.3, 5.9, 7.5, 6.9, 6.1, 6.5, 6.7, 6.5, 7.2, 7.0$$

Compute the mean and the median. Using the mean, what percentage of viewing time is spent on prime-time advertisements? What percentage of time is spent on the programs themselves?

55. A shipment of thirty boxes containing a certain type of electronic item is received. Upon inspection it was found that twelve boxes contained no defective items, eight boxes contained 1 defective item each, six boxes contained 2 defective items each, four boxes contained 3 defective items each, and the remaining boxes contained 4 defective items each. What are the mean, median and mode of this data?

56. The data on weight (in kilograms) of 1200 students in a college is shown in the following table.

Weight	Number of students
40-50	160
50-55	200
55-60	360
60-65	240
65-70	160
70-80	80

Compute the arithmetic mean, median and mode of the data.

57. A man travels from his home to office in a car. For his onward journey he gets a mileage of 8 miles per litre of petrol. On the return trip he gets a mileage of 12 miles per litre. Find the average mileage of the car for his to and fro journey. (Hint: Compute H.M.).

58. The increase in price of rice was 2% in 2006, 16% in 2007 and 2% in 2008. It is said that the average price rise during the period 2006-2008 was 4% and not 6.7%. Justify the statement and show how you would explain it to a layman.

59. Find the A.M., G.M. and H.M. of the values $a, ar, ar^2, \ldots, ar^{n-1}$ and show that A.M. $\times$ H.M. = G.M.^{2}.

60. Petrol is sold in three towns at the rate of $1.52, 1.58$ and 1.62 liters per 100 rupees. Find the average quantity of petrol in liters which can be bought for ₹100.

61. Suppose a person earned ₹50,000 in the year 2003 and ₹3,00,000 in the year 2011, what is his mean annual rate of increase over the period?

62. An incomplete frequency table is given as follows:

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
Frequency	9	25	?	35	21	12	?	130

Given that the median is 32, determine the missing frequencies.

Chapter 5

MEASURES OF DISPERSION AND SKEWNESS

"Statistics is aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other"- Horace Secrist.

5.1 INTRODUCTION

In the previous chapter we saw that averages provide useful summary information concerning the general level of data. However, this does not tell us anything about the spread of the data. For example, suppose you want to cross a river and it is known that the average depth of the river is 3 feet. Would you want to wade across on foot without knowing the maximum depth? Probably not. If the depth of the river varies between 2 feet and 4 feet you would probably agree to cross, what if you learned the river depth varies from 2 feet to 7 feet? Your decision would probably be not to cross. Before making a decision about crossing the river, you want to know both the typical depth and the variation in the depth of the river.

As another example, suppose a test is given to two groups of students and the scores obtained are given below.

Group I	:	44	42	40	40	38	36
Group II	:	70	65	40	40	15	21

We can see that both groups have the same arithmetic mean (40), the same median score (40) and the same mode (40). Though the averages are same the two sets of data are not identical. These examples clearly illustrate that a measure of central tendency describes only one of the important characteristics of the data. For a good understanding of the data, we must also know the extent of variability. The variability of the data is also referred to as the spread or dispersion or scatter of the data. Measures of dispersion are used to provide a numerical summary of the level of variation present in the measurements in respect of how the data cluster around their 'central' value.

The commonly used measures of variability are (1) the range (2) the mean deviation (3) the standard deviation and (4) the interquartile range.

5.2 THE RANGE

The simplest measure of dispersion is the range. The *range* is defined as the difference between the largest and the smallest values in the data.

Range = largest value - smallest value.

The range is widely used in statistical process control applications, weather reports, stock exchange quotations etc.

The main disadvantage of range is that it ignores the intermediate values. Consider for instance, the following two sets of data:

Set 1 : 151, 170, 156, 165, 150, 169, 174, 175, 171, 160
Set 2 : 150, 174, 174, 170, 175, 172, 175, 169, 173, 169.

On comparing the scattering of the above two sets of observations it is obvious that the observations in the first set are more scattered than that of the other set. But they both have the same range even if the two sets are markedly different with regard to scattering of observations.

By describing data solely in terms of extreme values, the presence of an outlier (unusually extreme value) will inflate the range out of proportion. This is another demerit of the range. However, it is simple in concept and easy to calculate.

The **MAX** and **MIN** functions identify the maximum and minimum values respectively in a set of values. Suppose the values are entered in the first raw of the worksheet from cells A1 to L1. Then type the formula '=MAX(A1:L1)-MIN(A1:L1)' in a blank cell and press Enter to obtain the range.

5.3 THE MEAN DEVIATION

The variability of the data depends upon the extent to which individual observations are spread about a central value. If the values are widely scattered then the distances to the values from this central value will be large and will have large variability. On the other hand, if the values are compactly located around a central value then the variability will be small.

Mean deviation about the mean value measures the average distance by which the values lie from their mean value. *Mean deviation about the mean value* is defined as the arithmetic mean of the absolute values of the deviations of the observations or

data points from their arithmetic mean.

$$\text{Mean deviation about mean} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

where $|x_i - \bar{x}|$ denotes the absolute difference between x_i and $\bar{x}$. That is, if $x_i - \bar{x} > 0$ then $|x_i - \bar{x}| = x_i - \bar{x}$ and if $x_i - \bar{x} < 0$ then $|x_i - \bar{x}| = \bar{x} - x_i$. For example, $|3-7| = |-4| = 4 = 7-3$.
For data presented in a frequency table,

$$\text{mean deviation about the mean value} = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{x}|}{N}$$

where $N = \sum_{i=1}^{k} f_i$.

$$\text{Mean deviation about the median} = \frac{\sum_{i=1}^{n} |x_i - \text{median}|}{n}.$$

Similarly, we can define mean deviation about the median or about any arbitrary value A.

Why do we consider absolute deviations? If we didn't then the positive and negative deviations may cancel each other and the resulting value would be a useless measure. Note that $\sum_{i=1}^{n} (x_i - \bar{x})$ is always zero whatever be the numbers $x_1, x_2, \ldots, x_n$.

It can be shown that the sum of absolute deviations of observations is minimum when it is taken from the median. That is $\sum_{i=1}^{n} |x_i - A|$ is minimum when A=median.

In contrast with the range, mean deviation takes into account the magnitudes of all the observations. Mean deviation from the point A is also a mathematical distance of the observations $x_1, x_2, \ldots, x_n$ from the fixed point A.

Example 5.3.1. The weights, in grams, of 10 pieces of scrap metal from a metal fabrication process are 2.63, 2.68, 2.57, 2.73, 2.56, 2.62, 2.38, 2.45, 2.35, 2.53. Compute the mean deviation about the mean and mean deviation about the median.
The arithmetic mean of the above values is $\bar{x} = 2.55$ and the median is $M = 2.565$. Now, let us form the following table for the purpose of computation.

x_i	2.63	2.68	2.57	2.73	2.56	2.62	2.38	2.45	2.35	2.53
$\lvert x_i - \bar{x}\rvert$	0.08	0.13	0.02	0.18	0.01	0.07	0.17	0.10	0.20	0.02
$\lvert x_i - M\rvert$	0.065	0.115	0.005	0.165	0.005	0.055	0.185	0.115	0.215	0.035

$$\sum_{i=1}^{n} |x_i - \bar{x}| = 0.98, \quad \sum_{i=1}^{n} |x_i - M| = 0.96.$$

$$\text{Mean deviation about the mean value} = \frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} = \frac{0.98}{10} = 0.098 \text{ gms.}$$

$$\text{Mean deviation about the median} = \frac{\sum_{i=1}^{n}|x_i-M|}{n} = \frac{0.96}{10} = 0.096 \text{ gms.}$$

Note that if you compute mean deviation about any arbitrary value A (other than the median), you will get a value greater than 0.096 gms. This is the meaning of saying that the mean absolute deviation is least when the deviations are taken from the median.

Computation Using MS Excel

The **AVEDEV** function in Excel produces the mean deviation about mean only. The steps for computation of mean deviation about median in Example 4.3.1 can be given as follows:

1. Enter the given ten values in the cells from A1 to A10.
2. Compute median by entering the formula '=MEDIAN(A1:A10)' in a blank cell, say A13. This gives the vaule 2.565.
3. Create the formula '=ABS(A1- 2.565)' in cell B1 and press Enter.
4. Copy the formula to the cells B2,...,B10 by dragging from the rightmost bottom corner of the selection box of cell B1.
5. Select cell B11, click Σ or type the formula '=sum(b1:b10)' and press Enter to obtain the sum of values in the cells B1,...,B10.
6. Select an empty cell and enter the formula '=B11/10' to obtain the final answer.

5.4 THE VARIANCE AND THE STANDARD DEVIATION

Since the sum of the deviations of observations from their arithmetic mean is zero we took absolute deviations to calculate mean deviation about the mean value. Another way to eliminate the signs of deviations is to square the individual deviations, thereby getting positive numbers for both positive and negative deviations.

The *variance* of a set of observations $x_1,\ldots,x_n$ is defined as the arithmetic mean of the squares of the deviations of the individual observations from their arithmetic mean.

$$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}. \tag{5.4.1}$$

Since the positive square root of the variance is another mathematical distance of the observations from $\bar{x}$ we take the positive square root to get another measure of

scatter. The positive square root of the variance is called the *standard deviation*.

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}.$$

The standard deviation of a set of observations will be always positive even if all the observations in the data are negative. Standard deviation provides a summary of variability or spread or scatter which reflects how scattered are the measurements from their arithmetic mean.

The formula for variance given in (5.4.1) can also be expressed as:

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2. \tag{5.4.2}$$

Proof:

$$\begin{aligned}
\text{Variance} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2)\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right], \text{since } \bar{x}^2 \text{ is free of } i \text{ and } \sum_{i=1}^{n}x_i = n\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2.
\end{aligned}$$

The formula in (5.4.2) is used for computational purpose as it is more convenient for manual computations.

For a frequency table with k class intervals the corresponding formula for variance is the following:

$$\text{Variance} = \frac{\sum_{i=1}^{k}f_i(x_i-\bar{x})^2}{N} = \frac{1}{N}\left[\sum_{i=1}^{k}f_ix_i^2\right] - \bar{x}^2$$

where

$$\sum_{i=1}^{k}f_i = N \text{ and } \bar{x} = \frac{\sum_{i=1}^{k}f_ix_i}{N}.$$

If there is no variation at all, that is, all the observations are equal, then the standard deviation is zero. If at least two observations are different then the standard deviation is a positive quantity.

Example 5.4.1. The heights of five people in centimeters are 183, 163, 172, 157, and 169. Find the standard deviation of this data.

Here

$$\Sigma x_i^2 = 142852,\ \bar{x} = 168.8 \text{ cm.}$$

$$\text{Variance } = \frac{1}{5}(142852) - 168.8^2 = 76.96.$$

$$\text{Standard deviation} = \sqrt{76.96} = 8.7727.$$

Properties of Standard Deviation

1. Standard deviation is unaffected by the change of origin or change of location of the observations. That is, the standard deviation is unaffected by adding or subtracting a constant value to each of the observations in a given set of data.
   Proof: Let $x_1, \ldots, x_n$ be a set of n observations and let $y_i = x_i + c,\ i = 1, \ldots, n$ be a new set of observations obtained by adding a constant 'c', c can be either positive or negative, to each x_i. Then the standard deviation of $y_1, \ldots, y_n$ will be the same as the standard deviation of $x_1, \ldots, x_n$. Observe that

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n (x_i + c)}{n} = \frac{1}{n}\left[\sum_{i=1}^n x_i + nc\right] = \frac{\sum_{i=1}^n x_i}{n} + c = \bar{x} + c.$$

$$\begin{aligned}
\text{Standard deviation of } y_i\text{'s} &= \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^n [(x_i + c) - (\bar{x} + c)]^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \\
&= \text{Standard deviation of } x_i\text{'s.}
\end{aligned}$$

   Hereafter we will abbreviate standard deviation as S.D.

2. Standard deviation is affected by change of scale. That is, when we multiply all the values in a set of data by a constant, the constant being positive or negative, to form the new set of values, then the standard deviation of the new set of values also gets multiplied by the absolute value of the constant.
   Proof: Let $x_1, \ldots, x_n$ be a set of observations. Define $y_i = cx_i,\ i = 1, \ldots, n$. Then the standard deviation of $y_1, \ldots, y_n$ is given by

$$\begin{aligned}
\text{S.D.}(y) &= \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum_{i=1}^n (cx_i - c\bar{x})^2}{n}} \\
&= \sqrt{\frac{c^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n}} = |c|\ \text{S.D.}(x);\quad \sqrt{c^2} = |c|.
\end{aligned}$$

Hence

$$\text{S.D.}(x) = |c| \text{ S.D.}(y).$$

Thus the variance and the S.D. are affected by change of scale.

The standard deviation is the least possible root mean square deviation of a set of observations because the sum of squares of deviations of the observations is minimum when the deviations are taken from the arithmetic mean.

Example 5.4.2. The data of breaking strengths (in pounds) of 120 cables is shown in the following table.

Breaking strength	Number of cables
140-150	16
150-160	20
160-170	36
170-180	24
180-190	16
190-200	8

Find the standard deviation by changing the origin and scale.

For calculating the standard deviation we shall form the following table.

Class mark (x_i)	f_i	$y_i = \frac{x_i - 165}{10}$	$f_i y_i$	y_i^2	$f_i y_i^2$
145	16	-2	-32	4	64
155	20	-1	-20	1	20
165	36	0	0	0	0
175	24	1	24	1	24
185	16	2	32	4	64
195	8	3	24	9	72
Total	120		28		244

Here, 165 is a convenient number as the constant to be subtracted. Any other number could have been taken. In this case the length of the class interval is an appropriate number as the scale constant 'c'.

$$\sum f_i y_i^2 = 244, \bar{y} = \frac{28}{120} = 0.2333.$$

$$S.D.(y) = \sqrt{\frac{244}{120} - 0.2333^2} = 1.4067.$$

Therefore, since c here is positive, $c = 10$,

$$S.D.(x) = 10 \times 1.4067 = 14.067 \text{ pounds.}$$

One can compute S.D.(x) directly by using the first two columns of the above table and can verify that both the answers are the same.

Computation Using MS Excel

The **STDEV** function in Excel produces the standard deviation of a given set of values. The steps for computation of standard deviation for the data in Example 5.4.2 is as follows:

1. Enter the class marks (x_i) in the cells A1 to A6 and corresponding frequencies (f_i) in the cells B1 to B6.

2. Type the formula '=A1*B1' in C1 and copy the formula to C2,...C6.

3. Type the formula '=A1^ 2*B1' in D1 and copy the formula to D2,...D6.

4. Select cells B7, C7 and D7, click Σ and press Enter to obtain the sums $\sum_{i=1}^{6} f_i$, $\sum_{i=1}^{6} f_i x_i$ and $\sum_{i=1}^{6} f_i x_i^2$ respectively.

5. Select an empty cell and enter the formula '=SQRT((D7/B7)-(C7/B7)^ 2)' to obtain the final answer.

Combined Variance

If two sets of data having respective sample sizes n_1, n_2, means $\bar{x}_1, \bar{x}_2$ and standard deviations s_1, s_2, then the standard deviation of the combined set of values is given by

$$s^2 = \frac{1}{n_1+n_2}[n_1(s_1^2+d_1^2)+n_2(s_2^2+d_2^2)]$$

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$ and $\bar{x} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$.

Proof: Let $x_{1i}, i = 1,2,\ldots,n_1$ and $x_{2j}, j = 1,2,\ldots,n_2$ be the two sets of observations. Then we have

$$\begin{aligned} s_1^2 &= \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 \\ s_2^2 &= \frac{1}{n_2}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2 \end{aligned}$$

where $\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}$ and $\bar{x}_2 = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2}$.

The average of the combined set of observations is

$$\bar{x} = \frac{\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j}}{n_1+n_2} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}.$$

The variance of the combined series is given by

$$s^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x})^2\right]. \tag{5.4.3}$$

Now

$$\begin{aligned}\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 &= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1+\bar{x}_1-\bar{x})^2\\ &= \sum_{i=1}^{n_1}[(x_{1i}-\bar{x}_1)^2+(\bar{x}_1-\bar{x})^2+2(\bar{x}_1-\bar{x})(x_{1i}-\bar{x}_1)]\\ &= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2+n_1(\bar{x}_1-\bar{x})^2, \text{ since } \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)=0\\ &= n_1s_1^2+n_1d_1^2.\end{aligned}$$

Similarly

$$\sum_{i=1}^{n_1}(x_{2j}-\bar{x})^2 = n_2s_2^2+n_2d_2^2.$$

On substituting in (5.4.3) we obtain

$$s^2 = \frac{1}{n_1+n_2}[n_1(s_1^2+d_1^2)+n_2(s_2^2+d_2^2)]. \tag{5.4.4}$$

Remark 5.4.1. The formula (5.4.4) can be generalized to the case of k sets of data as follows:

$$s^2 = \frac{1}{n_1+\cdots+n_k}[n_1(s_1^2+d_1^2)+\cdots+n_k(s_k^2+d_k^2)]$$

where $d_i = \bar{x}_i - \bar{x}$, $i = i,\ldots,k$ and $\bar{x} = \frac{n_1\bar{x}_1+\cdots+n_k\bar{x}_k}{n_1+\cdots+n_k}$.

Coefficient of Variation

What we have studied earlier are some measures of dispersion. A comparative idea of the degree of variation is not fully brought out by these measures of dispersion. For example, suppose we want to compare the variation in heights of boys aged 5 and 15 years with mean heights 100cms and 150cms respectively. Suppose that the corresponding standard deviations are obtained as 6cms and 9cms respectively. Since the mean height of boys aged 15 years is likely to be higher than that of boys aged 5 years the standard deviations as such cannot give a true comparative idea of the variation. A first look at the given information would lead us to believe that the variation in the heights of boys aged 15 years is greater. But the ratio of the standard deviation to the mean height in both the cases are equal to 0.6. Thus the relative variations, relative to the averages, are the same for both the groups.

As another example, suppose we want to compare the variation in salaries of professors in the U.S.A. and India. Since the salary in the U.S.A. is in dollars and that in India is in rupees, how can we compare the variation in salaries in the two countries? Hence to put variation in different sets of data in their proper perspective, a relative measure of dispersion is desirable. The most widely used relative measure of dispersion is the coefficient of variation introduced by Karl Pearson and it is defined as follows:

$$\text{Coefficient of variation } = \frac{\text{standard devation}}{\text{mean}} \times 100, \text{ when mean} \neq 0.$$

Thus, coefficient of variation is simply the standard deviation expressed as a percentage of mean.

The standard deviation is a measure of dispersion within the data, the value of which depends on the scale of measurement of the data. The coefficient of variation provides a measure of relative dispersion of data and is a measure of data variability which is free of the units of measurement as long as the scale of measurement is positive. In the ratio, standard deviation divided by the arithmetic mean, the scaling effect when the scaling constant is positive, will get canceled. The coefficient of variation does not change if the (linear) scale, but not the location of the data is changed. That is, if you take data $x_1,\ldots,x_n$ and transform it to new data, $y_1,\ldots,y_n$ using the mapping $y_i = ax_i + b$, the coefficient of variation of $y_1,\ldots,y_n$ will be the same as the coefficient of variation of $x_1,\ldots,x_n$ if $b = 0$ and $a > 0$, but not otherwise. So, the coefficient of variation would be the same for a set of length measurements whether they were measured in centimeters or inches (zero is the same on both scales). In this sense coefficient of variation is more useful when comparing different data sets which may differ in the scale of measurement. However, the coefficient of variation would be different for a set of temperature measurements made in Celsius and Fahrenheit (as the zero of the two scales is different). The coefficient of variation is also referred to as the *relative standard deviation* (RSD), relative to the arithmetic mean.

Coefficient of variation is used for comparing data sets. It is often required to compare the performance of two candidates in an examination or of two players, given their scores in various papers or games. The smaller the coefficient of variation, one may interpret as more consistent is the performance. The following example will illustrate the procedure.

Other relative measures of dispersion have been suggested but are not popular. They are the following:

Coefficient of range = $\dfrac{L-S}{L+S}$ where L is the largest value and S is the smallest value.

Coefficient of quartile deviation= $\dfrac{Q_3-Q_1}{Q_3+Q_1}$, where Q_1 is the lower quartile and Q_3 is the upper quartile.

$$\text{Coefficient of mean deviation} = \frac{\text{Mean deviation about median}}{\text{median}} \times 100.$$

Similarly we have other coefficients of mean deviations.

Example 5.4.3. Two candidates Manu and Tintu at the B.Sc. Examination obtained the following marks in ten papers.

Paper	I	II	III	IV	V	VI	VII	VIII	IX	X
Manu (x)	28	39	36	35	43	38	32	40	41	38
Tintu (y)	34	39	38	31	39	35	37	40	36	31

Which of them showed a more consistent performance?

We have to compute the arithmetic mean and standard deviation for each of the candidates so that corresponding coefficients of variation can be calculated.

$$\bar{x} = 37,\ \ \bar{y} = 36,\ \ S.D.(x) = 4.22,\ \ S.D.(y) = 10.42.$$

$$\text{C.V.}(x) = 11.4\%,\ \ \text{C.V.}(y) = 28.9\%.$$

Since coefficient of variation for Manu is smaller than that of Tintu, Manu is more consistent than Tintu.

5.5 THE INTERQUARTILE RANGE

Another common measure of variability is the *interquartile range*, which is defined as the difference between the upper quartile (Q_3) and the lower quartile (Q_1). The interquartile range measures the spread bounding the middle 50 percent of the values of the data. It tells nothing about the dispersion of values around the average. Half of the interquartile range is called the *quartile deviation*. The quartile deviation is also called *semi-interquartile range*. That is,

$$\text{quartile deviation} = \frac{1}{2}(Q_3 - Q_1).$$

Similarly the difference between 90th percentile (P_{90}) and 10th percentile (P_{10}) is a possible measure of scatter; so is the difference between the 99th percentile and the 1st percentile. The former measures the spread bounding the middle 80 percent of the values of the data, and the latter measures the spread bounding the middle 98 percent of them.

5.6 SKEWNESS IN THE DATA

Two sets of data may have the same average and the same measure of dispersion but still they may differ from each other very much. For example, the observations in one data may be spread in a symmetric manner about a central value, the other may be lacking in symmetry. But both sets of data can have the same mean and standard deviation. Thus the two numerical characteristics - central tendency and dispersion - are not sufficient to describe a data. We now describe another characteristic of the data called skewness.

If the observations in a data are spread in a perfect symmetrical manner about an average, then the mean and median are equal. If the data has a single mode and it is perfectly symmetric, then the mean,the median and the mode coincide. An example of a symmetric data presented in a frequency table is given below:

Height of students (in cms)	155-160	160-165	165-170	170-175	175-180
Number of students	6	14	25	14	6

Most of the data are not symmetric but are asymmetric. Asymmetry in the data can be of different types. To explain the idea consider the histogram of a data. Take the

median point of a data, so that the area obtained by taking the product of frequency and corresponding class interval on the left of the median is the same as that on the right of the median. If the frequencies and class intervals on the right of the median is an exact replica of those on the left of the median then the data is symmetric, if not the data is asymmetric. Asymmetry may happen when on one side the data is spread on a wider range so that the arithmetic mean of the whole data may lie on to the right or to the left of the median. If the arithmetic mean is shifted to the right of the median, as a result wider spread with smaller frequencies, then we say that the data is *skewed to the right* or *positively skewed*. If the same thing happens on the left of the median then we say that the data is *skewed to the left* or *negatively skewed*.

For a positively skewed data the right tail of the corresponding frequency curve or polygon is longer than the left tail and the opposite is the case when the data is negatively skewed.

Skewness is zero when the data is symmetric. "How much skewed?", cannot be answered by looking at these figures. There are different numerical measures of skewness proposed by various people.

The relative positions of the mean, the median and the mode give us the first measure of skewness. Karl Pearson suggested the following measures of skewness based on the relative position of the mean, the median and the mode.

$$\text{Karl Pearson's coefficient of skewness} = \frac{\text{mean-mode}}{\text{standard deviation}}. \tag{5.6.1}$$

The division with the standard deviation in the formula (5.6.1) is justified because it will make the measure of skewness free of the units in which the variable is expressed.

In cases where unique mode doesn't exist the following formula is used in place of (5.6.1).

$$\text{Coefficient of skewness} = \frac{3(\text{mean-median})}{\text{standard deviation}}.$$

In a symmetrical case, the lower and upper quartiles occur at equal distances on either side of the median (Q_2). Thus the quantity $(Q_3 - Q_2) - (Q_2 - Q_1)$ can be used for measuring skewness. To remove the scale factor we may divide by the interquartile range. Thus we have another measure of skewness.

$$\text{Bowley's coefficient of skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}.$$

Bowley's measure of skewness is based on the middle 50% of the observations because it leaves 25% of the observations on each extreme. As an improvement over Bowley's measure, Kelly has suggested a measure based on P_{90} and P_{10} so that only 10% of the observations on each extreme are ignored.

$$\text{Kelley's coefficient of skewness} = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}.$$

The resulting sign of the coefficient of skewness calculated using the above formulae agrees with the sign we have attached to skewness. There are other measures of

skewness which we shall discuss later. Usually, data on wages or salaries, consumption of electricity, weights of adult men etc. are skewed to the right. Data on the ages at death of inventors may be characteristically skewed to the left, since younger men do not often have enough inventions to their credit to be classified as "inventors".

We know that if the data is symmetric, then the mean, the median, and the mode coincide. Is it true that if the mean, median, and the mode coincide, then the data must be symmetric? The answer is no. Here are some examples.

Example 5.6.1. The number of flowers in 30 plants are noted and obtained the following data.

Number of flowers	0	1	2	3	4	5	6	7
Number of plants	2	3	9	10	4	2	1	1

Obviously the mode of the data is 3. Note that the frequencies of the values on the left of 3 is not the same as that on the right of 3. Hence the data is not symmetric.
Since the number of observations is 30, an even number, the median is the mean of 15th and 16th largest observations. Therefore, the median is also equal to 3.
Now

$$\text{arithmetic mean} = \frac{\Sigma f_i x_i}{N} = \frac{90}{30} = 3.$$

This means that for asymmetric data also

$$\text{arithmetic mean} = \text{median} = \text{mode}.$$

Example 5.6.2. A farmer has obtained girths of 7-year-old rubber trees planted in an acre of land and the data is summarized in the first two columns of the following table.

Girth (in inches)	Number of trees (f_i)	Cumulative Frequency	Class marks (x_i)	$f_i x_i$	frequency density
11-14	17	17	12.5	212.5	17/3
14-16	23	40	15	345	23/2
16-18	45	85	17	765	45/2
18-20	50	135	19	950	50/2
20-22	35	170	21	735	35/2
22-26	25	195	24	600	25/4
Total	195			3607.5	

$$\begin{aligned} \text{Arithmetic mean} &= \frac{\Sigma f_i x_i}{N} = \frac{3607.5}{195} = 18.5. \\ \text{Median} &= L_m + \frac{c}{f_m}\left(\frac{N}{2} - F\right). \end{aligned}$$

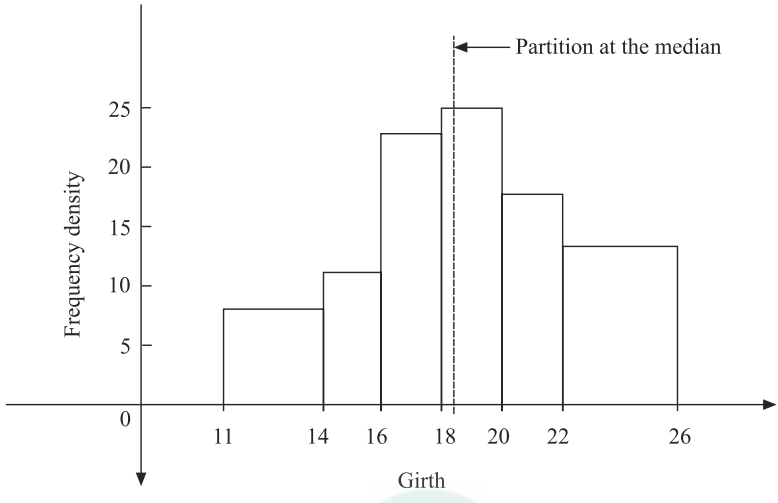
Here

$$L_m = 18,\ c = 2,\ f_m = 50,\ F = 85.$$

Thus,

$$\text{median} = 18.5.$$

Figure 5.1: Histogram of an asymmetric data with skewness zero



Again,

$$\text{mode } = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) c.$$

Here

$$L = 18,\ c = 2,\ f_0 = 45,\ f_1 = 50,\ f_2 = 35.$$

Thus,

$$\text{mode } = 18.5.$$

Hence,

$$\text{arithmetic mean} = \text{median} = \text{mode}.$$

Note that the areas of the histogram in Figure 5.1 to the left and to the right of the line at the median are equal.

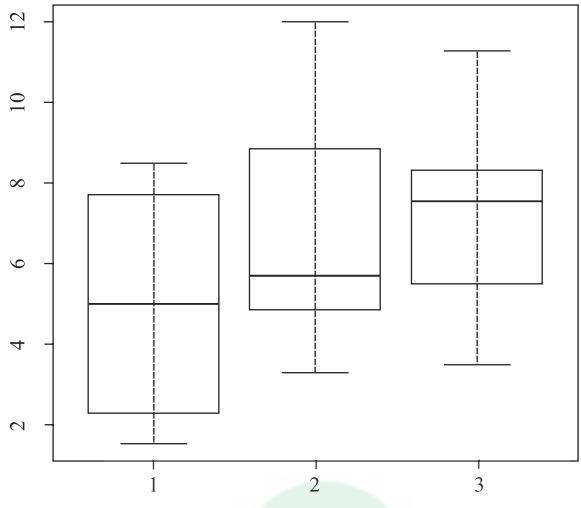
For symmetric data, the skewness is zero but the converse need not be true. From the above two examples it is evident that for asymmetric data also a measure of skewness can be zero.

5.7 BOX PLOT

Box plot is a data plot comprising tails and a box from the lower quartile to the upper quartile separated in the middle by the median for detecting data spread and patterning together with the presence of outliers (Outlier is a recorded value which differs markedly from the majority of the values collected). Line or whiskers, are used to connect the box edges to the minimum and maximum values in the data set.

Box plots illustrate the location of median, spread and skewness of a data and are useful for identifying outliers. The plot corresponds to a box, based on the lower quartile (Q_1) and upper quartile (Q_3), where the vertical crossbar inside the box marks the

Figure 5.2: Box plot of three data sets



position of the median. The median provides us with measure of central tendency. To measure dispersion, we can calculate the interquartile range from the graph, since the length of the box equals the interquartile range. Information concerning the skewness of the data can be gathered by seeing whether the median is closer to the lower or upper quartile. If the median is equidistant from the quartiles then the data is symmetric. If the median lies close to Q_3 then it is positively skewed and if median shifted towards Q_1 it is negatively skewed.

The first step in drawing a box plot is to create an appropriate scale along a horizontal line. Next, we draw a box that starts at Q_1 and ends at Q_3. Inside the box we place a vertical line to represent the median. Finally we extend lines from the box out to the minimum value and the maximum value. These lines outside of the box are sometimes called "whiskers" because they look a bit like a cat's whiskers. Box plots can also be drawn in a vertical manner as in the following example.

Box plots are particularly useful for comparing several groups of observations. A box plot is constructed for each group and these are displayed on a common scale. At least 10 observations per group are required in order for the plot to be meaningful.

Example 5.7.1. The box plots of the following three different data sets are shown in Figure 5.2.

Set 1: 4.6, 2.1, 1.5, 2.3, 5, 7.7, 7.9, 5.4, 8.5, 4.5.
Set 2: 5.7, 4.8, 3.8, 5.6, 4.9, 3.3, 6, 8.5, 9.3, 10.6, 12.
Set 3: 3.5, 7.5, 4.9, 4.2, 8.1, 5.7, 5.3, 6.4, 6.9, 7.8, 8.5, 8.1, 9.7, 10.5, 11.3.

Set 1 corresponds to the box plot of a symmetric data where the median lies in the middle of the box and the tails are of equal length. Note that for a symmetric data with single mode has mean = median = mode.

Right skew data are represented in Set 2 where the median lies close to Q_1 with the plot exhibiting a short left tail and long right tail with mean greater than median.

Set 3 corresponds to left-skew data with median close to Q_3 and mean less than the median.

5.8 HOW DESCRIPTIVE STATISTICS CAN BE MISUSED?

The famous British prime minister Benjamin Disraeli once said: "There are three kinds of lies- lies, damned lies, and statistics". The statement refers to the persuasive power of numbers, the use of statistics to bolster weak arguments, and the tendency of people to disparage statistics that do not support their positions. Sometimes deliberate distortions of data, but more commonly they involve either unintentional distortions or simply ineffective approaches to analysing numerical evidence. It is important for you to develop as much skill in rooting out statistical fallacies. Some organizations and people use statistics as a lamppost: for support, not for illumination. There are many misuses of descriptive statistics. The following are some errors that frequently occur in descriptive statistics.

Poor Presentation of Data

Sometimes graphs and charts are presented so as to give a misleading conclusion. Better presentations can reveal obscure facts in the data. Figures are sometimes presented in such a way that they seem comparable when in fact they are based on different concepts, time periods, areas, and so forth. Poor graphical and tabular presentations often lead both readers and writers to draw erroneous conclusions from their data.

Very Small Samples

Suppose you receive a report that a new drug for liver cancer will decrease the number of deaths due to liver cancer by 5 percent. But this figure of 5 percent might be based on the study of the effect of the drug on the lives of only 20 patients. If so, the figure of 5 percent may not be at all representative of the entire population of liver cancer patients (why?). Therefore it is generally wise to ask about the size of the sample on which a particular statistical result is based. Some people may select samples in such a way that the study may indicate what they want established, and the sample may not be at all a representative of the population. Clearly, statistics based on such improperly selected samples can be worse than valueless.

Improper Choice of Average

For a data concerning monthly income of residents belonging to a certain area, a real estate businessman may quote ₹25,000. His statement is perfectly true, but it is also misleading because the income data happens to be highly skewed to the right. Perhaps 2 percent of the residents earn about ₹50,000 per month, and the rest earn between ₹ 5000 and ₹15,000 per month. Thus the mean of ₹ 25,000 is not a very representative value. On the other hand, if somebody wants to impress the state government with the low income level of the area he may choose the mode as a measure of central tendency. Then the mode of the income data may be only ₹5000. Both are right; the

difference is due to the fact that different measures of central tendency are being used. Individuals and organizations sometimes choose the type of average that supports their case best, even if this information is misleading. The moral, of course, is clear: Whenever someone quotes an "average" be sure to find out what kind of measure it is and how representative it is likely to be.

Inappropriate Comparisons
Suppose you read in the newspaper that the number of deaths due to aeroplane crashes in 2008 is 20 times that in 1978. Before accepting this conclusion, it would be wise to question whether the number of air passengers and the number of flights in 1978 and 2008 are same or not. Also, do the figures include the deaths occurring due to terrorist attacks? Do the figures include the deaths occurred due to air force attack during a war?

Neglect of the Variation
Even if the proper kind of average is chosen, one has to take into account the variation of the individual values about the average. Be aware of these kinds of improper statistical procedures.

The web site *http://www.econoclass.com/misleadingstats.html* gives several examples of misleading statistics, and then an explanation of why it is misleading, and if possible suggests how to better present the data. The web sites
http://lilt.ilstu.edu/gmklass/pos138/datadisplay/sections/ goodcharts.htm and
http://lilt.ilstu.edu/gmklass/pos138/ datadisplay/badchart.htm
show how to create meaningful and readable graphs. Both sites give several examples of poorly constructed graphs.

Remark 5.8.1. Good data analysis entails little more than finding the best data relevant to a given study, making meaningful comparisons among the data, and drawing sound conclusions from the comparisons. To evaluate arguments based on numerical evidence, one must assess the reliability and validity of the individual measures used and validity of conclusions drawn from comparisons of the data.

Statistics in Excel
The site *http://home.ubalt.edu/ntsbarsh/excel/excel.htm* covers basic instructions for descriptive statistics and other analysis tools. It is a useful reference for students interested in exploring the capabilities of Excel for themselves.

5.9 EXERCISES

1. Sum of square of the deviations about arithmetic mean is:
   (a) maximum; (b) minimum; (c) zero; (d) none of these.

2. The standard deviation of a set of 10 observations is 3. If 4 is added to each observation in the data, the standard deviation of the new set of observations is:
   (a) 7; (b) 12; (c) $\sqrt{34}$; (d) 3.

3. If variance of a set of data is zero and its mean is 5. What is its mode?
   (a) 5; (b) 0; (c) 25; (d) 10.

4. If a constant value 5 is subtracted from each observation of a set, the variance is:
   (a) reduced by 5; (b) reduced by 25; (c) unaltered; (d) none of these.

5. The standard deviation of marks of 50 students is 12. Every student is later awarded 3 more marks. The standard deviation of the new set of marks is:
   (a) 12+$\sqrt{3}$; (b) 12; (c) 15; (d) none of these.

6. The arithmetic mean of two nonnegative integers is 5. Their standard deviation will be:
   (a) less than 5; (b) at most 5;
   (c) greater than five; (d) nothing can be said based on the given information.

7. If the first 25 percent observations of a data set is 20 or less and the last 25 percent observations of the data set is 50 or more, the quartile deviation (semi inter-quartile is) is :
   (a) 25; (b) 35; (c) 15; (d) 30.

8. If each value of a set of observations is divided by 5, its coefficient of variation is reduced by:
   (a) 0%; (b) 5%; (c) 10%; (d) 20%.

9. If a constant value 10 is subtracted from each value of a set of observations, the coefficient of variation will be:
   (a) decreased in comparison to original value; (b) same as original value;
   (c) increased in comparison to original value; (d) none of these.

10. For a positively skewed data, the inequality that holds is:
    (a) $Q_1 + Q_3 > 2Q_2$; (b) $Q_1 + Q_2 > 2Q_3$;
    (c) $Q_1 + Q_3 > Q_2$; (d) $Q_3 - Q_1 > Q_2$.

11. For a positively skewed data, which of the following holds?
    (a) median $>$ mode; (b) mode $>$ mean;
    (c) mean $>$ median; (d) mean $>$ mode.

12. The standard deviation of a set of values will be:
    (a) positive when the values are positive; (b) always positive;
    (c) positive when the values are negative; (d) all of these.

13. Which measure of dispersion has a different unit other than the unit of measurement of the variable:
    (a) range; (b) mean deviation; (c) standard deviation; (d) variance.

14. Which measure of dispersion is most affected by extreme values?
    (a) range; (b) mean deviation; (c) standard deviation; (d) quartile deviation.

15. The arithmetic mean of five numbers is 10. If each number is squared, then the mean of the squares of the numbers is:
    (a) 100; (b) greater than 100; (c) less than 100; (d) any of these.

16. If the standard deviation of a set of numbers is 4 and their arithmetic mean is 15, then the arithmetic mean of the squares of the numbers is:
    (a) 225; (b) 229; (c) 221; (d) 241.

17. If the constant value 5 is subtracted from each value of a data, then the coefficient of variation will be:
    (a) decreased in comparison to original value; (b) same as original value;
    (c) increased in comparison to original value; (d) none of these.

18. What does the length of the box in a box plot represent?
    (a) the range; (b) the interquartile range;
    (c) the median; (d) the mean.

19. Random digit dialing was used to select households in a particular state. An adult in each household contacted was asked whether the household had adequate health insurance. A critic of the poll said that the results were biased because households without telephones were not included in the survey. As a consequence, the estimated percentage of households that had adequate health insurance was biased upward. What type of bias was the critic concerned about?
    (a) measurement bias; (b) non-response bias;
    (c) response bias; (d) selection bias.

20. Let s_1 and s_2 denote the standard deviations of the two sets of data given below respectively,

Set 1	1	4	8	3	7	12
Set 2	1.1	4.4	8.8	3.3	7.7	13.2

    then
    (a) s_2 is 10% more than s_1; (b) s_2 is 11% more than s_1;
    (c) s_2 is 21% more than s_1; (d) s_2 is equal to s_1.

21. Let M_1 and M_2 be the means and s_1 and s_2 be the standard deviations respectively of the two sets of data given below,

Set 1	10	15	20	25	30
Set 2	6	11	16	21	24

    then
    (a) $M_1 > M_2, s_1 > s_2$; (b) $M_1 > M_2, s_1 < s_2$;
    (c) $M_1 < M_2, s_1 < s_2$; (d) $M_1 > M_2, s_1 = s_2$.

22. Why do we need a measure of variation?

23. Think of an example of a data where the range would be a misleading measure of variation.

24. A typist typed ten pages. The time spent in minutes for typing these pages are 5, 6, 5, 7, 6, 5, 4, 7, 6, 8. Find the mean deviation.

25. The following is the data on the waiting time that Manju spent at a bus stop for the morning bus to go to her office. Find the standard deviation of the waiting time.

Waiting time	:	0-5	5-10	10-15	15-20	20-25
Number of days	:	7	15	9	6	3

26. For the set of numbers 10, 7, 6, 12, 5, 10, 8, 11, 9 and 14, find
(a) $\sum_{i=1}^{5} |x_i -$ median$|$; (b) $\sum_{i=1}^{5} |x_i - \bar{x}|$; (c) the range; (d) the mean deviation about 10.

27. The times (in minutes) that eight customers spent in a supermarket are 6, 3, 1, 10, 2, 5, 4, and 6. Find

(a) the arithmetic mean $\bar{x}$ of the time a customer spent in the supermarket.

(b) standard deviation.

28. What is the variance of a set of scores when all the observations are identical?

29. After grading a test, Professor Jose announced to the class that the mean score was 58 points with a standard deviation of 14 points. However, Joe, one of the students, brought to his attention that one of the questions in the test was incorrect and impossible to solve. As a result Professor Jose agreed to add 10 points to the score of each student. Find the mean, the variance and the standard deviation of the new set of scores.

30. Suppose that every value in a data set is multiplied by a nonzero constant a. Exactly how does this affect the mean, variance and standard deviation, when compared with the original data set? Consider the cases $a > 0$ and $a < 0$ separately. In particular, what happens when a = 1?

31. For an irrigation cum drinking water supply dam at a certain place the normal water level, for proper irrigation and proper supply of drinking water, is 10 meters on 1st January. Taking normal level as zero, the water levels (in meters) for the last 12 years are 8.5, 7.8, 11.3, 10.5, 9.3, 13.2, 7.3, 8.9, 10.7, 12.6, 9.6 and 12.1. Compute average deviation of the water level.

32. Merly has recorded the price (in rupees) of a certain item at two different supermarkets over a period of twenty weeks. She has compiled the following information regarding the average prices and the standard deviations:

	Mean price	Standard deviation
Supermarket A	28.75	6.54
Supermarket B	27.91	2.70

If Merly is a bargain hunter, which store should she keep her eyes on?

33. The following are the scores made by two batsmen Alan and Milan in a series of innings:

Alan	:	25	40	90	83	100	12	6	45	68	26
Milan	:	65	81	5	27	112	36	45	20	15	84

Who is the more consistent player?

34. The following table gives the cumulative frequency of the marks of a random sample of 150 students in the Statistics examination:

Marks obtained									
equal to or less than	:	15	20	25	30	35	40	45	50
Number of students	:	11	18	46	85	120	135	144	150

(a) Find the range of marks obtained by the middle 50% of the students.

(b) If only 75% of the students are to be allowed to pass, taking the lower 25% as failures, what should be the minimum marks for passing?

(c) Compute quartile deviation.

35. In a nutrition experiment it was found that the coefficient of variation of gains in weight was 10%. In another experiment, the mean gain of weight was 5 kg and the standard deviation was 1kg. Does the second experiment agree with the first?

36. Seven buses are expected to arrive at a certain bus stand at 6 a.m., 7 a.m., 8 a.m., 9 a.m., 10 a.m., 11 a.m., 12 noon. They arrive at 6:04 a.m., 7:10 a.m., 8:02 a.m., 8:56 a.m., 10:05 a.m., 11:10 a.m., 11:55 a.m. respectively. Find the mean absolute deviation of the amount of time by which a bus is late. (If a bus arrives ahead of schedule, say, by 4 minutes, then take the late arrival time negative as -4.)

37. Ten pieces of rubber sheet, each of which weighs 1 kilograms after drying in a smoke house, loses the following amounts of moisture (in grams).

$$158, 160, 156, 153, 163, 148, 156, 163, 152, 164.$$

Find the mean absolute deviation of the amount of moisture lost.

38. The price of pepper (in rupees) at a certain market during six weeks are 298, 309, 303, 292, 288, 284, 292 and 295. Find the number of weeks when the price is within one standard deviation on either side of the mean price.

39. A shopkeeper sells coffee powder in packets, each containing 200 grams. The amounts of coffee powder in grams by which 50 packets differs from the nominal 200 grams were as follows (If a packet weighs less, say, by 4 grams, then the measurement is taken to be negative as -4.):
3,2,-1,5,-4,3,4,-4,0,1,-3,7,3,2,5,-4,0,1,-2,-1,3,5,-6,0,-5,2,3,3,-5,-2,-1,-2,3,4,1,4,2,-4,0,-1,
-6,0,0,1,2,-1,-3,-2,0,1.
Calculate the mean and the standard deviation of the amount by which the quantity of coffee powder in the packet differs from its nominal weight. Comment on the result.

40. The following are the diastolic pressure of 100 persons.
100, 85, 65, 160, 155, 120, 130, 85, 80, 65, 70, 90, 65, 90, 80, 130, 135, 128, 145, 158, 165, 80, 82, 66, 72, 86, 83, 126, 135, 138, 76, 84, 86, 127, 160, 80, 88, 90, 100, 95, 138, 146, 124, 100, 86, 80, 86, 136, 168, 165, 69, 85, 126, 110,

115, 132, 84, 70, 68, 145, 90, 96, 118, 88, 160, 155, 150, 80, 86, 94, 100, 70, 68, 116, 140, 139, 110, 85, 80, 89, 85, 76, 115, 130, 145, 150, 86, 80, 96, 78, 76, 120, 110, 82, 83, 94, 100, 74, 79, 135.

(a) Compute the mean and standard deviation of diastolic pressure.

(b) Construct a frequency table with a suitable number of classes and compute the mean and the standard deviation for the grouped data.

(c) Compare the values obtained in (a) and (b).

(d) Draw a box plot and comment on the skewness of the data.

(e) Compute Bowley's coefficient of skewness.

41. By the end of the year, Milan forgets the scores he received on the four quizzes (each worth 100 points) he took in the school. He only remembers that their average score was 80 points, standard deviation 10 points, and that 3 out of the 4 scores were the same. From this information, compute all four missing quiz scores. There are two possible solutions to this problem. Find them both.

42. A data set entirely consists of x ones and $n-x$ zeros. If p denotes the sample proportion of ones, obtain the sample mean and the sample variance in terms of p.

43. Calculate the mean deviation from the mean and the standard deviation of the set of numbers

$$a, a+d, a+2d, \ldots, a+2md.$$

Prove further that the latter is greater than the former.

44. The sum and sum of squares corresponding to length (x) in cms and weight (y) in kilograms of 10 tapioca tubers are given below:

$$\Sigma x = 420, \quad \Sigma x^2 = 4240, \quad \Sigma y = 4.3, \quad \Sigma y^2 = 8.1.$$

Which is more varying, the length or weight?

45. Measure the heights of a sample of students from your class and calculate the sample mean and standard deviation of the values. Now form the new set of "standardized" values by subtracting the sample mean from each value and dividing the resulting number with the standard deviation. Calculate the sample mean and standard deviation of the new set of values. Repeat the same several times by taking a different sample each time. Briefly explain what is happening, and why?.

46. The mean square deviation of a set of values $x_1, \ldots, x_n$ about a point a is defined to be

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2.$$

The mean square deviations about -1 and 1 of a set of observations are 7 and 3 respectively. Find the variance of the set of observations.

47. The following data give the arithmetic mean and standard deviation of three subgroups. Calculate the arithmetic mean and standard deviation of the whole group.

Subgroup	Number of labourers	Mean wage	Standard deviation
A	50	160	10
B	100	200	15
C	120	350	50

48. In an investigation of the risk factors for cardiovascular disease, levels of serum cotinine a metabolic product of nicotine were recorded for a group of smokers and a group of nonsmokers. The relevant frequency table is given below. (American Journal of Public Health, Volume 82, January 1992, 33-36.)

Cotinine Level (ng/ml)	Smokers	Nonsmokers
(0, 13)	78	3300
(13, 50)	133	72
(50, 100)	142	23
(100, 150)	206	15
(150, 200)	197	7
(200, 250)	220	8
(250, 300)	151	9
300 and above	412	11
Total	1539	3445

Answer each of the following:

(a) Is it fair to compare the cotinine levels for smokers and nonsmokers based on the absolute frequencies in each interval? Why or why not?

(b) For the smokers, compute a table of relative frequencies and a table of cumulative relative frequencies.

(c) What percentage of smokers have cotinine levels between 125 and 210 ng/ml?

(d) For the smokers, what level of cotinine (ng/ml) represents the upper one-fifth percentile?

(e) For all individuals in this study, smoking status is self-reported. Do you think any of the subjects might be misclassified? Why or why not?

(f) Compute the group mean serum cotinine level measurements for both groups, smokers and nonsmokers. (For the last interval take the midpoint of the interval to be 340 ng/ml.) In which interval does the median of each group lie? Compare the two groups.

49. Find the missing information from the following data:

	Group I	Group II	Group III	Combined
Sample size	-	50	80	200
Arithmetic mean	70.5	-	63.8	67.2
Standard deviation	4.8	6.3	-	5.2

50. A magazine once published a report saying that farmers lead other groups in the consumption of alcohol. As evidence, it pointed to the fact that a rehabilitation center in Kerala treated more farmers than other occupational groups for alcoholism. What sort of pitfall is present here?

Chapter 6

INTRODUCTION TO PROBABILITY THEORY

"Statistics is a body of methods for making wise decisions in the face of uncertainty"- W.A Wallis and H.V.Roberts.

6.1 INTRODUCTION

In our daily life we hear statements such as: 'Rain is likely today', 'The next child of the lady is probably a girl' and so on. Unpredictability lies in the above statements and such unpredictable behaviour is usually described as 'random'. Our present task is to learn how to deal with randomness and how to quantify it. This is necessary for making wise decisions in the face of uncertainty. The theory of probability was developed for this purpose and its domain of applications now extends over almost all disciplines. The theory of errors, actuarial mathematics and statistical mechanics are examples of some of the important applications of probability theory developed in the 19th century.

Fundamental principles of probability theory were formulated in the middle of 17th century in France. It is hard to believe, a gambler's dispute led to the origin of the mathematical theory of probability. In 1654, Chevalier de Méré (1607-1648), a French nobleman with an interest in gambling posed some questions to Blaise Pascal(1623-1662) who was one of the great thinkers of that time. The questions were the following:

1. How many throws of two dice are required for the occurrence of at least one "double 6" ?

2. How to share the wagered money between two gamblers if the game interrupted untimely?

These questions led to an exchange of letters between two great scientists Pascal and Pierre de Fermat (1601-1665) in which the fundamental principles of probability theory were formulated for the first time. The subject developed rapidly during the 18th century. The major contributors during this period were Jacob Bernoulli (1654-1705) and Abraham de Moivre (1667-1754). Pierre Simon (Marquis de) Laplace (1749-1827) applied

probabilistic ideas to many scientific problems. Many scientists have contributed to the theory since Laplace's time; among the most important are Pafnuty Lvovich Chebyshev (1821-1894), Andrei Andreyevich Markov (1856-1922), Richard von Mises (1883-1953) and Andrey Nikolaevich Kolmogorov (1903-1987).

The search for widely acceptable definition took nearly three centuries and was marked by much controversy. The matter was finally settled in the 20th century when the Russian mathematician A.N. Kolmogorov introduced axiomatic definition of probability in 1933.

This chapter is intended to describe basic concepts and results of probability theory. A number of examples and exercises are included.

Elementary Set Theory

Before introducing the basic terms of probability theory, we have to be familiar with the basic concepts of set theory which we now briefly review.

By a *set* we mean a well defined collection of objects called the elements of the set. Some familiar sets to which we will refer repeatedly include:

(i) The collection of voters in Kerala.

(ii) The set of natural numbers, $N = \{1, 2, 3, \ldots\}$.

(iii) The set of integers, $Z = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$.

(iv) The set of real numbers, $\Re = (-\infty, \infty)$.

The totality of all elements which might conceivably belong to any of the sets under discussion is called *universal set*, denoted by U. In the first example, the set of all people in Kerala constitutes the universal set. Sets are denoted by using uppercase letters and the elements are represented using lowercase letters. A set A is called a subset of a set B, written $A \subset B$, if every element of A is also an element of B. For example,

$$N \subset Z \subset \Re.$$

The empty set $\emptyset$, the set which contains no elements, is a subset of every set.

Two sets can be placed in *one-to-one correspondence* if their elements can be paired in such a way that each element of either set is associated with a unique element of the other set. A set whose elements can be arranged in a one-to-one correspondence with a subset $\{1, 2, 3, \ldots, n\}$ of positive integers is called a *finite set*. All other sets are called *infinite sets*. Any infinite set whose elements can be arranged in a one-to-one correspondence with the set $\{1, 2, 3, \ldots\}$ of natural numbers is said to be *denumerable* or *countably infinite*. All other infinite sets are said to be *uncountably infinite*. The set of real numbers and the set of irrational numbers are uncountable sets. A set is countable if it is either finite or denumerable. The set of even natural numbers and the set of rational numbers are denumerable sets and the set of of voters in Kerala is a finite set.

Given a set A we can form another set, called the *complement* of A, denoted as A' or $\overline{A}$ or A^c, which contains precisely those elements which are not in A but in the universal set U. That is, A^c will occur if and only if A does not occur.

Given two sets A and B, the *union* $A \cup B$ and the *intersection* $A \cap B$ are defined as follows. $A \cup B$ contains those elements which are either in A or B or in both. $A \cap B$ contains those elements which are in both A and B. If $A \cap B = \emptyset$, then A and B are disjoint. $A \cap B^c$ consists of all elements in A but not in B. The operations of union and intersection both satisfy commutative and associative laws, and they are related by the distributive laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

and

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

The definitions of union and intersection can be extended to a collection of sets, whether that collection be countable or uncountable. To belong to the union of a collection of sets, an element must belong to at least one of the sets. To belong to the intersection, an element must belong to every one of the sets. The following laws called De Morgan's laws relate complement to union and intersection.

$$(A \cap B)^c = A^c \cup B^c$$

and

$$(A \cup B)^c = A^c \cap B^c.$$

De Morgan's laws extend to the case of countable collection of sets as follows:

$$\left(\bigcup_{k=1}^{\infty} A_k \right)^c = \bigcap_{k=1}^{\infty} A_k^c$$

and

$$\left(\bigcap_{k=1}^{\infty} A_k \right)^c = \bigcup_{k=1}^{\infty} A_k^c.$$

Thus the complement of a union is always the intersection of the complements of sets, and the complement of an intersection is always the union of the complements of sets. The first property states that an object is not in any of the sets in the collection if and only if it is in the complement of each set; the second property states that an object is not in every set in the collection if it is in the complement of at least one set.

A *symmetrical difference* between two sets, denoted as $A \triangle B$, is defined as

$$A \triangle B = (A \cap \bar{B}) \cup (B \cap \bar{A}).$$

Sets and their complements, unions and intersections, can be visualized by means of Venn diagrams. John Venn (1834-1923) was a mathematician and priest who lived and died in England. In Venn diagrams a set is represented by the interior of any closed curve which is contained within any other closed curve that represents the universal set. The complement of the set is then represented by the exterior of the inner curve within the outer curve. Venn diagrams are shown in Figure 6.1.

6.2 RANDOM EXPERIMENT

When a scientist announces a discovery, other scientists in different parts of the world will be able to verify his/her findings for themselves. If the results of two scientists appear to disagree it means that the experimental conditions were not quite the same in the two cases. If the same results are obtained when an experiment is repeated under the same conditions, we say that the experiment is repeatable. It is the repeatable nature of science that permits the use of scientific theory for predicting what will be observed under specified conditions. However, there are also experiments whose results vary, in spite of all efforts to keep the experimental conditions constant. Familiar examples are provided by gambling games: throwing dice, tossing coins etc. can all be thought of as "experiments" with unpredictable results. More important and interesting instances occur in many contexts. For example, seeds that are apparently identical will produce plants of differing height, electric bulbs that are identically manufactured and used differ in their life lengths etc. We shall refer to experiments where the outcomes are not predetermined or deterministic, and thus do not always yield the same result when repeated under the same conditions, as *random experiments*. Probability theory and Statistics are the branches of Mathematics that have been developed to deal with random phenomena.

In the case of a *deterministic experiment* we can determine the outcome of the experiment in advance based on the outcomes of the past repetitions of the experiment. For example, if the experiment is to see whether a silver coin will sink in water then the outcome is pre-determined. The coin will sink in water. It is a deterministic experiment. The experiments conducted in science laboratories are deterministic in nature as the outcomes of these experiments will be the same if they are repeated under identical conditions.

Now consider an experiment of throwing a coin. Then 'head' or 'tail' will turn up and the outcome in a particular trial is not pre-determined even if we have repeated the experiment several times in identical conditions. We may make every effort to homogenize the experimental conditions, by always placing the coin in the same position, always applying the same force to throw, always throwing it against the same spot on the floor, and so on. In spite of all such efforts, the outcome of this experiment varies irregularly from repetition to repetition and hence it is a random experiment.

A *random experiment* (trial or run) is an experiment having more than one possible outcomes and can be repeated (reproduced) an arbitrary number of times (at least theoretically) under an invariable set of conditions. In each repetition of the experiment, one and only one of the possible outcomes takes place. The object of observation in a random experiment may be a certain process or a physical phenomenon or an operational system. Note that an experiment is a process- natural or set up deliberately- that has an observable outcome. In the deliberate setting, the word experiment and trial are synonymous. The following are some examples of random experiment.

1. Counting the number of deaths that occurred in Kottayam municipality on a particular day.

2. Observing the number of people in Kerala infected by swine flu on a particular day.
3. Noting the number of artificial insemination done in cows in a particular locality on a particular day.
4. Noting the weights of fishes caught by a fisherman on a certain day.
5. Noting the life length of an electric bulb.
6. Noting the quantity of gold sold in a day in a particular town.
7. Noting the consumption of electricity in your college on a certain day.

Sample Space

The set of all possible outcomes that may occur as a result of a particular random experiment, usually denoted by S, is called the 'sample space' of the random experiment. For example, the sample space for the coin tossing experiment is

$$S = \{\text{head, tail}\} \text{ or } \{H, T\} \text{ or } \{0, 1\}, \ldots.$$

One can represent S graphically also. The elements of S are called *sample points* or *elementary outcomes* or *elementary events* and they may be real numbers, alphabets, living things, objects etc. It is desirable to use a sample space whose elements cannot be "subdivided" into more elementary kinds of outcomes; that is, an element of a sample space should not represent two or more outcomes which are distinguishable in some fashion. As in the coin tossing experiment, there may be several natural choices for S; select the simplest one that contains enough information to describe the outcomes of interest. A sample space may be discrete or continuous.

Discrete Sample Space

A sample space is called *discrete* if the outcomes are countable or enumerable which means that the outcomes can be arranged in one-to-one correspondence with the set of natural numbers. That is the outcomes in a discrete sample space can be numbered $1, 2, \ldots$. Obviously, a discrete sample space may contain finite or countably infinite number of sample points.

Example 6.2.1. If I plant ten mango seeds and count the number that germinate, then the sample space is $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

Example 6.2.2. Suppose that the random experiment is to count the number of repairs of a particular machine. Then the sample space is $S = \{0, 1, 2, \ldots\}$.

Continuous Sample Space

If the sample space has continuum of points, which are uncountable, then it is called a *continuous* sample space. Continuous sample spaces arise in practice whenever the outcomes of experiment are measurements of physical properties such as temperature, speed, pressure, length, volume etc. that are measured on continuous scales.

Example 6.2.3. The random experiment is to measure the duration between one repair and the next breakdown of a particular equipment. Then the sample space S is the set of all positive real numbers.

Events

A specified collection of outcomes in the sample space S or any subset of S is called an *event*. Upper case letters such as A, B, C etc are used to denote events. Obviously, an event is a subset of a sample space. The event coinciding with the empty set $\emptyset$ is called an *impossible event*, and the event coinciding with the entire set S is called a *certain* or *sure event*.

Example 6.2.4. Consider a random experiment of rolling a die and observing the number of points on the upward face of the die. Then the sample space is $S = \{1,2,3,4,5,6\}$. Define :

A: the event that the outcome is even,
B: the event that the outcome is odd,
C: the event that the outcome is prime.

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

Sample points comprising of an event are said to be *favourable* to that event. In the above example, if the face 5 has turned up, then we say that B has occurred and there are 3 outcomes favourable to each of the events A, B and C.

6.2.1 Algebra of Events

The universal set in set theory can be identified with the sample space and the concept of subsets can be identified with events. The usual operations on sets - complement, union and intersection - have simple interpretations in terms of occurrence of events. As in set theory we often use Venn diagrams (see Figure 6.1) to represent sample space and events.

If A is an event, then the *complement* of A with respect to S (that is, those elements in S which are not in A) can be interpreted as non-occurrence of the event A, and vice versa. Obviously, when A is an event then its complement, denoted by $\bar{A}$ or A^c or A', is also an event. It is depicted in Figure 6.1(a).

Note that

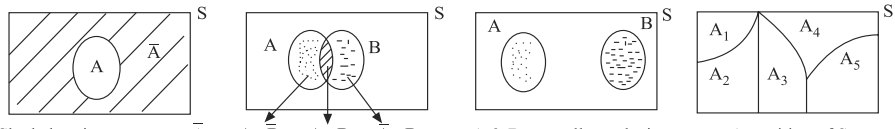
$$A \cap A^c = \emptyset \text{ and } A \cup A^c = S.$$

If A and B are two events, then the *union* of A and B, denoted as $A \cup B$, is an event which occurs when either A or B or both occurs. Shaded regions in Figure 6.1(b)-(c) represent union of two events A and B. Similarly the *intersection* of A and B, denoted as $A \cap B$, is an event which occurs when both A and B occur. This is shown in Figure 6.1(b).

Two events A and B are said to be *disjoint* or *mutually exclusive* if the random experiment cannot result in their joint occurrence. This is shown in Figure 6.1(c). Note that $A \cap \bar{B}$ and $\bar{A} \cap B$, shown in Figure 6.1(b), are disjoint. Events $\{A_1, A_2, \ldots\}$ are said to be disjoint or mutually exclusive if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

A set of events $\{A_1, A_2, \ldots\}$ is said to be '*totally exhaustive*' or to be '*complete*' if the union of $A_1, A_2, \ldots$ is S. In Example 6.2.4 the events A and B are mutually exclusive and totally exhaustive. Obviously, the individual outcomes or elementary events in a sample space S constitute a set of mutually exclusive and totally exhaustive events. A

Figure 6.1: Compound events



set of events $\{A_1, A_2, \ldots\}$ said to constitute a *partition of* S if $A_1, A_2, \ldots$ are disjoint and their union is S. In other words, no matter what the outcome of the experiment, one and only one of $A_1, A_2, \ldots$ occurs. This is shown in Figure 6.1(d). Events which are combinations of two or more events formed by taking unions, intersections and complements are called *compound* events. Therefore, it is necessary that the definition of events should be such that the above operations on events can be carried out. More precisely, the set of all events, denoted by $\mathcal{A}$, must satisfy the following three axioms.

Axiom 1: The sample space S is an event.
Axiom 2: If A is an event, then A^c is also an event in S.
Axiom 3: If $\{A_1, A_2, \ldots\}$ is any countable set of events in S, then the union $A = \bigcup_{i=1}^{\infty} A_i$ is also an event in S.

From the first two axioms it is obvious that $\emptyset = S^c$ is also an event. From the last two axioms it follows that $\bigcup_{i=1}^{\infty} A_i^c = \left(\bigcap_{i=1}^{\infty} A_i\right)^c$ is an event so that $\bigcap_{i=1}^{\infty} A_i$ is also an event. Thus $\mathcal{A}$ is a set of sets which is closed under countable unions (intersections) and complements. In set theory, a family which satisfies the above three axioms is called a σ-*algebra* (or a σ-*field* or a *Borel field*). Therefore, by an observable event we mean a subset of S such that it is a member of the σ-field of events, denoted by $\mathcal{A}$.

To illustrate these ideas consider Example 6.2.4. Take $\mathcal{A}$ as the set of all subsets of S. The events A and B are disjoint but B and C are joint. Since $A \cup B = S$ and $A \cap B = \emptyset$, A and B together form a partition of S.

The class of events observable in a given random experiment is, in general, smaller than the set of all subsets of S (If there are n elements in S, then the set of all subsets of S will contain 2^n elements). Obviously $\mathcal{A} = \{\emptyset, S\}$ is a σ-field and it is called the trivial σ-field. For random experiments with a finite S it is quite usual to take $\mathcal{A}$ to be the set of all subsets of S. For such random experiments any subset of S may be interpreted as an observable event. When S is infinite(countable or uncountable) any subset of S need not be an observable event as $\mathcal{A}$ may contain only some smaller collection of subsets of S.

Remark 6.2.1. The definition of an event with respect to a σ-algebra agrees with the notion of an event introduced earlier as an observed result of a random experiment. It is obvious that if any two events are observable in a given random experiment, then according to the meaning of algebra of events (sets) their union, intersection and com-

plement must also be observable in the random experiment. In other words, the field of events must be closed under algebraic operations on events. These conditions are precisely satisfied by the σ-field of events.

Example 6.2.5. Consider Example 6.2.4. It is easy to check that $\mathcal{A}_1 = \{\emptyset, A, B, S\}$ is σ-algebra but $\mathcal{A}_2 = \{\emptyset, A, B, C, S\}$ is not. $\mathcal{A}_2$ is not a σ-field because $\bar{C} = \{1,4,6\}$, $B \cup C = \{1,2,3,5\}$, $A \cup C = \{2,3,4,5,6\}$ etc. do not belong to $\mathcal{A}_2$.

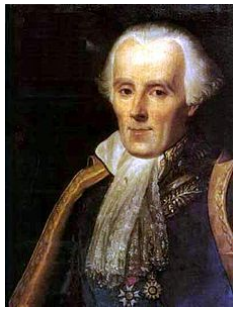
6.3 PROBABILITY

Given a random experiment and a sample space of possible outcomes of the experiment, it is natural to ask "what is the chance of a particular event to occur?" It is here we are looking for a numerical value that will be a measure of the chance of occurrence of the event. Probability of an event A, denoted by $P(A)$, is a number in the closed interval [0,1] which is assigned on the basis of the relative likelihood of the occurrence of that event.

There are three different approaches to probability-classical, empirical and axiomatic approaches. Let us study these three approaches and practise how to assign a measure of certainty to random events.

6.3.1 Mathematical or Classical or 'a priori' Definition of Probability

The classical definition of probability is identified with the works of a French mathematician and an astronomer Pierre Simon Laplace (1749-1827). It is applicable only to those random experiments in which S is a finite set of symmetric elementary outcomes. Symmetry of outcomes is assumed with respect to the physical characteristics and other factors which may affect the outcomes. If there is symmetry in the outcomes of the random experiment then the outcomes are called *equally likely* because there is no reason to suspect that one outcome is more likely or less likely to occur than any other outcome.



Pierre Simon Laplace (1749-1827)

Don't assume that the outcomes are equally likely unless either you are told to, or there is some physical reason for assuming it. In the mango seeds example, it is most unlikely. In the coins example, the assumption will hold if the coin is 'fair': this means that there is no physical reason for it to favour one side over the other. The first edition of Laplace's book *Théorie Analytique des Probabilités* appeared in 1812.

Let us define an event E that may consists of a single sample point or a group of sample points in a sample space consisting of a countable number of symmetric sample points. Letting $n(A)$ be the number of sample points in an event A, we shall state the classical definition or a priori definition of probability.

Definition 6.3.1. If a random experiment has $n(S)$ symmetric elementary outcomes and $n(E)$ of which are favourable to an event E, then

$$P(E) = \frac{n(E)}{n(S)} = \frac{\text{Number of outcomes favourable to E}}{\text{Total number of outcomes in } S}.$$

Remark 6.3.1. To calculate the 'a priori' probability we do not actually perform the random experiment but it is necessary to know that the possible outcomes are symmetric or equally likely. The latin phrase a priori means 'without investigation or sensory experience'.

Limitations of 'a priori' Definition of Probability

(i) We should not use the 'a priori' definition if we do not know that the possible outcomes are symmetric.

(ii) The definition is valid when the sample space is finite.

(iii) Since both $n(E)$ and $n(S)$ are integers, $P(E)$ is reduced to a rational number. Irrational numbers are not taken care of by this definition.

6.3.2 Empirical or Statistical or Relative Frequency Definition of Probability

If the sample space of the random experiment is not symmetric then the classical definition of probability cannot be applied. In such circumstances the only way to calculate the probability is to empirically run the experiment, and count the relative frequency of the events of interest and use it as an approximation of the probability.

The principal founder and proponent of the frequency interpretation of probability is Richard von Mises (1883-1953). Consider a series of repetition of a particular random experiment and we are interested in a certain event E. If E occurs r times in n trials, then the ratio $\frac{r}{n}$ is called the *relative frequency* or the *frequency ratio* of the event E. If we observe the frequency ratio $\frac{r}{n}$ for various values of n it can be seen that the ratio shows a tendency to settle down to a constant value as the number of trials gets larger and larger. This tendency is sometimes called *statistical regularity*. Formally, the *empirical definition* of probability is as follows:

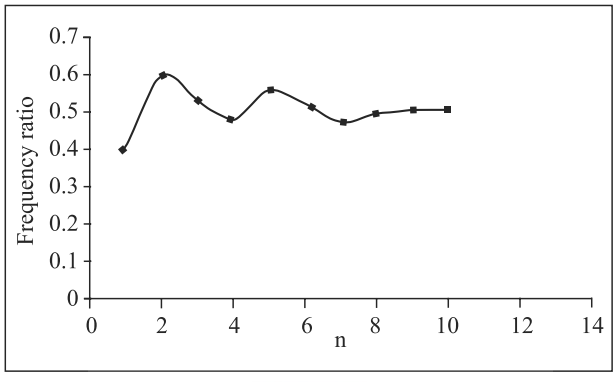
Richard von Mises (1883-1953)

Definition 6.3.2. If an event E occurs r times out of n independent trials of a random experiment then the limiting value of the ratio $\frac{r}{n}$ as $n \to \infty$ if it exists is called the probability of the event E.

For example, consider the tossing of an unbiased coin and let A be the event that a throw results in a head. The frequency ratio $\frac{r}{n}$ will fluctuate with small values of n

but as n increases there will be tendency for the ratio to approach 0.5. This is shown graphically in Figure 6.2.

Figure 6.2: Convergence of relative frequency



The word independent in the definition means that the outcome of any trial does not depend on the results of previous trials. Even if statistical definition overcomes most of the limitations of mathematical definition of probability it requires repetition of the random experiment a large number of times. Here probability only has meaning for events from experiments which could in principle be repeated arbitrarily many times under essentially identical conditions. Then, the probability of an event is simply the "long-run proportion" of times that the event occurs under many repetitions of the experiment. It is reasonable to suppose that this proportion will settle down to some limiting value eventually, which is the probability of the event.The larger the value of n, the better is the estimate of probability. However, we cannot assert that the limit of $\frac{r}{n}$ exists in a mathematical sense. This definition of probability is known as the *frequentist interpretation* of probability. Probabilities measured by the frequency interpretation are referred to as *objective probabilities*.

6.3.3 Axiomatic Definition of Probability

The Russian mathematician Andrey Nikolaevich Kolmogorov (1903-1987) presented a definition of probability in terms of three axioms. In 1933, Kolmogorov published his book *Foundations of the Theory of Probability* that laid the foundation of modern axiomatic theory of probability. A quotation attributed to Kolmogorov is: "Every mathematician believes that he is ahead over all others. The reason why they don't say this in public, is because they are intelligent people."

A.N. Kolmogorov (1903-1987)

Definition 6.3.3. For each event A with respect to the sample space S, there exists a real number, denoted by $P(A)$, that satisfies the following axioms:
Axiom 1: $0 \leq P(A) \leq 1$
Axiom 2: $P(S) = 1$
Axiom 3: For any sequence of mutually exclusive events $A_1, A_2, \ldots$ defined with respect to S

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

If there are only finitely many events $A_1, A_2, ..., A_n$ then Axiom 3 becomes

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \Sigma_{i=1}^{n} P(A_i).$$

The first axiom states that probability is a nonnegative number in $[0,1]$. Axiom 2 states that probability of a sure event(sample space) is unity. Axiom 3 states for any sequence of mutually exclusive events, the probability of at least one of the events occurring is the sum of the probabilities of individual events. Fundamental results in probability originated from this definition.

6.3.4 Theorems of Probability

Based on the three axioms of probability, we can derive many other rules. Among the immediate application of the axioms we prove the following theorems:

Theorem 6.3.1. *If A and A' are complementary events in a sample space S, then*

$$P(A') = 1 - P(A).$$

Proof. By the definition of a complement, A and A' are mutually exclusive and $A \cup A' = S$. By Axiom 2

$$\begin{aligned} P(S) &= 1 \\ P(A \cup A') &= 1 \\ P(A) + P(A') &= 1 \text{ by Axiom 3} \\ P(A') &= 1 - P(A). \end{aligned}$$

Theorem 6.3.2.

$$P(\emptyset) = 0 \textit{ for any sample space } S.$$

Proof. Since $S \cup \emptyset = S$ and S and $\emptyset$ are mutually exclusive we have

$$\begin{aligned} P(S) &= P(S \cup \emptyset) \\ &= P(S) + P(\emptyset) \text{ by Axiom 3} \end{aligned}$$

and hence $P(\emptyset) = 0$.

Remark 6.3.2. $P(A) = 0$ does not necessarily imply that A is an empty set. In practice, we often assign a probability of 0 to events which, in colloquial terms, would not happen in a million cases. For instance, we assign probability of 0 to the event that a coconut falls exactly on the head of a person who passes beneath a coconut tree.

Theorem 6.3.3. *If A and B are events in a sample space S and $A \subset B$, then $P(A) \leq P(B)$.*

Proof. Since $A \subset B$ we can write

$$B = A \cup (A' \cap B)$$

where A and $A' \cap B$ are mutually exclusive. Hence by Axiom 3, we have

$$\begin{aligned} P(B) &= P(A) + P(A' \cap B) \\ &\geq P(A). \end{aligned}$$

This theorem states that if the event A is a subset of the event B, then $P(A)$ is not greater than $P(B)$. For instance, in a random draw the probability of getting a heart from an ordinary deck of 52 playing cards is not greater than the probability of drawing a red card, namely, $\frac{1}{4}$ compared to $\frac{1}{2}$.

Theorem 6.3.4. *If A and B are any two events in a sample space S, then*

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

Proof. We can write $A \cup B$ as the union of two mutually exclusive events $A \cap B'$ and B. The shaded regions in Figure 6.1 (b)-(c) represent $A \cup B$.

$$A \cup B = (A \cap B') \cup B.$$

Note that if A and B are disjoint then $A \cap B' = A$. Since $(A \cap B')$ and B are mutually exclusive, by using Axiom 3 we can write

$$P(A \cup B) = P(A \cap B') + P(B). \tag{6.3.1}$$

Now A can be written as the union of two disjoint sets as

$$A = (A \cap B') \cup (A \cap B).$$

Thus, by Axiom 3, we have

$$\begin{aligned} P(A) &= P(A \cap B') + P(A \cap B) \quad \text{and hence} \\ P(A \cap B') &= P(A) - P(A \cap B). \end{aligned}$$

Now substituting for $P(A \cap B')$ in (6.3.1) we obtain

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

In particular, if A and B are mutually exclusive then $P(A \cap B) = P(\emptyset) = 0$. In this case

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

The above theorem is called the *addition theorem* of probability for two events. Addition theorem for the case of three events is as follows:

Theorem 6.3.5. *If A, B and C are any three events in a sample space S, then*

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

Proof. On writing $A \cup B \cup C$ as $A \cup (B \cup C)$ and using the previous theorem, we get

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

Now $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and hence

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

Hence it follows that

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

Further generalization of addition theorem to k events is now straightforward and the resulting formula can be proved in a similar manner.

For computing probabilities of events using classical definition one has to find the number of all possible outcomes of the random experiment and the number of outcomes favourable to the event under consideration. It in turn requires the use of counting techniques. Before we consider problems in probability let us consider some important principles of counting.

Some Counting Principles

Principle 1 (Multiplication principle)
If an operation (job) consists of k tasks, of which the first task can be done in n_1 ways, second can be done in n_2 ways and so on, then the number of ways in which the operation can be done is $n_1 \times n_2 \times \ldots \times n_k$.

Example 6.3.1. How many ways can a student answer a true-false test consisting of 10 questions?

Each question can be answered in 2 ways. Therefore there are

$$\underbrace{2 \times 2 \times \ldots \times 2}_{10 \text{ factors}} = 2^{10} = 1024$$

different ways in which one can answer the test, and only one of these corresponds to the case where each answer is correct.

Remark 6.3.3. Given n distinct symbols which may be repeated any number of times, the number of arrangements of symbols of length r is n^r.

Example 6.3.2. Suppose a population contains N numbered units and a sample of n units was drawn one at a time, with each unit being placed back into the population before the next unit is drawn. Sampling of this type is called *sampling with replacement*. Then the total number of possible samples is N^n.

Remark 6.3.4. Suppose we have k trials each having $n(S_1),\ldots,n(S_k)$ possible outcomes respectively, then the resulting sample space consists of $n(S_1)\times n(S_2)\times\ldots\times n(S_k)$ possible outcomes. In particular if the random experiment is to toss three unbiased identical dice together or tossing of an unbiased die thrice, then the number of possible outcomes is $6\times 6\times 6=6^3$.

Example 6.3.3. A coin and a die are tossed together. The coin can fall in two ways $\{H,T\}$ and the die in six ways $\{1,2,3,4,5,6\}$. Then there are $2\times 6=12$ possible pairs of outcomes and the sample space is the following:

$$\begin{aligned} S &= \{H,T\}\times\{1,2,3,4,5,6\} \\ &= \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}. \end{aligned}$$

Principle 2 (Permutations)
The number of arrangements (permutations) of n distinct things in a row is

$$n(n-1)\cdots 1=n!.$$

where $n!$ is read as "n factorial". The number of arrangements of n distinct things in a circle is $(n-1)!$.

Example 6.3.4. How many permutations are possible with the letters abc?

The possible arrangements are abc, acb, bac, bca, cab and cba. Hence the number of distinct permutations is six. We could have arrived at this answer without actually listing the different permutations. Since there are three choices of the first, two choices for the second and only one choice for the last position, the total number of arrangements is

$$3\times 2\times 1=3!.$$

Principle 3 (Permutation of elements with repetitions)
The number of distinct arrangements of n objects in which n_1 are alike, n_2 others are alike, $\ldots,n_k$ of k^{th} kind are alike and $n_1+\cdots+n_k=n$ is given by $\frac{n!}{n_1!n_2!\ldots n_k!}$.

Example 6.3.5. How many signals can be obtained using 2 red flags, one white flag and one green flag?

Denoting the four flags by R,R,W and G we find by enumeration the different signals as:

RRWG, RRGW, RWRG, RWGR, RGRW, RGWR,
GRRW, GRWR, GWRR WRRG, WRGR, WGRR.

Now applying the above formula, the answer is

$$\frac{4!}{2!1!1!}=\frac{24}{2}=12.$$

Principle 4 (Permutation of n distinct elements taken r at a time $(r\leq n)$)
Given n distinct symbols, the number of distinct arrangements of symbols of length $r(r\leq n)$ is $nP_r=n(n-1)\cdots(n-(r-1))=\frac{n!}{(n-r)!}$.

Example 6.3.6. How many numbers with three distinct digits are possible using the digits 1, 2, 3, 4 and 5?

There are 5 distinct digits and we are picking 3. Observe that the order is important, since, for example, 234 and 324 are different even though they have the same digits. Applying the above formula, there are $5P_3 = 5 \times 4 \times 3 = 60$ distinct numbers.

Principle 5 (Combinatorial principle)
The number of combinations of r symbols selected without repetition from n distinct symbols is

$$\binom{n}{r} = \frac{nP_r}{r!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}, \quad 0! = 1 \text{ by convention.}$$

Example 6.3.7. How many ways are there for choosing a set of 3 books from a set of 4 books?

Let A,B,C and D be the books. Since the order of choosing the books is not important the following are the ways of choosing 3 books from 4 books:

$$ABC, ABD, BCD, ACD.$$

By applying the above formula, the number of ways in which 3 books can be selected from 4 books is

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = 4.$$

If the order of books is important, then the combination ABC itself can be arranged in 3! ways and they are the following:

$$ABC, ACB, BAC, BCA, CAB, CBA.$$

That is if the order of books selected is also important then there are

$$4 \times 3! = 24 = 4P_3$$

different ways. Since a combination of r things can be arranged in $r!$ different ways, it is obvious that

$$nP_r = r! \times nC_r.$$

Remark 6.3.5. The number of different possible samples of size n under *sampling without replacement* scheme from a population consisting of N units is NC_n.

Example 6.3.8. How many different seating arrangement of 7 students are possible, if three students are seated first?

The first three positions can be filled in 3! ways and then arranging the remaining students in the remaining four positions in 4! ways. These two operations can be simultaneously done in $3! \times 4! = 144$ ways.

Example 6.3.9. Six boys and four girls are asked to sit in a row such that no two girls may sit together. In how many ways they can be placed?

Six boys can be seated in 6! different ways. Corresponding to any one arrangement of boys we have # in the following array to include a girl:

$$\#B_1\#B_2\#B_3\#B_4\#B_5\#B_6\#.$$

Out of these seven positions, girls can be placed in any four positions in $7P_4$ ways. Hence the required answer is $6! \times \frac{7!}{3!} = 6,04,800$.

6.4 PROBABILITY PROBLEMS

We will now look at some examples of probability problems.

Example 6.4.1. A fair coin is tossed 5 times. What is the probability of getting exactly three heads?

There are $2^5 = 32$ possible outcomes in the sample space of this random experiment. Note that the sample space is a set of symmetric elementary outcomes. There are $\binom{5}{3} = 10$ outcomes favourable to the event of getting exactly three heads, namely, $HHHTT, HTHTH, \ldots$. Hence the required probability is $\frac{10}{32} = \frac{5}{16}$.

Example 6.4.2. A bag contains 10 identical balls of which 6 are white and 4 are red. What is the probability that a ball selected at random is white?

Since balls are identical in shape and one ball is to be drawn randomly from 10 balls any one of the 10 balls may be drawn with the same chance. Therefore, the sample space S is symmetric and the total number of outcomes in S is 10. The number of outcomes favourable to the event that the ball drawn is white is 6. So the required probability is $\frac{6}{10}$.

Example 6.4.3. Three unbiased dice are rolled. What is the probability that they all show different faces?

Let A be the event that all dice show different faces.
The total number of outcomes in S is 6^3.
However, for outcomes in A the upper faces are required to be different. Thus when the upper face of one die is freely chosen in 6 ways, the upper face of the next can be freely chosen in 5 ways (different from the first choice). The last may be chosen in 4 ways. Hence the number of favourable outcomes is $6 \times 5 \times 4$.
Required probability is

$$P(A) = \frac{6 \times 5 \times 4}{6^3} = \frac{5}{9}.$$

Example 6.4.4. At a parking area there are 100 vehicles, 60 of which are cars, 30 are vans and the remainder are buses. If every vehicle is equally likely to leave, find the probability of:

(a) A: van leaving first.

(b) B: bus leaving first.

(c) C: car leaving second if either a bus or van had left first.

(a) Let S be the sample space. Then

$$n(S) = 100, n(A) = 30.$$

Probability of a van leaving first is

$$P(A) = \frac{n(A)}{n(S)} = \frac{30}{100} = \frac{3}{10}.$$

(b) Here

$$n(B) = 100 - 60 - 30 = 10.$$

Probability of a bus leaving first is

$$P(B) = \frac{n(B)}{n(S)} = \frac{10}{100} = \frac{1}{10}.$$

(c) If either a bus or van had left first, then there would be 99 vehicles remaining, 60 of which are cars. Let S be the corresponding sample space. Then

$$n(S) = 99, n(C) = 60.$$

Probability of a car leaving after a bus or van is

$$P(C) = \frac{n(C)}{n(S)} = \frac{60}{99} = \frac{20}{33}.$$

Example 6.4.5. Two squares are chosen at random on a chessboard. What is the probability that they have a side in common?

The number of ways of choosing the first square is 64. The number of ways of choosing the second square is 63. There are a total of $64 \times 63 = 4032$ ways of choosing two squares.
If the first square happens to be any of the four corner ones, the second square can be chosen in 2 ways. If the first square happens to be any of the 24 squares on the side of the chess board, the second square can be chosen in 3 ways. If the first square happens to be any of the 36 remaining squares, the second square can be chosen in 4 ways.
Hence the desired number of combinations is $(4 \times 2) + (24 \times 3) + (36 \times 4) = 224$.
Therefore,

$$\text{the required probability} = \frac{224}{4032} = \frac{1}{18}.$$

Example 6.4.6. 8 students come to a hostel, George and Vishnu among them. There are two rooms for 2 students each and one room for four persons. What is the probability of George and Vishnu finding themselves in the room for four?

Number of ways of accommodating the students is

$$n(S) = \frac{8!}{2!2!4!} = 420.$$

Let A denotes the event that George and Vishnu find themselves in the room for four. The number of outcomes favourable to the event A is

$$n(A) = \frac{6!}{2!2!2!} = 90.$$

The required probability is

$$P(A) = \frac{n(A)}{n(S)} = \frac{90}{420} = \frac{9}{42}.$$

Example 6.4.7. Suppose you are a prisoner sentenced to death. The Emperor offers you a chance to live by playing a simple game. He gives you 50 black marbles, 50 white marbles and 2 empty bowls. He then says, "Divide these 100 marbles into these 2 bowls. You can divide them any way you like as long as you use all the marbles. Then I will blindfold you and mix the bowls around. You then can choose one bowl and remove one marble. If the marble is WHITE you will live, but if the marble is BLACK... you will die."
How do you divide the marbles into two bowls so that you have the greatest probability of choosing a WHITE marble?

Place 1 white marble in one bowl, and place the rest of the marbles in the other bowl (49 whites, and 50 blacks). Dividing the marbles in this way you begin with a 50:50 chance of choosing the bowl with just one white marble, therefore life! But even if you choose the other bowl, you still have almost a 50:50 chance of picking one of the 49 white marbles.

Example 6.4.8. A lot of 100 items is inspected as follows: 5 items are chosen at random and tested; If all 5 are good, the lot is accepted. If there are 20 defective items in the lot, what is the probability of accepting the lot?

The number of all possible combinations of 5 items chosen from 100 is $\binom{100}{5}$. The number of possible combinations of 5 good items is $\binom{80}{5}$. Therefore, the probability of choosing 5 good items, that is, accepting the lot is

$$\frac{\binom{80}{5}}{\binom{100}{5}} = 0.32.$$

Example 6.4.9. Find the probability of getting a total score of 15 points in a throw of three identical dice.

This can be solved by the enumeration method by considering all cases of throws where the numbers add up to 15, for example (6, 6, 3), (6, 5, 4), (6, 4, 5), (6, 3, 6) etc. This would be a little tedious process. Instead we associate the probability with the coefficient and score with the index of x in the expansion of $\left(\frac{1}{6}x + \frac{1}{6}x^2 + \cdots + \frac{1}{6}x^6\right)^3$.

Since there are three dice, the expression is raised to power three. The probability of the score 15 is the coefficient of x^{15} in the above expansion, that is, in $\frac{1}{6^3}x^3\left(\frac{1-x^6}{1-x}\right)^3$.

$$\begin{aligned}
\text{Required probability} &= \text{coefficient of } x^{15} \text{ in } \frac{1}{6^3}x^3(1-x^6)^3(1-x)^{-3} \\
&= \text{coefficient of } x^{12} \text{ in } [\frac{1}{6^3}(1-3x^6+3x^{12}-x^{18}) \\
&\quad \times(1+\frac{3x}{1!}+\frac{3.4}{2!}x^2+\cdots)] \\
&= \frac{1}{6^3}\left[\frac{3.4.5...14}{12!}-3\frac{3.4.5...8}{6!}+3\right] \\
&= \frac{1}{6^3}(91-84+3) \\
&= \frac{5}{108}.
\end{aligned}$$

Example 6.4.10. A family was randomly chosen from a set of families having four children. What is the probability that there will be at least one boy, assuming boys and girls are equally likely?

Instead of listing the 16 outcomes BBBB, BBBG etc we simply consider:

$$\begin{aligned}
P(\text{at least one boy}) &= 1-P(\text{no boys}) \\
&= 1-P(GGGG) \\
&= 1-\frac{1}{16} \\
&= \frac{15}{16}.
\end{aligned}$$

Remark 6.4.1. Often when you work out the probability of an event, you sometimes do not need to work out the probability of an event occurring, in fact you need the opposite, the probability that the event will not occur. For example, consider the outcomes favourable to the event " at least one". Then it is always easy to consider the outcomes favourable to the complement event "none". Then

$$P(\text{at least one}) = 1-P(\text{none}).$$

Example 6.4.11. In a certain club the probability that a man picked at random is a lawyer is 0.64, the probability that he is a liar is 0.75, and the probability that he is a lawyer and liar is 0.50. Find the probability that:
(i) he is a lawyer or liar; (ii) he is neither lawyer nor a liar.

Let A denotes the event that the selected man is a lawyer and B the selected man is a liar.
We wish to find the probabilities (i) $P(A\cup B)$ and (ii) $P(\overline{A\cup B})$.

Since $P(A) = 0.65,\ P(B) = 0.75$ and $P(A \cap B) = 0.50$ we get

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.65 + 0.75 - 0.50 \\ &= 0.90. \\ P(\overline{A \cup B}) &= 1 - P(A \cup B) = 1 - 0.90 = 0.10. \end{aligned}$$

Example 6.4.12. Three letters written to Annie, Lisha and Merry are to be placed into three envelopes addressed to them. Assuming the letters are placed in the envelopes randomly, what is the probability that

(a) each lady gets the letter written to her?

(b) only Lisha gets the letter written to her?

(c) only Annie and Lisha get the letters written to them?

Let E_A, E_L and E_M be the envelopes and A, L, M be the letters corresponding to Annie, Lisha and Merry.

(a) Since there are 3 letters and the total number of placements of 3 letters in 3 envelopes is $3! = 6$. They are the following:

Placement→ / Envelope ↓	1	2	3	4	5	6
E_A	A	A	L	L	M	M
E_L	L	M	A	M	L	A
E_M	M	L	M	A	A	L

Out of the above 6 placements, only placement 1 is correct. Hence the probability that each girl gets the letter written to her is $\frac{1}{6}$.

(b) Only Lisha gets the letter written to her implies that the other two letters are misplaced. Out of the total six placements only the fifth one is favourable to this event. Hence required probability is $\frac{1}{6}$.

(c) If Annie and Lisha get the letters written to them, then Merry will automatically get the letter written to her. Therefore the required probability is 0.

Example 6.4.13. Suppose that k balls are placed at random into n urns; what is the probability that each ball is in a different urn?

The sample space is

$$S = \{(u_1, u_2, \ldots, u_k) : 1 \leq u_i \leq n, 1 \leq i \leq k\}$$

where u_i represents the number of the urn into which the i^{th} ball is placed. Since each ball may be placed into any of the n urns, we have $n(S) = n^k$. If E is the event "each ball in a different urn", then

$$n(E) = n(n-1)\cdots(n-k+1).$$

Since there is no reason to believe that any particular arrangement of k balls should occur more or less frequently than any other, we may assume that there is symmetry in S with respect to elementary events.
Hence

$$P(E) = \frac{n(E)}{n(S)} = \frac{n(n-1)\dots(n-k+1)}{n^k}.$$

Example 6.4.14. Suppose that we have recorded the birthdays of k people; what is the probability that at least two of these people have the same birthday?

This is just an application of the preceding problem with $n = 365$.
Instead of calculating $P(B)$, where B is the event "at least two of the people sampled have the same birthday", we will calculate $P(\bar{B})$, where $\bar{B}$ is the event "none of the people samples have the same birthday".

$$P(\bar{B}) = \frac{n(\bar{B})}{n(S)} = \frac{365 \times 364 \times \cdots \times (365-k+1)}{365^k}.$$

Hence

$$P(B) = 1 - \frac{365 \times 364 \times \cdots \times (365-k+1)}{365^k}. \tag{6.4.1}$$

Remark 6.4.2. Intuitively, it may seem that at least 365 people should be present to have a chance of 50% that at least two of them will have the same birthday. However, using the formula (6.4.1) it can be shown that only $k = 23$ persons need to be present to have a 50.7% chance that at least two persons have the same birthday. Only 55 people are needed so that the chance is 98.6%. In other words, if you have a group of 55 people, it is almost certain that at least two persons will have the same birthday.

Example 6.4.15. Suppose that there are six gamblers, each of whom bet on a different number from 1 to 6. The dealer has 3 dice, which are fair, meaning that the chance that each face shows up is exactly $\frac{1}{6}$. The dealer says: "You can choose your bet on a number from 1 to 6. Then I'll roll the 3 dice. If none shows the number you bet, you'll lose ₹1. If one shows the number you bet, you'll win ₹1. If two or three dice show the number you bet, you'll win ₹3 or ₹5, respectively." Is it a fair game?

If he rolls three different numbers, for example $(1,2,3)$, the three gamblers who bet 1, 2, 3 each wins ₹1 while the three gamblers who bet 4, 5, 6 each loses ₹1.
If two of the dice he rolls show the same number, for example $(1,1,2)$, the gambler who bet 1 wins ₹3, the gambler who bet 2 wins 1, and the other four gamblers each loses 1.
If all three dice show the same number, for example $(1,1,1)$, the gambler who bet 1 wins ₹5, and the other five gamblers each loses ₹1.
In each case, the dealer neither wins nor loses. Hence it is a fair game.

Example 6.4.16. Two balls are selected at random from an urn containing ten balls numbered 1, 2, …, 10. Calculate the probability of the event E "sum of the two balls is odd".

We shall consider three different ways in which the sample could have been drawn and calculate $P(E)$ in each case.

(a) Suppose first that the two balls are drawn together or they are drawn without replacement. Then there are $\binom{10}{2} = 45$ possible combinations of two balls, which constitutes the sample space. Thus $n(S)$=45.
Now E occurs when one ball is odd and the other ball is even which can happen in $\binom{5}{1} \times \binom{5}{1} = 25$ ways since there are five ways to choose one of the five odd balls and five ways to choose one of the five even balls. That is $n(E) = 25$.
Therefore,

$$P(E) = \frac{n(E)}{n(S)} = \frac{25}{45} = \frac{5}{9}.$$

(b) If the two balls are drawn with replacement, then, the resulting sample space S' contains $n(S') = 10^2 = 100$ possible outcomes. The event E can be considered as a union $A \cup B$, in which
A: first ball is odd, second ball is even
B: first ball is even, second ball is odd.
Now $n(A) = n(B) = 5^2 = 25$ and since $A \cap B = \emptyset$, we have $n(A \cup B) = n(A) + n(B) = 50$. Therefore,

$$P(E) = P(A \cup B) = \frac{n(A \cup B)}{n(S')} = \frac{50}{100} = \frac{1}{2},$$

this time.

Example 6.4.17. A survey was taken on 60 classes at a college to find the total number of left-handed students in each class. The table below shows the results:

Number of left-handed students(x)	0	1	2	3	4	5
Number of classes(f)	2	4	10	24	16	4

A class was selected at random. Find the probability that

(a) (i) A: the class has two left-handed students.

(ii) B: the class has at least three left-handed students.

(b) Given that the total number of students in the 60 classes is 1800, find the probability that a student randomly chosen from these 60 classes is left-handed.

(a) (i) Let S be the sample space. Then

$$n(S) = 60, n(A) = 10.$$

Hence

$$P(A) = \frac{n(A)}{n(S)} = \frac{10}{60} = \frac{1}{6}.$$

(ii) Now

$$n(B) = 24 + 16 + 4 = 44.$$

Therefore

$$P(B) = \frac{n(B)}{n(S)} = \frac{22}{60} = \frac{11}{30}.$$

(b) Firstly, we shall find the total number of left-handed students in the college. Total number of left-handed students $= \Sigma fx = 0+4+20+72+64+20 = 180$. Here, the sample space S contains all the students in the 60 classes. Let C be the event that a student is left-handed. Then

$$n(S) = 1800, n(C) = 180.$$

Hence

$$P(C) = \frac{180}{1800} = \frac{1}{10}.$$

Example 6.4.18. The performance of 500 mobile phones manufactured by a certain company was observed over a period of five years. The number of mobile phones which went out of order within the first year, second year etc. were as follows:

Year	1	2	3	4	5	Above 5 years	Total
Number of mobile phones	5	10	47	175	225	38	500

What is the probability that a randomly chosen mobile phone manufactured by this company would be out of order (i) within two years, (ii) after three years but within five years?

Let the events corresponding to (i) and (ii) are denoted by A and B respectively. From the above table we find that the number of mobile phones which went out of order within two years is 5+10=15. Assuming symmetry in the sample space we have

$$P(A) = \frac{n(A)}{n(S)} = \frac{15}{500} = 0.03.$$

The number of mobile phones which went out of order after three years but within five years is 175+225=400, from the above table. Hence

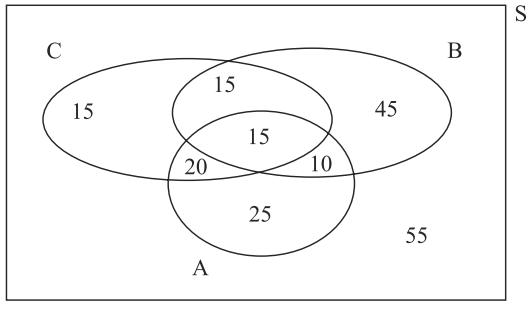
$$P(B) = \frac{n(B)}{n(S)} = \frac{400}{500} = 0.80.$$

Example 6.4.19. A survey conducted among 200 rubber farmers in Kottayam district revealed that they have different hybrids of rubber trees in their plantations. Largely planted hybrids are RRII - 105, PB - 260 and RRIM-600 and let them be denoted by A, B and C respectively. The following are the number of plantations having one, two or three of these hybrids:

Hybrids	A	B	C	AB	AC	BC	ABC
Number of farmers	70	85	65	25	35	30	15

A farmer is randomly chosen from these 200 rubber farmers. If all the farmers have the equal chance of getting selected find the probability that he planted (i) A alone (ii) at least two of the hybrids A or B or C (iii) both B and C (iv) none of the hybrids A, B or C.

The accompanying Venn diagram would give the answers at once.



(i) From the diagram we get the number of farmers who planted A alone is 25 and hence

$$P(A \text{ alone }) = \frac{25}{200} = \frac{1}{8}.$$

(ii) Number of farmers who planted at least two among A, B, C is 20+10+15+15=60.

$$\text{Required probability} = \frac{60}{200} = \frac{3}{10}.$$

(iii) Number of farmers planted both B and C is 30. Note that those who planted B and C consist of those who panted A also and those who do not planted A. Therefore,

$$P(BC) = \frac{30}{200} = \frac{3}{20}.$$

(iv) Number of farmers planted none of A or B or C is 55.

$$\text{Required probability} = P(\overline{A \cup B \cup C}) = \frac{55}{200} = \frac{11}{40}.$$

Example 6.4.20. A swindler once approached an honest man with a die. He handed over him the die and told him about the bet. If the man rolled a ONE, he wins, and gets back twice the amount of his bet. If not, the swindler would keep the bet.

"But...my chances are only one out of six," retorted the man.

"True," grinned the swindler, "But I'll give you three tries to get a one."

The man considered. Three tries, with each try having a $\frac{1}{6}$ chance of winning. So his chance of winning is $\frac{3}{6} = \frac{1}{2}$. Why not give it a try?

Is the bet really fair? If not, what are the chances of the man winning?

As you have guessed, the bet is not fair. He had calculated the probability wrongly. Probability does not accumulate, like $\frac{1}{6} \times 3$.

The probability of the man not getting a ONE in three throws is: $\frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$, which is $\frac{125}{216}$. This is the probability of the swindler winning.

Hence, the remaining fraction, $\frac{91}{216}$, is the actual chance of the man winning.

Example 6.4.21. A certain product was found to have two types of minor defects. Suppose the probability that a randomly chosen item has only a type 1 defect is 0.2 and the probability that it has only a type 2 defect is 0.3. Also, the probability that it

has both defects is 0.1. Find the probabilities of the following events:

A : It has either a type 1 defect or a type 2 defect or both.
B : It item does not have either of the defects.
C : It has defect 1, but not defect 2.
D : It has exactly one of the two defects.

Let D_1 and D_2 denote the events that the item has defects 1 and 2 respectively. Given that $P(D_1) = 0.2, P(D_2) = 0.3$ and $P(D_1 \cap D_2) = 0.1$. Then, by the definitions of the events and the properties of probabilities, it follows that:

$$\begin{aligned} P(A) &= P(D_1 \cup D_2) = P(D_1) + P(D_2) - P(D_1 \cap D_2) = 0.2 + 0.3 - 0.1 = 0.4. \\ P(B) &= P[(D_1 \cup D_2)^c] = 1 - P(D_1 \cup D_2) = 1 - 0.4 = 0.6. \\ P(C) &= P(D_1 \cap \overline{D}_2) = P(D_1) - P(D_1 \cap D_2) = 0.2 - 0.1 = 0.1. \\ P(D) &= P(A) - P(D_1 \cap D_2) = 0.4 - 0.1 = 0.3. \end{aligned}$$

6.4.1 Probability for Uncountable Outcomes

Example 6.4.22. A hole is detected in an underground water pipeline of length 10,000 meters. The hole is equally likely to be located anywhere and if we decide to assign probabilities proportional to lengths then the probability of the hole lying in any region $A \subset [0, 10,000]$ is

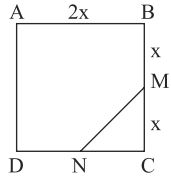
$$P(A) = \frac{\text{length of } A}{10,000}.$$

Find the probabilities of the following events:
B: the hole is located within 250 meters on either side of the center of the pipeline.
C: the hole is located within 500 meters of either end of the pipeline.

By this rule of assigning probabilities proportional to lengths, it follows that

$$\begin{aligned} P(B) &= \frac{\text{Length of the interval}[4750, 5250]}{10,000} = \frac{500}{10,000} = 0.05. \\ P(C) &= \frac{\text{Length of the region}[0, 500] \cup [9500, 10,000]}{10,000} = \frac{1000}{10,000} = 0.10. \end{aligned}$$

Example 6.4.23. *ABCD* is a square. *M* is the midpoint of *BC* and *N* is the midpoint of *CD*. A point is selected at random in the square. What is the probability that it lies in the triangle *MCN*?



Based on the given information, it is reasonable to assign probabilities proportional to the areas. Let $2x$ be the length of the square.

$$\begin{aligned} \text{Area of square} &= 2x \times 2x = 4x^2. \\ \text{Area of triangle } MCN &= \frac{1}{2}x^2. \end{aligned}$$

If we decide to assign probabilities proportional to area, then

$$\begin{aligned} P(\text{point randomly chosen lies in the triangle MCN}) &= \frac{1}{2}x^2 \div 4x^2 \\ &= \frac{1}{8}. \end{aligned}$$

Example 6.4.24. Two friends who take train to their jobs from the same station arrive to the station uniformly randomly between 9 and 9.20 in the morning. They are willing to wait for one another for 5 minutes, after which they take a train together or alone. What is the probability of their meeting at the station?

In a Cartesian system of coordinates (x,y), a square of side 20 (minutes) represents all the possibilities of the morning arrivals of the two friends at the railway station. The shaded region A is bounded by two straight lines, $y = x+5$ and $y = x-5$, so that inside A, $|x-y| = 5$. It follows that the two friends will meet only provided their arrivals x and y fall into region A. The probability of this happening is given by the ratio of the area of A to the area of the square.

$$\text{Area of the square} = 400.$$

$$\text{Area of the shaded region A} = 400 - [(\frac{1}{2} \times 15 \times 15) + (\frac{1}{2} \times 15 \times 15)] = 175.$$

$$\text{Required probability} = \frac{175}{400} = \frac{7}{16}.$$

6.4.2 Odds Ratio

For any event A, let $P(A) = p$, thus $P(A^c) = q = 1-p$. The odds of event $A = \frac{p}{q} = \frac{p}{1-p}$, $p \neq 1$, that is, the probability that A does occur, divided by the probability that it does not occur. Note that if odds = 1, then A and A^c are equally likely to occur. If odds > 1 (likewise, < 1), then the probability that A occurs is greater (likewise, less)

than the probability that it does not occur. For example, suppose that the probability of contracting a certain disease in a particular group of high risk individuals is $P(D^+) = 0.80$, so that the probability of being disease-free is $P(D^-) = 0.20$. Then the odds of contracting the disease in this group is equal to $0.80/0.20 = 4$(or 4 to 1). That is, within this group, the probability of disease is four times larger than the probability of no disease. Likewise, if in a reference group of low risk individuals, the prevalence of the same disease is only $P(D^+) = 0.05$, so that $P(D^-) = 0.95$, then their odds $= 0.05/0.95 = 1/19$. As its name suggests, the corresponding *odds ratio* between the two groups is defined as the ratio of their respective odds, that is, $\frac{4}{1/19} = 76$. That is, the odds of the high-risk group contracting the disease are 76 times larger than the odds of the low-risk reference group. Odds ratios have nice properties, and are used extensively in epidemiological studies.

Example 6.4.25. A card is drawn from a well shuffled pack of 52 cards. A gambler bets that it is either a heart or an ace. What are the odds against his winning the bet?

Let A be the event that the card drawn is a heart and B be the event that it is an ace. Since there are 13 hearts and 4 aces, we have

$$P(A) = \frac{13}{52}, P(B) = \frac{4}{52} \text{ and } P(A \cap B) = \frac{1}{52}.$$

Then

$$\begin{aligned} P(\text{the gambler wins}) = P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} \\ &= \frac{4}{13}. \\ P(\text{the gambler loses}) &= 1 - P(A \cup B) \\ &= 1 - \frac{4}{13} \\ &= \frac{9}{13}. \end{aligned}$$

A probability of $\frac{9}{13}$ is equivalent to odds of $\frac{9}{13}$ to $\frac{4}{13}$, which is usually expressed as 9 to 4.

6.5 SUBJECTIVE PROBABILITY

If the experiment can be repeated a number of times under similar conditions or the sample space consists of only a finite number of equally likely and mutually exclusive outcomes then we can assign probabilities to events associated with this experiment. There are situations which cannot be repeated, in which one would like to assign probabilities. For example, a new cinema may have different probability of 'box office success' if it is released this month rather than next month. This is an "experiment" that cannot

be performed over and over under similar conditions as the conditions of the film market is not static. If the cinema is not released this month but next, the experiment will be performed under different conditions and thus will be a different experiment.

There are also situations in which chances are computed by repeated experimentation, but in which one would like to be able to apply them to a particular case in which really "chance" may not be appropriate. For instance, one is told that a certain medical treatment has a 80 percent record of success; can he then apply this to himself, or is the chance aspect of the treatment overshadowed by particular circumstances in his own case which are more pertinent? Some people advocates subjective or personal concept of probability, as opposed to the more common view that probability is objective, something attributable to the phenomenon itself without reference to the person computing it. Both views have their uses.

In dealing with events associated with the above type of experiments, decision makers sometimes use a *subjective* or *personal definition of probability*. According to this definition, *the probability of an event is the degree of confidence or belief on the part of the decision maker that the event will occur.* In assigning probability in this way, the decision maker or statistician must use his knowledge, experience and intuition together with whatever objective information can be collected with regard to the event under consideration. The subjective method is used as a last resort when the other approaches are not practical.

Probabilistic statements can be interpreted in different ways. For example, how would you interpret the following statement? There is a 40 percent chance of rain today. The interpretation will vary depending on the context in which the statement is made. If the statement was made as part of a forecast by the weather forecasting station, then something like the following interpretation might be appropriate: In the recent history of this region, of all days on which present atmospheric conditions have been experienced, rain has occurred on approximately 40 percent of the times.This is an example of the frequentist interpretation of probability.

Suppose, however, that you had just looked at the sky, wondering if you should carry an umbrella to the college, and asked your mother if she thought that it was going to rain. Unless your mother is a metereologist, it is not posssible that she possesses the knowledge required to make a frequentist statement! If her response was a casual "I'd say that there's a 40 percent chance," then something like the following interpretation might be appropriate: I believe that it might very well rain, but that it's a little less likely to rain than not. This is an example of the subjectivist interpretation of probability. With this interpretation, a probability expresses the strength of one's belief.

It seems logical to suppose that a "measure of belief" should satisfy all of the axioms of probability. Hence, whether we interpret probability as a long-run relative frequency of occurrence or measure of belief, its mathematical properties remain unchanged.

6.6 EXERCISES

1. A menu lists two soups, three meat dishes and five desserts. How many different meals are possible consisting of one soup, one meat dish and a dessert?

2. A die is rolled four times and a coin is tossed twice. Find how many outcomes are there in the sample space of this random experiment.

3. Ten people have gathered at a party. Find the number of handshakes that take place if each person shakes hands with everyone else in the group.

4. Using seven consonants and five vowels, how many words consisting of four consonants and three vowels can be formed?

5. A student is to answer 10 out of twelve questions in an examination. How many choices has he? How many if he must answer the first three questions? How many if he must answer at least four of the first five questions?

6. Out of five men and five women, how many ways are there to form a committee consisting of three women and two men?

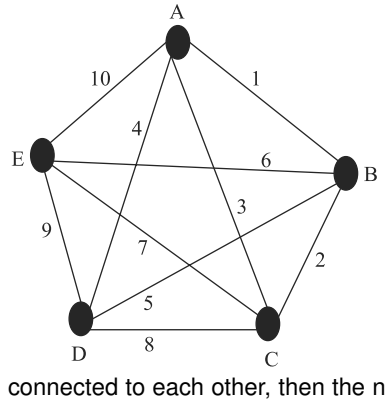
7. A family consisting of an old man, six adults and four children is to be seated in a row for dinner. The children wish to occupy the two seats at each end and the old man refuses to have a child on either side of him. In how many ways can the seating arrangements be made for the dinner.

8. A telephone number consists of seven digits, the first three representing the exchange. How many different telephone number are possible within a particular exchange?

9. In how many ways can five students line up to get on a bus? In how many ways can they line up, if two of the students refuse to follow each other?

10. An agriculturist wants to test 3 fertilizers, each at 4 levels of concentration, on 5 different types of soil. How many different tests must he make?

11. A group of points A, B, C, D and E are connected to each other with lines as below. How many triangles do they form?



Hint: If n points are all connected to each other, then the number of geometrical figures having m sides formed is equal to $\binom{n}{m}$.

12. In a group of 100 college students 70 are enrolled in a course in psychology 58 are enrolled in a course in sociology, and 45 are enrolled in both. How many of these students are not enrolled in either course?
Hint: Draw a suitable Venn diagram and fill in the numbers associated with the various regions.

13. A red flag, a white flag and a green flag are to be arranged in a row for a signal. How many distinct signals are possible?

14. A quizmaster has developed 5 questions designed to ask for a certain competition. He will select 3 of these questions. How many different arrangements are there for the order of the 3 selected questions?

15. Voters must choose from among three candidates for president, two candidates for secretary and four candidates for treasurer, and must vote Yes or No on a public issue. How many different ways can voters mark their ballots?

16. A political organization has 60 men and 40 women members. How many four person delegations to the national convention are possible? How many if the delegation must contain two men and two women? How many if the delegation cannot be constituted entirely of members of the same sex?

17. Describe the sample space for the following random experiments.

    (a) Shooting at the center of a circle with radius 50cm and noting the distance by which he misses the target if it hits within the circle.

    (b) A machine produces items which may be classified as good or defective. Three such items are inspected and classified.

    (c) Number of accidents on Delhi-Agra highway in a day are recorded.

    (d) Tossing a coin until first head turns up.

    (e) Noting the voltage in an electric terminal at the time instant t.

    (f) Looking the number of rubber trees lost in a plantation due to wind in an year.

    (g) Noting the blood pressure of a person at the time instant t.

    (h) Counting the number of fishes per $1000m^2$ randomly chosen area of Vembanadu lake.

    (i) Looking the amount with the gambler at the end of the nth game.

18. Give some examples of random experiments from the fields of social sciences, insurance and banking.

19. A coin is tossed once. Then if it comes up head, a die is thrown once; if it comes up tail, it is tossed twice more. List

    (i) the ten elements of the sample space S.

    (ii) the elements of S corresponding to the event that exactly one head occurs.

    (iii) the elements of S corresponding to the event that at least two tails occur or a number greater than 4 occurs.

20. A box contains 10 balls of which 5 are red and 5 are white. If two balls are randomly drawn with replacement what is the probability that two white balls are drawn?
(a) 1 (b) 0.25 (c) 0.20 (d) 0.50.

21. If three fair coins are tossed, the probability of two heads is:
(a) $\frac{3}{8}$ (b) $\frac{2}{3}$ (c) $\frac{1}{8}$ (d) $\frac{1}{4}$.

22. A bag contains six red balls, four blue balls and two yellow balls. If two balls are drawn at random without replacement, the probability that one ball will be red and the other will be blue is:
(a) 0.364 (b) 0.333 (c) 0.182 (d) none of these.

23. A bag contains six red balls, four blue balls, and two yellow balls. If two balls are drawn at random with replacement, the probability either ball is red is:
(a) 0.227 (b) 0.250 (c) 0.750 (d)none of these.

24. A and B are mutually exclusive events, and $P(A) = 0.25$, $P(B) = \frac{1}{3}$. $P(A$ or $B)$ is:
(a) 0.583 (b) 0.5 (c) 0.083 (d) none of these.

25. The probability of the intersection of two mutually exclusive events is always:
(a) 0.5 (b) 0 (c) 1 (d) none of these.

26. For any two events A and B, $P(A \cap B^c)$ is equal to:
(a) $P(A) - P(B)$ (b) $P(B) - P(A)$ (c) $P(B) - P(AB)$ (d) $P(A) - P(AB)$.

27. A survey conducted in a city found that 47% of teenagers have a part time job. The same survey found that 78% plan to attend college. If a teenager is chosen at random, what is the probability that the teenager has a part time job and plans to attend college?
(a) 60% (b) 31% (c) 37% (d) data is insufficient

28. From a pack of 52 cards, two cards are drawn at random. The probability that one is an ace and the other is a king is:
(a) $\frac{2}{13}$ (b) $\frac{1}{169}$ (c) $\frac{16}{169}$ (d) $\frac{8}{663}$.

29. The data reveals that 10 per cent patients die in a particular type of operation. A doctor performed 9 operations and all of them survived. If the 10th patient also being operated, what is the probability that the patient will survive?
(a) 0 (b) 0.90 (c) 1 (d) none of these.

30. The probability that a leap year will have 53 Sundays is :
(a) $\frac{1}{7}$ (b) $\frac{2}{7}$ (c) $\frac{2}{53}$ (d) $\frac{52}{53}$.

31. A balanced coin is tossed six times. The probability of obtaining heads and tails alternately is:
(a) $\frac{1}{64}$ (b) $\frac{1}{2}$ (c) $\frac{1}{32}$ (d) none of these.

32. Three identical houses were available in a locality for allotment. Three persons applied for a house. If all the three houses are equally preferable to these persons what is the probability that all the three persons applied for the same house?
(a) $\frac{1}{3}$ (b) $\frac{2}{9}$ (c) $\frac{1}{27}$ (d) 1.

33. In the problem of question 32, the probability that each of the three applied for a different house is:
(a) $\frac{1}{9}$ (b) $\frac{1}{27}$ (c) 1 (d) $\frac{2}{9}$.

34. A train is scheduled to leave the station at 7am. However, it is equally likely to actually leave the station at any time from 6:55 to 7.15 am. What is the probability it will depart the station early?
(a) 0.25 (b) 0.33 (c) 0.67 (d) 0.75.

35. A fair coin is tossed repeatedly until a head is obtained. The probability that the coin has to be tossed at least four times is:
(a) $\frac{1}{2}$ (b) $\frac{1}{4}$ (c) $\frac{1}{6}$ (d) $\frac{1}{8}$.

36. Sibil flipped a coin and it landed on head. If he flips it again, it will:
(a) likely be head; (b) unlikely be head
(c) certainly be head; (d) have an equal chance to be head or tail.

37. Two dice are rolled. Let A be the event of getting odd numbers on both and B be the event of getting a pair of same numbers on both dice. Are A and B mutually exclusive?

38. Four dice are rolled simultaneously. What is the number of possible outcomes in which at least one of the die shows 6?

39. A ball is drawn at random from an urn containing 30 balls numbered 1,2,...,30. The number is recorded and the ball is replaced. A second ball is drawn at random and its number recorded. What is the probability that the two balls do not have the same number?

40. Six married couples are standing in a room. If two people are chosen at random find the probability that

    (a) they are married to each other;

    (b) one is male and the other is female.

41. There are three traffic lights on your way home. As you arrive at each light assume that it is either red (R) or green (G) and that it is green with probability 0.7. List the elements of the sample space. Is the sample space symmetric?

42. Stephy has 9 pairs of dark blue socks and 9 pairs of black socks. He keeps them all in the same bag. If he picks out three socks at random what is the probability that he will get a matching pair?

43. Two students and two professors are arranged randomly in a row for a panel discussion. What is the probability that the students and professors alternate in the row?

44. The key to a locked door is one of 12 keys in a cabinet. If a person selects two keys at random from the cabinet and takes them to the door, what is the probability that he can open the door?

45. Compute the probability that, out of five people, at least two have the same birth month.

46. A personal assistant randomly puts n letters in n addressed envelopes. What is the probability that all letters are wrongly placed?

47. Nicy and Anju play a game where they simultaneously exhibit their right hands with one, two, three, four or five fingers extended. Write down the sample space and find the probabilities of the following events.

    (a) Nicy and Anju extend the same number of fingers.

    (b) Anju shows an odd number of fingers.

    (c) Nicy and Anju together extend five fingers.

48. The letters of the word STATISTICS are arranged in some order. What is the probability that the three S's are consecutive?

49. If the letters of the word CINEMA are arranged in all possible ways, find the probability that:

    (i) the word ends in a vowel.

    (ii) the word starts with a consonant and ends in a vowel.

50. A secretary types four letters to four people and addresses the four envelopes. If she inserts the letters at random, each in a different envelope, what is the probability that exactly two letters will go into the right envelope?

51. If $P(A) = \frac{1}{2}, P(B) = \frac{1}{4}, P(A \cup B) = \frac{3}{8}$, what is the value of $P(A \cap B)$?

52. There are 100 students in a lecture hall, 40 know Hindi, 50 know English and 35 know Malayalam. 20 students know English and Hindi, 8 know English and Malayalam, 10 know Hindi and Malayalam. 5 students know Hindi, English and Malayalam. If one of the students is picked at random from the hall then what is the probability that he knows:

    (i) either Hindi or English;

    (ii) only English;

    (iii) none of these languages.

53. Suppose I flip two fair identical coins without letting you see the outcome, and I tell you that at least one of the coins came up heads. What is the probability that the other coin is also a head?

54. A waiting room has six seats arranged in a row. Suppose that three people enter the room and choose their seats at random. What is the probability that they sit with no empty seats between them? What is the probability that there is exactly one empty seat between any two of them?

55. If 4 married couples are arranged randomly in a row, find the probability that no husband sits next to his wife?

56. A group of $n > 2$ people sit randomly at a round table. What is the probability that a certain two people will sit next to each other?

57. When two fair identical coins are tossed the result can be two heads, one head and one tail, or two tails, and hence each of these events has a probability of $\frac{1}{3}$. What is wrong with this argument? What is the correct argument?

58. A group of 6 men and 6 women is randomly divided into two groups of size 6 each. What is the probability that both groups will have the same number of men?

59. A criminologist says that the probability that a college drop-out will be arrested for theft is 0.10. Again, the probability that he will be arrested for either theft or homicide (or both) is 0.12.

    (i) Is the probability that he will be arrested for homicide equal to 0.02? Why or why not?

    (ii) Is the probability that he will be arrested for homicide less than 0.02? Why or why not?

60. Two people alternately toss an unbiased coin until one of them gets a head. What is the probability that the player making the first toss wins the game?

61. This question is about the paradox of the Chevalier De M*éré*. The first was "the probability of obtaining at least one 6 when a balanced die is rolled 4 times". The second was "the probability of obtaining at least one double-6 when two balanced identical dice are rolled 24 times".
    He thought that the two compound events had the same probability, namely, $\frac{4}{6}$ and $\frac{24}{36}$, respectively. By calculating the correct values, establish that he was wrong in both cases.

62. There are six men and twelve women. Half of the men have gray hair and so do half of the women. What are the chances that a person chosen at random will be a man, a person with gray hair or both?

63. How many times would you have to flip a fair coin to have a 90% chance of obtaining at least one head?

64. How many fair dice must be thrown to give a better than 50% chance that at least one of the dice will show a one?

65. Five people, designated as A, B, C, D and E are arranged in linear order. Assuming that each possible order is equally likely, what is the probability that

    (i) there is exactly one person between A and B;

    (ii) there are exactly two people between A and B?

66. Prove the following relations

    (i) If $E \subset F$, then $F^c \subset E^c$.

    (ii) $F = (F \cap E) \cup (F \cap E^c)$, and $E \cup F = E \cup (E^c \cap F)$.

67. Two basket-ball players in turn throw a ball until the first hit. The first player to throw the ball into the basket wins. The events are A_k=$\{$the first player throws the ball into the basket on his k-th throw$\}$, B_k=$\{$the second player throws the ball into the basket on his k-th throw$\}$; A=$\{$the first player wins$\}$, B=$\{$the second player wins$\}$. The first player is the first to throw. Find the composition of the set of elementary outcomes and write events A and B in terms of the algebra of events.

68. Let E, F and G be three events. Find expressions for the events so that of E, F and G:
(a) both E and G but not F occur; (b) only E occurs;
(c) at least two of the events occur; (d) all three occur;
(e) at most two of them occur; (f) exactly two of them occur;
(g) at most one of them occurs; (h) at least one of the events occurs.

69. Give verbal descriptions and simplest expressions of the following events:
(a) $(E\cup F)\cap(E\cup F^c)$ (b) $(E\cup F)\cap(F\cup G)$ (c) $(E\cup F)\cap(E^c\cup F)\cap(E\cup F^c)$.

70. Write down some examples of sigma fields other than the collection of all subsets of a given sample space S.

71. (The game of rencontre) An urn contains n tickets numbered 1,2,...,n. The tickets are shuffled thoroughly and then drawn one by one without replacement. If the ticket numbered r appears in the r-th drawing, this is termed as a *match* (French: rencontre). Show that the probability of at least one match is

$$1-\frac{1}{2!}+\frac{1}{3!}-\cdots+\frac{(-1)^{n-1}}{n!}\to 1-e^{-1} \text{ as } n\to\infty.$$

72. If A and B are disjoint events, are A^c and B^c disjoint? Are $A\cap C$ and $B\cap C$ disjoint? What about $A\cap C$ and $B\cap C$?

73. If $A_n\subset A_{n-1}\subset\cdots\subset A_1$, show that $\bigcap_{i=1}^{n} A_i = A_n$, $\bigcup_{i=1}^{n} A_i = A_1$.

74. If $A, B_1, B_2, \ldots$ are arbitrary events show that

$$A\cap\left(\bigcup_i B_i\right) \;=\; \bigcup_i (A\cap B_i).$$

This is the distributive law with infinitely many factors.

75. Show that the probability that exactly one of the events E or F occurs equals

$$P(E)+P(F)-2P(E\cap F).$$

76. Let A and B be observable events of a random experiment. Show that

$$P(A\triangle B)=2P(A\cup B)-P(A)-P(B).$$

77. If $P(E) = 0.9$ and $P(F) = 0.8$, show that $P(E \cap F) \geq 0.7$. In general, prove Bonferroni's inequality, namely,

$$P(E \cap F) \geq P(E) + P(F) - 1.$$

78. Prove that $P(E \cap F^c) = P(E) - P(E \cap F)$.

79. Use induction to generalize Bonferroni's inequality to n events. That is, show that

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) \geq P(E_1) + \cdots + P(E_n) - (n-1).$$

Chapter 7

CONDITIONAL PROBABILITY AND BAYES' THEOREM

"Statistics is not so simple a subject that it can be codified in terms of a simple recipe that yields satisfactory methods in all problems"- J.C.Kiefer.

7.1 INTRODUCTION

If you were to make a bet on the outcome of World Cup Cricket Series, on whom you bet and how much you bet could very well be different if you were to place your bet after the league matches had been played, rather than before the series began. Information of what happened in the league matches would almost certainly affect your assessment of who was likely to win the series.

Similarly, the fact that a certain event A has already occurred will most likely affect the chances of another event B occurring. For example, if we were to roll a fair die twice, then there is one chance in eighteen of getting a sum 11 because exactly 2 of the 36 possible outcomes have sum 11. Now after rolling the die once we had seen a 1 or 2, then we would be certain that the sum of the two rolls of the die could not possibly be 11. However suppose that we had seen a 5 in the first toss. How should we revise our estimate of the likelihood of getting sum11? If the first roll resulted in a 5, then only the following 6 elements of the original sample space remain relevant:

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

We know that the other 30 elements did not occur. Of the 6 elements that might have occurred, only one, $(5,6)$, has sum 11. Therefore we can now revise our estimate of the chance of obtaining sum 11 as 1 in 6, which is higher than 1 in 18.

In many circumstances it happens that we acquire partial knowledge of the outcome of a random experiment before the complete result becomes known. In some

random experiments we may see that the final result of the experiment is revealed to us only in part though the experiment may be completed. Examples will make clear that when such partial information is obtained, the original probability should be modified.

Example 7.1.1. A card is picked at random from a deck of well shuffled 52 playing cards and it is given that the card is red. What is the probability that it is an ace?

Given that the card is a red one and hence we need only to concentrate on 26 red cards. Of these there are 2 aces. Therefore, the probability of picking an ace given the information that the card is red is $\frac{2}{26} = \frac{1}{13}$.
Alternatively we can obtain this answer using conditional probability as follows:
Let A be the event of getting an ace and B be the event of getting a red card. Then the probability of getting an ace given that a red card is chosen, denoted by $P(A|B)$, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/52}{26/52} = \frac{1}{13},$$

where $A \cap B$ can be identified as the event that the selected card is an ace as well as a red card.

Example 7.1.2. A market research organization has services under warranty provided by 30 new car dealers in a certain city and the data is summarized in the following table.

Business experience	Good service	Poor service
5 years or less	10	2
More than 5 years	6	12

If a person buys a car from a randomly chosen car dealer from these 30 dealers, what is the probability that he will get good service under warranty? Also, if a person randomly selects one of the dealers who has been in business for more than 5 years, what is the probability that he gets one who provides good service under warranty?
Let A denotes the selection of a dealer who provides good service under warranty. By "randomly" we mean that all possible selections are equally likely and hence

$$P(A) = \frac{n(A)}{n(S)} = \frac{10+6}{30} = \frac{8}{15}$$

To answer the second question, let B denote the selection of a dealer who has been in business for more than 5 years. Now we want to calculate the probability of A given the information that the event B has occurred and let it be denoted by $P(A|B)$. Therefore we limit ourselves to the reduced sample space which consists of the last line of the table, namely, $6+12 = 18$ dealers who have been in business for more than 5 years. Of these, 6 provide good service under warranty. Hence

$$P(A|B) = \frac{6}{18} = \frac{1}{3}.$$

Note that

$$P(A|B) = \frac{6/30}{18/30} = \frac{P(A \cap B)}{P(B)}.$$

7.2 CONDITIONAL PROBABILITY

In Example 7.1.1 and Example 7.1.2 we expressed the conditional probability in terms of two probabilities defined on the whole sample space S. Generalizing from these examples, let us make the following definition of conditional probability.

Definition 7.2.1. If A and B are any two events with respect to a sample space S and $P(B) \neq 0$, then the conditional probability of A given B is defined as

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

Based on the definition of conditional probability we can write the following theorem and it is called the *multiplication theorem* of probability for the case of two events.

Theorem 7.2.1. *If A and B are any two events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then*

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

Similarly we can obtain the multiplication theorem for the case of three events as follows:

Theorem 7.2.2. *If A_1, A_2 and A_3 are any three events in a sample space S such that $P(A_1) \neq 0$ and $P(A_1 \cap A_2) \neq 0$, then*

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2).$$

Proof. Writing $A_1 \cap A_2 \cap A_3$ as $(A_1 \cap A_2) \cap A_3$ and using the multiplication theorem for two events twice, we get

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P[(A_1 \cap A_2) \cap A_3] \\ &= P(A_1 \cap A_2)P(A_3|A_1 \cap A_2) \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2). \end{aligned}$$

Further generalization of multiplication theorem to k events is now straightforward and the resulting formula can be proved by mathematical induction.

Example 7.2.1. Three cards are drawn without replacement from an ordinary deck of well shuffled playing cards. Find the probability of not getting a spade.

Let A_i, $i = 1,2,3$ denote the event that the i-th drawn card is not a spade. Then we are looking for

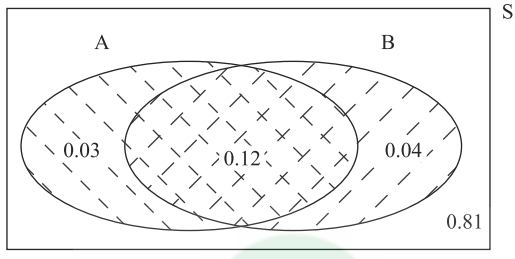
$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \\ &= \frac{39}{52} \times \frac{38}{51} \times \frac{37}{50}. \end{aligned}$$

To find$P(A_2|A_1)$ we consider the restricted sample space. If the first card is not a spade, there remains 51 cards in the deck including 13 spades so that the probability of not getting a spade on the second draw is $\frac{38}{51}$. Similarly $P(A_3|A_1 \cap A_2) = \frac{37}{50}$.

Notice that the above probability can be obtained directly using combinatorial reasoning as $\binom{39}{3}/\binom{52}{3}$. If the cards were drawn with replacement, the probability would be quite different, $\frac{39}{52} \times \frac{39}{52} \times \frac{39}{52} = \frac{27}{64}$.

Example 7.2.2. A health study produced a data on lung cancer in relation to smoking status of 100 persons is given below. Suppose a person is randomly chosen from the population. Let A denote the event of getting a person having lung cancer and B denote the event of getting a smoker.

	Lung cancer (A)	No lung cancer (A^c)	Total
Smoker (B)	12	4	16
Nonsmoker (B^c)	3	81	84
Total	15	85	100



Assuming symmetry in the sample space we have,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{n(A \cap B)}{n(B)} = \frac{12}{16} = 0.75.$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{n(A \cap B)}{n(A)} = \frac{12}{15} = 0.80.$$

Note that $P(A|B)$ is much greater than $P(A)$ and similarly $P(B|A)$ is much greater than $P(B)$. Further, $P(A|B) \neq P(B|A)$ in general.

Example 7.2.3. A bowl contains two white (W) beads and three black (B) beads. One bead is drawn at random, and then another is drawn at random from those remaining. Then we have the following probabilities for each of the different events.

The phrase "drawn at random" means that we are assuming symmetry in the sample space and hence the available beads at each drawing are to be assigned equal probabilities. For instance, the probability that the first bead is black is 3/5, the ratio of the number of black beads to the number of beads. At the second drawing the probabilities depend on what happened during the first drawing; if the first bead actually is black, then there are two white and two black beads left in the bag. These four remaining beads can be assigned equal probabilities by assuming symmetry, so that

$$P(\text{second white given first black}) = \frac{2}{4}.$$

Then,

$$\begin{aligned} P(\text{first black and second white}) &= P(\text{second white given first black})P(\text{first black}) \\ &= \frac{2}{4} \times \frac{3}{5} = \frac{3}{10}. \end{aligned}$$

In analogous fashion one can compute the following probabilities:

$$\begin{aligned} P(\text{first } W \text{ and then } B) &= \frac{2}{5}\times\frac{3}{4}=\frac{3}{10}, \\ P(\text{first } W \text{ and then } W) &= \frac{2}{5}\times\frac{1}{4}=\frac{1}{10}, \\ P(\text{both } B) &= \frac{3}{5}\times\frac{2}{4}=\frac{3}{10}. \end{aligned}$$

Note that the total of these probabilities is, of course, unity; but they are not equal. Notice, incidentally, that

$$P(\text{first white}) = P(\text{first } W \text{ and then } B)+P(\text{first } W \text{and then } W)=\frac{2}{5},$$

and

$$P(\text{first black}) = P(\text{first } B \text{ and then } W)+P(\text{both } B)=\frac{3}{5}.$$

Notice also that

$$P(\text{second white})=P(\text{first } B \text{ and then } W)+P(\text{both } W)=\frac{2}{5}.$$

That is, the probability of a white bead on the second draw, without the knowledge of what happened on the first draw, is the same as though all the beads were still in the bag!

7.2.1 Independent Events

If the occurrence or nonoccurrence of either of the events A and B does not affect the probability of occurrence of the other, then $P(A|B)=P(A)$ and $P(B|A)=P(B)$. Hence

$$\begin{aligned} P(A\cap B) &= P(B)P(A|B) \\ &= P(B)P(A). \end{aligned}$$

We shall use this condition as our formal definition of independence of events.

Definition 7.2.2. Two events A and B are independent if and only if

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

This result can be extendable to the case of k events. That is, if $A_1,A_2,\ldots,A_k$ are independent or *mutually independent* if $P(A_i\cap A_j)=P(A_i)P(A_j)$ for all $i\neq j$,$P(A_i\cap A_j\cap A_k)=P(A_i)P(A_j)P(A_k)$ for all three distinct events and so on, and finally,

$$P(A_1\cap A_2\cap\ldots\cap A_k)=P(A_1)\ldots P(A_k).$$

If the property is satisfied only for all pairs, that is, $P(A_i\cap A_j)=P(A_i)P(A_j)$ for all $i\neq j$ then we say that these events are *pairwise independent*. Pairwise independence need not imply mutual independence.

Theorem 7.2.3. *If the two events A and B are independent, then (a) A and B' are independent (b) A' and B are independent and (c) A' and B' are independent.*

Proof.
(a) We can write $A = (A \cap B') \cup (A \cap B)$ where $A \cap B'$ and $A \cap B$ are mutually exclusive. Hence by Axiom 3, we get

$$\begin{aligned} P(A) &= P(A \cap B') + P(A \cap B) \\ &= P(A \cap B') + P(A)P(B), \text{ since } A \text{ and } B \text{ are independent.} \end{aligned}$$

Hence it follows that

$$P(A \cap B') = P(A) - P(A)P(B) = P(A)[1 - P(B)] = P(A)P(B').$$

That is, A and B' are independent.
Similarly we can establish (b) by interchanging A by B and B by A in the above steps.
(c) Since $A' \cap B' = \overline{A \cup B}$, we have

$$\begin{aligned} P(A' \cap B') &= P(\overline{A \cup B}) \\ &= 1 - P(A \cup B) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \\ &= 1 - P(A) - P(B) + P(A)P(B) \\ &= [1 - P(A)][1 - P(B)] \\ &= P(A')P(B'). \end{aligned}$$

Hence A' and B' are independent.

Theorem 7.2.4.

$$P[(A_1 \cup A_2)|B)] = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B);\ P(B) \neq 0.$$

Proof. By definition

$$P[(A_1 \cup A_2)|B)] = \frac{P[(A_1 \cup A_2) \cap B]}{P(B)};\ P(B) \neq 0.$$

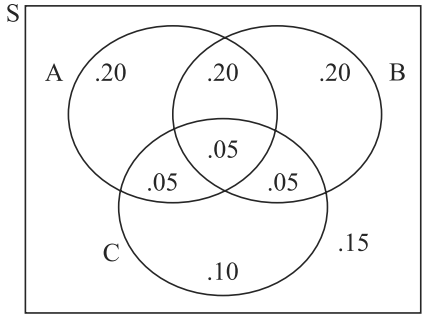
Now

$$\begin{aligned} P[(A_1 \cup A_2) \cap B] &= P[(A_1 \cap B) \cup (A_2 \cap B)] \\ &= P(A_1 \cap B) + P(A_2 \cap B) - P(A_1 \cap A_2 \cap B). \end{aligned}$$

Hence

$$\begin{aligned} P[(A_1 \cup A_2)|B)] &= \frac{P(A_1 \cap B) + P(A_2 \cap B) - P(A_1 \cap A_2 \cap B)}{P(B)} \\ &= P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B). \end{aligned}$$

Remark 7.2.1. It can be shown that for any event A, all of the elementary properties of probability covered in the last chapter, extend to conditional probability. As another example, since we know that $P(A^c) = 1 - P(A)$, it therefore also follows that $P(A^c|B) = 1 - P(A|B)$. However, there is one important exception! We know that if A and B are two independent events, then $P(A \cap B) = P(A)P(B)$. But this does not extend to conditional probabilities! In particular, if C is any other event, then $P(A \cap B|C) \neq P(A|C)P(B|C)$ in general. Venn diagram given below illustrates this, for three events A, B and C:



From the Venn diagram, we have

$$P(A) = P(B) = 0.50 \quad \text{and} \quad P(A \cap B) = 0.25.$$

Also,

$$\begin{aligned} P(A|C) &= \frac{P(A \cap C)}{P(C)} = \frac{0.10}{0.25} = \frac{2}{5}. \\ P(B|C) &= \frac{P(B \cap C)}{P(C)} = \frac{0.10}{0.25} = \frac{2}{5}. \\ P(A \cap B|C) &= \frac{P(A \cap B \cap C)}{P(C)} = \frac{0.05}{0.25} = \frac{1}{5}. \end{aligned}$$

Confirm that $P(A \cap B) = P(A)P(B)$ but $P(A \cap B|C) \neq P(A|C)P(B|C)$. In other words, two events that may be independent in a general population, may not necessarily be independent in a particular subgroup of that population. Note that even if two events are independent they need not be disjoint. Independence is a product probability property and should not be interpreted as one has nothing to do with the other. The events depend on each other through the relation $P(A \cap B) = P(A)P(B)$ but in the conditional space the probability of an event is free of the conditions imposed on the other event.

Mutual Independence of three events

Let A, B and C be any three events with respect to the sample space S of a random experiment. These events are said to be independent if

$$P(A \cap B) = P(A)P(B),\ P(B \cap C) = P(B)P(C),\ P(A \cap C) = P(A)P(C) \ \text{ and}$$

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

For mutual independence, we need all these conditions to hold. If, only the first three relations are satisfied by the events then we say that they are *pairwise independent*. The following example illustrates that pairwise independence does not guarantee mutual independence.

Example 7.2.4. An unbiased coin is tossed twice and define the events:
A: the first toss is head, B: the second toss is head and C: the outcomes of the two tosses are same. Show that the conditions of pairwise independence are satisfied but not of mutual independence.
Here

$$S = \{HH, HT, TH, TT\}, A = \{HH, HT\}, B = \{HH, TH\} \text{ and } C = \{HH, TT\}.$$

Since the coin is unbiased we will assume symmetry in the sample space and assign equal probabilities to the sample points. Then

$$\begin{aligned} A\cap B=\{HH\} \Longrightarrow P(A\cap B)=0.25 &= P(A)P(B), \\ A\cap C=\{HH\} \Longrightarrow P(A\cap C)=0.25 &= P(A)P(C) \text{ and} \\ B\cap C=\{HH\} \Longrightarrow P(B\cap C)=0.25 &= P(B)P(C). \end{aligned}$$

Now $A\cap B\cap C=\{HH\}$ and hence

$$P(A\cap B\cap C)=0.25\neq P(A)P(B)P(C)=0.125.$$

Thus pairwise independence does not guarantee mutual independence.

Example 7.2.5. Tom, Merly, Jose, Manju and Metty decide to go to their favourite restaurant to share some food. They sit down at a round table for five without anyone having preference of any seat, and as soon as they do, Manju notes, "We sat down around the table in age order! What are the odds of that?"

Imagine they sat down in age order, with each person randomly picking a seat. The first person has 5 chairs to select from a total of 5 chairs. The second oldest can sit to his right or left, since these five can sit either clockwise or counterclockwise. The probability of picking a seat next to the first person is thus $\frac{2}{4}$, or $\frac{1}{2}$. The third oldest now has three chairs to choose from, one of which continues the progression in the order determined by the second person, for a probability of $\frac{1}{3}$. This leaves two seats for the fourth oldest, or a $\frac{1}{2}$ chance. The youngest would thus be guaranteed to sit in the right seat, since there is only one seat left. Thus it happens with probability $1\times\frac{1}{2}\times\frac{1}{3}\times\frac{1}{2}\times 1=\frac{1}{12}$, due to conditional probability.

Example 7.2.6. A radar center consists of two units operating independently. The probability that one of the units detects an incoming missile is 0.99, and the probability that the other unit detects it is 0.95. What is the probability that

(i) both units will detect?

(ii) at least one will detect?

(iii) neither will detect?

Let A and B respectively denote the two radar units. Then, given that $P(A)=0.99, P(B)=0.95$.
(i)

$$\begin{aligned} P(\text{both will detect}) &= P(A\cap B) \\ &= P(A)P(B), \text{ since } A \text{ and } B \text{ are independent} \\ &= (0.99)(0.95) \\ &= 0.9405. \end{aligned}$$

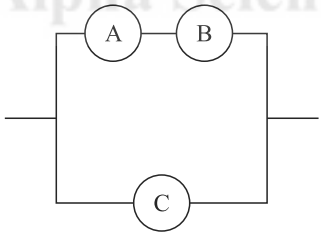
(ii)

$$\begin{aligned} P(\text{at least one will detect}) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A)P(B), \text{ since } A \text{ and } B \text{ are independent} \\ &= 0.99 + 0.95 - (0.99)(0.95) \\ &= 0.9995. \end{aligned}$$

(iii)

$$\begin{aligned} P(\text{neither will detect}) &= P(\overline{A \cup B}) \\ &= 1 - P(A \cup B) \\ &= 1 - 0.9995 \\ &= 0.0005. \end{aligned}$$

Example 7.2.7. The circuit of an electrical equipment is shown in the diagram. It works so long as current can flow from left to right. The three components work in an independent manner. The probability that component A works is 0.8; the probability that component B works is 0.9; and the probability that component C works is 0.75. Find the probability that the equipment works. Let A, B and C respectively represent the events 'component A works', 'component B works', and 'component C works'. Now the equipment will work if either A and B are working, or C is working. Thus the event we are interested in is $(A \cap B) \cup C$.



Now

$$\begin{aligned} P((A \cap B) \cup C) &= P(A \cap B) + P(C) - P(A \cap B \cap C), \text{ by addition theorem} \\ &= P(A)P(B) + P(C) - P(A)P(B)P(C), \text{ by mutual independence} \\ &= (0.8)(0.9) + (0.75) - (0.8)(0.9)(0.75) \\ &= 0.93. \end{aligned}$$

Are mutually exclusive events independent?

It is not uncommon for people to confuse the concepts of mutually exclusive events and independent events. Consider the events "getting three in the first throw" and "getting five in the second throw" of a balanced die. The probability of getting a three in the first throw and getting five in the second throw is the result of getting $(3,5)$ when two balanced dice are rolled together.

$$P[(3,5)] = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = P[3] \times P[5].$$

The two events are independent, since whatever happens to the first die cannot affect the throw of the second, the probabilities are therefore multiplied. Now consider the events "getting three" and "getting five" when a balanced die is rolled once. If '3' turns up then '5' cannot occur and vice versa. Thus these two events are mutually exclusive. When calculating the probabilities for exclusive events you add the probabilities. When calculating the probabilities for independent events you multiply the probabilities. So, if A and B are mutually exclusive, they cannot be independent. If A and B are independent, they cannot be mutually exclusive.

7.2.2 Independent Trials

If the outcomes of a particular random experiment have no influence on the outcomes of another trial or random experiment then we say that the trials are independent. Events corresponding to independent trials are always independent. If $E_1, E_2, \ldots, E_r$ are events corresponding to the independent trials $1, 2, \ldots, r$ respectively, then

$$P(E_1 \cap E_2 \cap \ldots \cap E_r) = P(E_1)P(E_2)\ldots P(E_r).$$

Example 7.2.8. A fair coin and a balanced die are tossed together. What is the probability of getting a tail and the face 5?

Let A denote the event of getting a tail and B denote the event of getting the face 5. Obviously A and B are events of two independent component experiments. Hence

$$\text{Required probability } = P(A \cap B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{6}.$$

If we compute this probability without using independence property, then we will proceed as follows:
Here we have the sample space S

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

so that $n(S) = 12$ and $n(A \cap B) = 1$. Thus

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{12}.$$

Example 7.2.9. One urn contains 6 white and 10 black balls. A second urn contains 8 white and 12 black balls. One ball is randomly drawn from each of the urns. The balls in each urn are identical except for colours. What is the probability that the balls drawn are both white?

The probability of a white ball from the first urn is $\frac{6}{16}$ and from the second urn is $\frac{8}{20}$. Now the colour of the ball drawn from the second urn in no way depends upon the colour of the ball drawn from the first urn. Hence the two events are independent and the required probability is the product $\frac{6}{16} \times \frac{8}{20} = \frac{3}{20}$.

Example 7.2.10. A fair coin is tossed until a head is obtained. What is the probability that the coin has to be tossed at least four times?

$$\begin{aligned} P(\text{tossing the coin at least four times}) &= P(TTTH \text{or} TTTTH\ldots) \\ &= P(TTTH)+P(TTTTH)+\cdots \\ &= P(T)P(T)P(T)P(H)+\cdots \\ &= \left(\frac{1}{2}\right)^4+\left(\frac{1}{2}\right)^5+\cdots \\ &= \left(\frac{1}{2}\right)^4\left(\frac{1}{1-\frac{1}{2}}\right) \\ &= \frac{1}{8}. \end{aligned}$$

Remark 7.2.2. Consider a random experiment having only two outcomes- one may be called SUCCESS and the other FAILURE. Let p and $1-p$ be the respective probabilities of a SUCCESS and a FAILURE. Such a random experiment is called a *Bernoulli trial.* The following are some examples of Bernoulli trial:

(i) Tossing a coin and observing its face.

(ii) Rolling a die and observing whether the number is greater than 4 or not.

(iii) Noting whether the next person coming to a shop is a male or female.

(iv) Selecting a student at random from a class and noting whether his weight is less than 50 kilograms or not.

(v) The next respondent of a particular study is a smoker or not.

(vi) The time one has to wait for getting service from a reservation counter is less than 15 minutes or not.

Example 7.2.11. A fair coin is tossed independently twice. Here, getting head may be labeled as success and $p=\frac{1}{2}$. Then we have the following probabilities:

$$\begin{aligned} P(\text{both tail}) &= P(FF)=\frac{1}{2}.\frac{1}{2}=\left(\frac{1}{2}\right)^2 \\ P(\text{one head}) &= P(FS \text{ or } SF)=P(FS)+P(SF)=\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}=\binom{2}{1}\left(\frac{1}{2}\right)^2 \\ P(\text{both head}) &= P(SS)=\frac{1}{2}\times\frac{1}{2}=\left(\frac{1}{2}\right)^2. \end{aligned}$$

Similarly, if the coin is tossed independently thrice, then

$$\begin{aligned} P(\text{one head}) &= P(FFS \text{ or } FSF \text{ or } SFF) \\ &= P(FFS)+P(FSF)+P(SFF) \\ &= \left(\frac{1}{2}\right)^3+\left(\frac{1}{2}\right)^3+\left(\frac{1}{2}\right)^3 \\ &= \binom{3}{1}\left(\frac{1}{2}\right)^3. \end{aligned}$$

Now generalizing the above computation of probabilities of events associated with Bernoulli trials one can obtain the following:

Remark 7.2.3. If a Bernoulli trial is conducted independently n times where p remains the same from trial to trial, then the probability of getting exactly x successes is:

$$\binom{n}{x} p^x (1-p)^{n-x} \tag{7.2.1}$$

where x may be any of the values $0, 1, \ldots, n$. The function (7.2.1) is called binomial probability function. Using (7.2.1) the answer in Example 6.4.1 is $\binom{5}{3}\left(\frac{1}{2}\right)^3(1-\frac{1}{2})^{5-3} = \frac{5}{16}$.
If a Bernoulli trial is conducted independently until first success occurs, then the probability of getting first success at the rth trial is given by

$$P(\text{(r-1) failures before the first success}) = p(1-p)^{r-1} \tag{7.2.2}$$

where $r = 1, 2, \ldots$. The function (7.2.2) is called geometric probability function. Using (7.2.2) the answer in Example 7.2.10 can be given as

$$\sum_{r=4}^{\infty} \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{r-1} = \frac{1}{8}.$$

Similarly, the probability of k failures before the first success is $p(1-p)^k,\ k = 0, 1, \ldots$.

Example 7.2.12. A box contains 6 gold coins, 4 silver coins and 3 bronze coins. If three coins are taken together at random, what is the probability that they are all of different material? What is the probability that they are all of the same material?

In this case the sampling is done without replacement and the sample is unordered. So $n(S) = \binom{13}{3} = 286$. The event that the three coins are all of different material can occur in $6 \times 4 \times 3 = 72$ ways, since we must have one of the six gold coins, and so on. So the probability is $72/286$. The event that the three coins are of the same material can occur in $\binom{6}{3} + \binom{4}{3} + \binom{3}{3} = 20 + 4 + 1 = 25$ ways, and the probability is $25/286$.

Example 7.2.13. Two people alternately roll an unbiased die until one of them gets a 5. What is the probability that the player making the first toss wins the game?

If the face 5 turns up then it can be considered as a success and failure if any other face comes up. So it is a Bernoulli trial with probability of success $\frac{1}{6}$. Hence

$$P(\text{the player making the first toss wins}) = P(\text{getting '5' in the 1st toss or 3rd,...}).$$

Now using the formula (7.2.2),

$$\begin{aligned} \text{required probability} &= \frac{1}{6}\left(\frac{5}{6}\right)^{1-1} + \frac{1}{6}\left(\frac{5}{6}\right)^{3-1} + \cdots \\ &= \frac{2}{3}. \end{aligned}$$

7.3 THEOREM OF TOTAL PROBABILITY AND BAYES' FORMULA

A plot of land has been divided into five subplots of unequal area denoted by A_1, A_2, A_3, A_4 and A_5. Each of the subplots contains part of a rock. The owner of the land wishes to know what percentage of the plot is occupied by the rock, so he asks each of the individual tenants to determine the percentage of land that is occupied by the rock. But since the five subplots have different sizes, the owner cannot simply average the five percentages obtained from his tenants. Rather he must weight these percentages by the percentage of the entire plot occupied by each tenant's subplot. If we let $P(A_i)$ denote the percentage of the entire plot occupied by subplot A_i and $P(R|A_i)$ the percentage of subplot A_i covered by the rock, then the percentage $P(R)$ of the entire plot that is covered by the rock is

$$P(R) = P(A_1)P(R_1) + P(A_2)P(R_2) + P(A_3)P(R_3) + P(A_4)P(R_4) + P(A_5)P(R_5).$$

In the above, $P(A_i)$ can be viewed as probabilities of the partition parts and $P(R|A_i)$ as conditional probabilities. In a similar fashion the probability of an event can be calculated by finding the conditional probabilities of the event relative to the parts of some partition of the sample space S and then taking a weighted average of these probabilities, the weights being the probabilities of the parts of the partition.

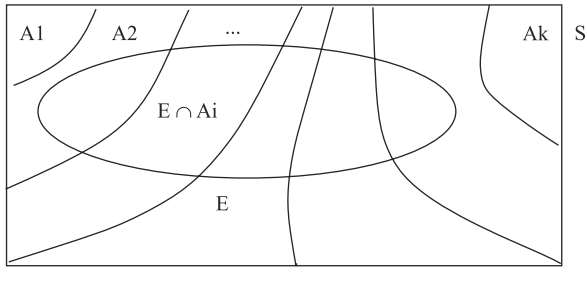
Definition 7.3.1. The events $A_1, A_2, \ldots, A_k$ constitute a partition of a sample space S if
(i) $A_i \cap A_j = \emptyset$ for all $i \neq j$,
(ii) $\bigcup_{i=1}^{k} A_i = S$ and
(iii) $P(A_i) > 0$, for $i = 1, 2, \ldots, k$.

Now we shall state the *total probability theorem* (or *rule*).

Theorem 7.3.1. *If $E \subset S$ and the events $A_1, A_2, \ldots, A_k$ form a partition of a sample space S, then*

$$P(E) = P(A_1)P(E|A_1) + P(A_2)P(E|A_2) + \cdots + P(A_k)P(E|A_k).$$

Figure 7.1: Partitioning event E into k mutually exclusive events



Proof. Since A_i's constitute a partition of S, $A_i \cap A_j = \emptyset$ for all $i, j = 1, 2, \ldots, k, i \neq j$. Hence we can express the event E as a union of k disjoint events (Figure 7.1) as follows:

$$E = (E \cap A_1) \cup (E \cap A_2) \cdots (E \cap A_k).$$

Thus

$$\begin{aligned} P(E) &= P(E\cap A_1)+P(E\cap A_2)+\ldots+P(E\cap A_k) \text{ by Axiom 3} \\ &= P(A_1)P(E|A_1)+\cdots+P(A_k)P(E|A_k), \text{ using multiplication theorem.} \end{aligned}$$

Example 7.3.1. Suppose the probability that a postgraduate student in Statistics prepares for Indian Statistical Service (I.S.S.) examination is 0.40. Given that he prepares, the probability is 0.80 that he will pass the examination. Given that he does not prepare, the probability is 0.30 that he will pass the examination. What is the probability that he will pass I.S.S. examination?

Let A denotes the event that the student will pass I.S.S. examination and B be the event that the student prepares for I.S.S. examination. Then A can be written as

$$A = (A\cap B)\cup(A\cap B').$$

Notice that $(A\cap B)$ and $(A\cap B')$ are mutually exclusive. Therefore

$$\begin{aligned} P(A) &= P(A\cap B)+P(A\cap B') \\ &= P(A|B)P(B)+P(A|B')P(B') \\ &= (0.80)(0.40)+(0.30)(0.60) \\ &= 0.50. \end{aligned}$$

Example 7.3.2. Suppose that in a college 35 percent of the faculty are seniors, 40 percent are juniors, 15 are freshmen and the remaining are guest lecturers. Assume that there is no overlap in these categories. If we are told that 10 percent of the seniors, 50 percent of the juniors, 70 percent of the freshmen and 5 percent of the guest lecturers are doctoral degree holders determine the probability that a faculty member chosen at random has doctoral degree.

Here the following partition of the sample space occurs:
A_1: the event of getting a senior faculty A_2: the event of getting a junior faculty
A_3: the event of getting a freshman A_4: the event of getting a guest lecturer.
Given that $P(A_1)=0.35$, $P(A_2)=0.40$, $P(A_3)=0.15$ and $P(A_4)=0.10$. If E is the event "faculty member has doctoral degree" then we are also given that $P(E|A_1)=0.10$, $P(E|A_2)=0.50$, $P(E|A_3)=0.70$ and $P(E|A_4)=0.05$. By the total probability law, we have

$$P(E)=(0.35\times0.10)+(0.40\times0.50)+(0.15\times0.70)+(0.10\times0.05)=0.345.$$

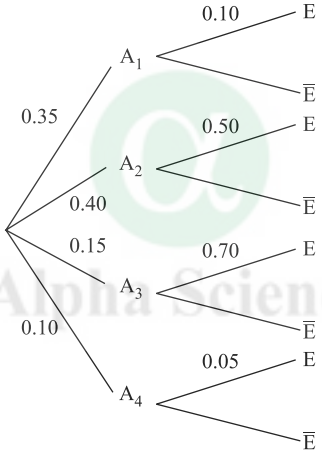
From Figure 7.2, notice that the four terms in the sum for $P(E)$ are the probabilities along the four branches of the tree that lead to E.

In the partition theorem we have obtained the probability for the 'effect' E given the various 'causes' $A_1,A_2,\ldots,A_k$. But in some situations we know the probability of the effect, for various possible causes (the partition), and we are interested in determining how likely each cause is, knowing the effect to have occurred. Now consider the following situations:

1. Suppose that three machines producing the same item, each producing a certain percentage of defective items. If an item picked at random from the production unit is found to be defective then what is the probability that this item is produced by a particular machine?

2. A doctor has noticed certain symptoms E in a patient and the doctor wants to determine its cause C_i (the disease which is causing these symptoms). Note that there may be several diseases $C_1, \ldots, C_k$ with the same symptoms. Usually the doctor will have information regarding $P(E|C_i)$, how likely these symptoms are, given that the patient has a certain disease C_i. But what the doctor really wants to know is the reverse probability $P(C_i|E)$, how likely it is that a certain disease which is causing these symptoms to appear.

Figure 7.2: Tree diagram



Thomas Bayes (1702-1761), an English presbyterian minister and mathematician, examined these types of problems. His works on probability theory were published in a posthumous scientific paper in 1764. Now let us consider the following example.

Thomas Bayes (1702-1761)

Example 7.3.3. Suppose that four classes in a college contain Christian, Hindu and Muslim students in the numbers given below:

Class	Christian	Hindu	Muslim	Total
1	8	6	2	16
2	5	4	3	12
3	7	4	1	12
4	6	6	4	16

A class is picked at random and a student is randomly chosen from that class.

(a) Find the probability that the selected student is a Muslim.

(b) Given that the student is Muslim, what is the probability that the student came from class 2?

Let M be the event that the selected student is a Muslim and C_i denotes the event that the selected student belongs to class i, $i = 1,2,3,4$. (a) Applying the partition theorem we can write

$$P(M) = P(M|C_1)P(C_1) + P(M|C_2)P(C_2) + P(M|C_3)P(C_3) + P(M|C_4)P(C_4).$$

Now, since a class is picked at random, $P(C_1) = P(C_2) = P(C_3) = P(C_4) = \frac{1}{4}$. Also, it is easily seen that $P(M|C_1) = \frac{2}{16}$, $P(M|C_2) = \frac{3}{12}$, $P(M|C_3) = \frac{1}{12}$ and $P(M|C_4) = \frac{4}{16}$. Therefore,

$$\begin{aligned} P(M) &= \frac{2}{16} \times \frac{1}{4} + \frac{3}{12} \times \frac{1}{4} + \frac{2}{12} \times \frac{1}{4} + \frac{4}{16} \times \frac{1}{4} \\ &= \frac{17}{96}. \end{aligned}$$

(b) Here we require $P(C_2|M)$. Using the formula for conditional probability, this is given by

$$P(C_2|M) = \frac{P(C_2 \cap M)}{P(M)}.$$

But we know that $P(C_2 \cap M) = P(M|C_2)P(C_2)$. Hence

$$P(C_2|M) = \frac{P(M|C_2)P(C_2)}{P(M)}.$$

$$\begin{aligned} P(C_2|M) &= \frac{P(M|C_2)P(C_2)}{P(M|C_1)P(C_1) + P(M|C_2)P(C_2) + P(M|C_3)P(C_3) + P(M|C_4)P(C_4} \\ &= \frac{P(M|C_2)P(C_2)}{P(M)} \\ &= \frac{3/48}{17/96} \\ &= \frac{6}{17}. \end{aligned}$$

A generalization of the result in (b) of the above example is called *Bayes' formula* (or *rule*, or also *theorem*) and it is given in the following theorem:

Theorem 7.3.2. *If $A_1, A_2, \ldots, A_k$ is a partition of a sample space S and $P(A_i) \neq 0$ for $i = 1,2,\ldots,k$ then for any event E with respect to S such that $P(E) \neq 0$,*

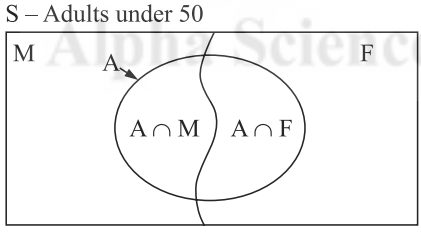
$$P(A_i|E) = \frac{P(A_i)P(E|A_i)}{P(A_1)P(E|A_1) + P(A_2)P(E|A_2) + \cdots + P(A_k)P(E|A_k)}.$$

for $i = 1,2,\ldots,k$.

Bayes' formula is the starting point for an entire statistical philosophy known as *Bayesian statistics*. The computations involved in Bayes' formula can be made simple by drawing a tree diagram as in Figure 7.2. If the event E has occurred, then only the branches leading to E remain relevant. These branches then constitute our conditional sample space. If we are calculating $P(A_i|E)$, then we are interested in the likelihood of the i-th branch leading to E relative to all k branches that lead to E. The numerator in Bayes' formula is the probability along this i-th branch, while the denominator is the sum of probabilities along all branches in our conditional sample space.

Remark 7.3.1. It may happen that the probabilities of the partition events A_i's are unknown and that the very purpose of the experiment is to either approximate or obtain exact values for these probabilities. Initially we assign or guess or estimate the probabilities of the partition events $A_1,\ldots,A_k$ based on a certain 'a priori' idea even before performing the random experiment. Hence the probabilities $P(A_1),\ldots,P(A_k)$ are called the *a priori*, or *prior* probabilities. After the experiment has been run once, or perhaps several times, and we have observed the event E, we will probably wish to revise our estimates for these probabilities and obtain the *a posteriori*, or *posterior* probabilities $P(A_1|E),\ldots,P(A_k|E)$.

Example 7.3.4. Suppose that, for adults under age 50, we are interested in comparing sleep disorders (A) between males(M) and females(F). It is known that 70% of males and 40% of females have sleep disorders. Compute the posterior probabilities. Assume equal number of males and females in the population.



We have the prior probabilities

$$P(M) = P(F) = 0.50$$

and

$$P(A|M) = 0.70, \quad P(A|F) = 0.40.$$

We actually wish to calculate the probability of each gender, given A. That is, the posterior probabilities $P(M|A)$ and $P(F|A)$.

$$\begin{aligned} P(M|A) &= \frac{P(A|M)P(M)}{P(A|M)P(M)+P(A|F)P(F} \\ &= \frac{0.70\times 0.50}{0.70\times 0.50+0.40\times 0.50} \\ &= \frac{0.35}{0.55} \\ &= 0.64. \end{aligned}$$

Similarly,

$$\begin{aligned} P(F|A) &= \frac{P(A|F)P(F)}{P(A|M)P(M)+P(A|F)P(F} \\ &= \frac{0.20}{0.55} \\ &= 0.36. \end{aligned}$$

Thus, the additional information that a randomly selected individual has sleep disorder increases the likelihood of being male from a prior probability of 50% to a posterior probability of 64%, and likewise, decreases the likelihood of being female from a prior probability of 50% to a posterior probability of 36%. That is, knowledge of event A can alter a prior probability $P(B)$ to a posterior probability $P(B|A)$, of some other event B.

Example 7.3.5. Suppose that three identical boxes each contain two coins, Box 1 contains two gold coins, Box 2 contains two silver coins, and Box 3 contains a gold coin and a silver coin. A box is chosen at random and then a coin is taken out at random and is found to be a gold coin. Find the probability that the other coin in the box is also a gold coin.

Let E be the event that the chosen coin was gold and A_i denotes the event that the coin was chosen from Box i, $i = 1,2,3$. With the help of Bayes' formula we have

$$P(A_1|E) = \frac{\frac{1}{3}\times 1}{\frac{1}{3}\times 1+\frac{1}{3}\times 0+\frac{1}{3}\times\frac{1}{2}} = \frac{2}{3}.$$

Hence $P(A_1|E) = \frac{2}{3}$ and not $\frac{1}{2}$ as reasoned earlier. The logic of the answer is quite simple: two of the three gold coins in the three boxes are in Box 1.

Clinical tests are frequently used in medicine and epidemiology to diagnose the presence (T^+) or absence (T^-) of a particular condition, such as pregnancy or disease. Definitive disease status (either D^+ or D^-) is often subsequently determined by some other costlier means such as surgical procedures or autopsy. Different measures of the test's merit can then be estimated using various conditional probabilities. An example of this kind is given below:

Example 7.3.6. A study was conducted to detect the sensitivity of a clinical test to detect breast cancer. For this, a random sample of 400 females aged above 30 years have undergone the clinical test and subsequent diagnosis to confirm the disease status of each. The data obtained is given below:

	Diseased (D^+)	Nondiseased(D^-)	Total
Test Positive(T^+)	32	18	50
Test Negative(T^-)	8	342	350
Total	40	360	400

Assuming symmetry in the sample space we have,

$$P(T^+|D^+) = \frac{n(T^+\cap D^+)}{n(D^+)} = \frac{32}{40} = 0.80, \quad P(T^+|D^-) = \frac{n(T^+\cap D^-)}{n(D^-)} = \frac{18}{360} = 0.05,$$

$$P(T^-|D^+) = \frac{n(T^- \cap D^+)}{n(D^+)} = \frac{8}{40} = 0.20 \text{ and } P(T^-|D^-) = \frac{n(T^- \cap D^-)}{n(D^-)} = \frac{342}{360} = 0.95.$$

Also,

$$P(D^+) = \frac{40}{400} = 0.10 \qquad \text{and} \qquad P(D^-) = \frac{360}{400} = 0.90.$$

A physician may be interested in the probabilities $P(D^+|T^+)$ and $P(D^-|T^-)$. Using Bayes' formula,

$$\begin{aligned} P(D^+|T^+) &= \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \\ &= \frac{0.80 \times 0.10}{0.80 \times 0.10 + 0.05 \times 0.90} \\ &= 0.64. \end{aligned}$$

Similarly,

$$\begin{aligned} P(D^-|T^-) &= \frac{P(T^-|D^-)P(D^-)}{P(T^-|D^-)P(D^-) + P(T^-|D^+)P(D^+)} \\ &= \frac{0.95 \times 0.90}{0.95 \times 0.90 + 0.20 \times 0.10} \\ &= 0.977. \end{aligned}$$

Usually, $P(T^+|D^+)$ is known as *sensitivity* or 'true positive rate' of the test and $P(T^-|D^-)$ is known as the *specificity* or 'true negative rate' of the test. Here we have estimated the prior probabilities, namely $P(D^+)$ and $P(D^-)$, from the sample information. Instead, one can use the values from outside published sources and references for a more accurate estimate of prior probabilities. $P(D^+|T^+)$ and $P(D^-|T^-)$ are the posterior probabilities. From the above, a positive test result increases the probability of having this disease from 10% to 64%; a negative test result increases the probability of not having the disease from 90% to 97.7%.

7.4 EXERCISES

1. Two events A and B are such that if B occurs, the probability of A is unchanged. The events are said to be
   (a) Mutually exclusive; (b) Exhaustive; (c) Independent; (d) Equally likely.

2. If you and two friends all randomly answered a multiple choice question that had 4 choices, what is the chance that your two friends both got it right and you got it wrong?
   (a) $\frac{1}{16}$ (b) $\frac{1}{64}$ (c) $\frac{3}{16}$ (d) $\frac{3}{64}$.

3. There is 80% chance that a problem will be solved by a statistics student and 60% chance is there that the same problem will be solved by the mathematics student. The probability that at least one of them will solve the problem is:
   (a) 0.48 (b) 0.92 (c) 0.10 (d) 0.75.

4. We are told that in a random experiment there are five possible outcomes. Which of the following statements is true ?

(a) If, after 20 trials, one outcome has not been observed then the probability that it will occur in the next trial is increased.

(b) If, after 20 trials, one outcome has been observed more often than the others then the probability that it will not occur in the next trial is increased.

(c) If, after 20 trials, one outcome has not been observed then the probability that it will occur in the next trial is unchanged.

(d) If the outcomes are equally likely then the trials are independent.

The students in a college were surveyed. Each student was asked whether he or she used a helmet during bike journey. This information and the gender of the student was recorded as follows:

Gender	Use helmet?		Total
	Yes	No	
Male	69	31	100
Female	82	23	105
Total	151	54	205

Assume symmetry in the sample space and answer the next three questions.

5. What is the probability that a randomly selected student is a male who does not use his helmet?
(a) $\frac{23}{105}$ (b) $\frac{31}{100}$ (c) $\frac{31}{205}$ (d) $\frac{54}{205}$.

6. What is the probability that a randomly selected student is a female given that the person is a helmet user?
(a) $\frac{82}{105}$ (b) $\frac{82}{205}$ (c) $\frac{82}{151}$ (d) $\frac{151}{205}$.

7. Is the use of helmet independent of gender?

   (a) No, because the probability that a randomly chosen student is a female does not equal the probability of female given helmet use;

   (b) No, because the number of females who use a safety belt is not equal to the number of males who use a helmet;

   (c) Yes, because the sample was randomly selected;

   (d) Yes, because both genders do use helmets more than they do not use helmets.

Suppose that 1% of a population has a particular disease. A new test for identifying the disease has been developed. If the person has the disease, the test is positive 94% of the time. If the person does not have the disease, the test is positive 2% of the time. Assume symmetry in the sample space and answer the next two questions.

8. What is the probability that a randomly selected person from this population tests positive?
(a) 0.0292 (b) 0.0094 (c) 0.096 (d) 0.96.

9. A person is randomly selected from this population and tested. He tests positive. Which of the following best represents the probability that he has the disease?
(a) 0.0094 (b) 0.32 (c) 0.34 (d) 0.94.

10. If A and B are two independent events with $P(A) = 0.3$ and $P(B) = 0.6$, then find $P(A \text{ and } B)$ and $P(A \text{ or } B)$.

11. If A and B are two independent events with $P(A) = 0.5$ and $P(A \text{ and } B) = 0.2$, then find (i) $P(B)$ and (ii) $P(A \text{ or } B)$.

12. If A and B are two events for which $P(A) = 0.2, P(B) = 0.8$ and $P(A \text{ or } B) = 0.84$. (i) Are A and B mutually exclusive? (ii) Are A and B independent?

13. Are mutually exclusive events independent?

14. Let events A, B, C be mutually independent. Then show that A and $B \cap C$ are independent, and A and $B \cup C$ are independent.

15. If A is a sure event and B is any other event, then show that A and B are independent.

16. If A is a null event and B is any other event, then show that A and B are independent.

17. Assume that each child who is born is equally likely to be a boy or girl. If a family has two children, what is the probability that both are girls given that
(i) the eldest is a girl; (ii) at least one is a girl.

18. Two balanced dice are rolled. What is the probability that at least one is a six? If the two faces are different, what is the probability that at least one is a six?

19. Raj and Das shoot at a target at the same time. Suppose Raj hits the target with probability 0.7, whereas Das independently, hits the target with probability 0.4.
(i) Given that exactly one shot hit the target, what is the probability that it was Das's shot?
(ii) Given that the target is hit, what is the probability that Das hit it?

20. Consider a random experiment of rolling a balanced die until it shows a 6. Calculate the probability that the event "more than six tosses are required before the first 6 occurs".

21. Two fair dice are rolled until sum 12 occurs. What is the probability that at most 9 rolls will be required?

22. In a certain community it is found that the sex ratio among children is 3 girls to 2 boys. If a family of 5 children is picked at random, calculate the probability that

    (i) the eldest is a boy and the rest are all girls.

    (ii) the first three children are of one sex and the rest of other sex.

23. Suppose that the probability that a train arrives at the railway station before 9.00 am is 0.7. A passenger takes a bus that arrives at the railway station by 9.00 am with a probability of 0.6. Find the probability that

(i) the train and the passenger both arrive by 9.00 am.

(ii) the train arrives by 9.00 am and the passenger arrives after 9.00 am.

24. In a population the proportion of persons of A, B, O and AB blood types are, respectively, 0.40, 0.20, 0.30, and 0.10. If two persons from this population marry each other, find the probability that

    (i) they are both of blood type A.

    (ii) neither is blood type B.

    (iii) one is blood type AB and the other is of blood type O.

    (iv) they are both of same blood type.

    (v) they are both of different blood type.

25. In a population it is known that 15 percent of the people have cardiac problems and 3 percent of the people have cancer. Assuming that incidence of heart disease and cancer are independent, what is the probability that a randomly chosen person: (i) has both ailments? (ii) does not have either ailments?

26. Suppose that electricity flows through two fuses A and B connected to a system. When an extra voltage comes fuse A will break with probability 0.8 and B will break with probability 0.9. Assuming that the fuses function independently, find the probability that the system is safe after a lightning when (i) the fuses are connected in series (ii) the fuses are connected in parallel.

27. A certain medical syndrome is usually associated with two overlapping sets of symptoms, A and B. Suppose it is known that:
    If B occurs, then A occurs with probability 0.80.
    If A occurs, then B occurs with probability 0.90.
    If A does not occur, then B does not occur with probability 0.85.
    Find the probability that A does not occur if B does not occur.

28. The progression of a certain disease is typically characterized by the onset of up to three distinct symptoms, with the following properties:
    Each symptom occurs with probability 0.60.
    If a single symptom occurs, there is a 0.45 probability that the two other symptoms will also occur.
    If any two symptoms occur, there is a 0.75 probability that the remaining symptom will also occur.
    Answer each of the following.

    (i) What is the probability that all three symptoms will occur?

    (ii) What is the probability that at least two symptoms occur?

    (iii) What is the probability that exactly two symptoms occur?

    (iv) Is the event that a symptom occurs statistically independent of the event that any other symptom occurs?

29. A family has $n(n > 1)$ children. Let A be the event that the family has at most one girl and B be the event that not every child is of the same sex. Determine the value of n for which A and B are independent events. Assume P(girl)=P(boy)=0.5.

30. An experiment succeeds twice as often as it fails. What is the probability that in the next five trials there will be four successes?

31. A fair die is thrown and the result is recorded. The same die is thrown a second time. Calculate the probability that the number obtained on the second toss exceeds the number obtained on the first toss.

32. The probability that the propulsion system of a missile functions properly is 0.90, and the probability that its guidance system functions properly is 0.75. If the two systems operate independently, what is the probability that a launch is successful?

33. Coin A is weighted in such a way that heads are three times as likely to occur as tails. Another coin B is weighted in such a way that tails are three times as likely to occur as heads. If both coins are tossed once, what is the probability that both coins come up heads? What is the probability that at least one comes up heads?

34. Suppose that A and B are independent and B and C are mutually exclusive. Are A and C necessarily mutually exclusive? If $P(A) = 0.2,\ P(B) = 0.5,\ P(C) = 0.3$ find the probability that

    (i) all the three events occur.

    (ii) B occurs given that A occurs.

    (iii) B occurs given that C occurs

35. In the Kerala legislative assembly, 80 percent of the legislators are from rural areas and 10 percent are under 40 years of age. If a legislator is chosen randomly, can we calculate the probability of choosing a legislator under 40 years of age from a rural area by (a) multiplying 0.80 and 0.10; (b) adding 0.80 and 0.10? Why or why not?

36. Box 1 contains 1000 transistors, of which 100 are defective and Box 2 contains 2000 transistors, of which 100 are also defective. A box is taken at random and two transistors are drawn from it, at random and without replacement.
    (a) Calculate the probability that both transistors are defective.
    (b) Given that both transistors are defective, what is the probability that they come from Box 1?

37. In an industrial unit manufacturing a certain item there are 3 machines: 60 percent of the items are produced by machine 1, 30 percent by machine 2, and 10 percent by machine 3. The past statistics show that the percentage of defective items produced by these machines are 3 percent, 2 percent and 1 percent respectively. If an item supplied to a customer is a defective one, what is the probability that it came from machine 2?

38. An observational study investigates the connection between aspirin use and three vascular conditions gastrointestinal bleeding, primary stroke, and cardiovascular disease using a group of patients exhibiting these disjoint conditions with the following prior probabilities: $P(\text{GI bleeding}) = 0.2$, $P(\text{Stroke}) = 0.3$, and $P(\text{CVD}) = 0.5$, as well as with the following conditional probabilities: $P(\text{Aspirin} \mid \text{GI bleeding}) = 0.09$, $P(\text{Aspirin} \mid \text{Stroke}) = 0.04$, and $P(\text{Aspirin} \mid \text{CVD}) = 0.02$.

(i) Calculate the following posterior probabilities:
$P(\text{GI bleeding} \mid \text{Aspirin})$, $P(\text{Stroke} \mid \text{Aspirin})$, and $P(\text{CVD} \mid \text{Aspirin})$.

(ii) Compare the prior probability of each category with its corresponding posterior probability. What conclusions can you draw?

39. A man takes part in a game show. At the end, he is presented with three boxes numbered 1, 2 and 3 and is asked to choose one among them. The grand prize is hidden, at random, in one of the boxes, while there is nothing inside the other two boxes. The game show host knows where the grand prize has been hidden. Suppose that the man has chosen Box 1 and that the host tells him that he did well in not choosing Box 3, because there was nothing inside it. He then offers the man the opportunity to change his choice and, therefore, to select Box 2 instead. What is the probability that the man will win the grand prize if he decides to stick with Box 1?

40. On the basis of a retrospective study, it is determined (from hospital records, tumor registries, and death certificates) that the overall five-year survival (event S) of a particular form of cancer in a population has a prior probability of $P(S) = 0.4$. Furthermore, the conditional probability of having received a certain treatment (event T) among the survivors is given by $P(T|S) = 0.8$, while the conditional probability of treatment among the non-survivors is only $P(T|S^c) = 0.3$. A cancer patient is uncertain about whether or not to undergo this treatment, and consults with her oncologist, who is familiar with this study. Compare the prior probability of overall survival given above with each of the following posterior probabilities, and interpret it.

(i) Survival among treated individuals, $P(S|T)$.

(ii) Survival among untreated individuals, $P(S|T^c)$.

41. The following data is taken from a study investigating the use of a technique called radionuclide ventriculography as a diagnostic test for detecting coronary artery disease [Source: Radiology, Volume 167, May 1988, 565-569].

	Diseased	Nondiseased	Total
Test Positive	302	80	382
Test Negative	179	372	551
Total	481	452	933

Assume symmetry in the sample space.

(i) Calculate the sensitivity and specificity of radionuclide ventriculography in this study.

(ii) For a population in which the prevalence of coronary artery disease is 0.10, calculate the predictive power of a positive test and the predictive power of a negative test, using radionuclide ventriculography.

42. State and prove multiplication theorem for the case of k events.

Chapter 8

INDEX NUMBERS

"Planning is the order of the day and without statistics planning is inconceivable"-L.H.C.Tippet.

8.1 INTRODUCTION

A lack of quantitative skills leaves you vulnerable to misinterpretation and wrong decision-making. Are your numerical skills good enough to protect you from spurious arguments made by politicians, the media and business men?The importance of rates, percentages and ratios in every sphere of human activity is known to us. They are mere numbers which convey a lot of information. In economics and business analysis ratios are used as a measuring tool for evaluating the performance of firm's financial operations.

The relationship that two similar variables bear to each other is termed as *ratio*. A widely used ratio is the sex ratio.The relationship of the number of females to the number of males in the population is given by the *sex ratio*, which states the number of females per 1000 males.

The ratio, per hundred, is generally referred to as a *percentage*. Thus percentages are merely special cases of the more general concept of ratios. When dealing with a small number of cases, the use of percentages alone leads to wrong conclusions. For example, consider the statement: $33\frac{1}{3}$ percent of the women students in a particular course are married. Of course the number of women students is an important information. This may be due to the fact that there were only three women students and one of them was married.

The term *rate* is usually used to express the amount or quantity of one variable considered in relation to one unit of a different variable. For example, the rate of speed of 60 kilometers per hour expresses the distance traveled in one unit of time. Some frequently used rates are birth rate, death rate, infant mortality rate and literacy rate. Crude *birth rates* are usually calculated by dividing the births during a year by the mid-year population for that year. A comparison of birth rates in relation to total population is not thoroughly satisfactory, since the proportion of people capable of reproduction may differ from time to time or place to place. It is called "crude" because it does not

take into account age or sex differences among the population. The crude death rate of a region for a given year is obtained by dividing the number of deaths occurring in that region during the year by the mid-year population of the region, and expressing the result per thousand. *Death rates* for specific classes of population (various age groups, males and females), and for specific diseases or causes are also very common. An intelligent comparison of the death rates of different categories involves consideration of the fact that the proportions of the components of the categories may differ. Crude birth rates of more than 30 per 1000 are considered high and rates of less than 18 per 1000 are considered low. Crude death rates of below ten are considered low while crude death rates above 20 per 1000 are considered high. *Infant mortality rate* is defined as the number of deaths of infants (one year of age or younger) per 1000 live births. It is used to compare the health and well-being of populations across and within countries.

We often study price movements in order to discover their cause or effect on the economy. Every economic factor whether it be price, value of money, production, sales or profits changes from time to time. So, we have to deal with the average of changes in a group of related variables, that may relate to periods of time or between places. Our administrators have to deal with the following type of questions:

1. How far the shortage of sand affected the construction sector?
2. Is the farmer getting fair price for his product?
3. Have the expenses in rubber plantation risen this year as compared with previous years?
4. To what extent the rise in the petroleum prices has affected the increase in the prices of food commodities?
5. Suppose a person purchased a "basket" of goods or services in the year 2002 and how much amount he required today to purchase the same "basket" of goods or services?

These are some of the many questions that the policy makers have to address on. These questions cannot be answered in terms of percentage or rate or ratio. Now index numbers are the most widely used statistical device for estimating trends in prices, wages, production and other economic aspects. The index numbers were first introduced in Italy in the year 1764. Though originally developed to measure the effect of changes in prices, there is hardly any field nowadays where index numbers are not used.

Index number is a statistical device for measuring changes or differences in magnitudes in a variable or a group of related variables over time. Index number is a combination of two words namely index meaning an indicator and number meaning a numerical value. Since it works in a similar way to percentage it makes such changes easier to compare. It works in the following way. Suppose that a cup of tea in a particular cafe cost ₹4 in 2005. In 2012, an identical cup of tea cost ₹6. How has the price changed between 2005 and 2012? The particular time period of 2005 which we've chosen to compare against, is called the *base period*. The variable for that period, in this case the ₹4, is then given a value of 100, corresponding to 100%. For comparison, it is

convenient to designate some year as base for which the index number is 100. The index can then be calculated for the later period of 2012 as a proportionate change as follows:

$$\text{Index number} = \frac{6}{4} \times 100 = 150.$$

The index number shows us that there has been a price increase of 50% since the base period. An index number for a single price change like this is called a *price relative*.

It is customary to compare changes in the price level with changes in bank deposits, quantity of production etc. Sometimes we want to compare the cost of living at different times or in different places. Usually, organizations compile their physical volume of trade, industrial production, stocks of goods, and so forth in order to keep at least of current business conditions. In all of the above mentioned situations an index number is calculated which expresses the relative change in the magnitudes of a variable or a group of variables. Since index numbers reveal the prevailing economic conditions they are called *economic barometers*. Indices can be calculated with any convenient frequency: yearly, monthly, daily, etc. Examples of each are Gross National Product, unemployment figures, stock market prices, etc.

Index values are measured in percentage points since the base value is always 100.

8.2 SIMPLEST TYPES OF INDEX NUMBERS

Price Relative

Price relative for a commodity is defined as the ratio of the price of the commodity in a given period or point of time 'k' to its price in another period or point of time '0', called the base period. For example, if the retail price of one liter of diesel in the year 2000 was ₹ 12.50 and that for the year 2012 was ₹49.50 then the price relative in percentage (of 2012 with respect to 2000) is given by

$$p_{0k} = \frac{p_k}{p_0} \times 100 = \frac{49.50}{12.50} \times 100 = 396\%.$$

That is, a price relative is the price of the current year expressed as a percentage of the price of the base year. The advantage of dealing with price relatives instead of prices is that these ratios do not depend upon the units such as dollar, rupee etc in which prices are expressed. But it should be ensured that the same units are used for both time periods. Since our purpose is to study intrinsic variations, the units (rupees, kilograms) should not affect the value of the index number. This is ensured by the use of price relatives instead of prices.

Quantity Relative

If we are interested in changes in quantity or volume of goods we consider quantity relatives. Quantity relative in percentage of the period 'k' with respect to the base period '0' is

$$q_{0k} = \frac{q_k}{q_0} \times 100.$$

Value Relative

Value relatives are used when we wish to compare changes in the money value of the transaction, consumption or sale in two different periods or points of time. When we multiply price per unit of the commodity with the corresponding quantity of the commodity produced or sold gives the total money values pq of the production or sale. *Value relative* in percentage of the period 'k' with respect to the base period '0' is

$$v_{0k} = \frac{p_k q_k}{p_0 q_0} \times 100.$$

Properties of Relatives

1. Identity property: $P_{kk} = 1$
2. Time reversal property: $P_{0k} \times P_{k0} = 1$

$$P_{0k} \times P_{k0} = \frac{p_k}{p_0} \times \frac{p_0}{p_k} = 1.$$

3. Circular or cyclic property: $P_{ab} \times P_{bc} \times P_{ca} = 1$

$$P_{ab} P_{bc} P_{ca} = \frac{p_b}{p_a} \times \frac{p_c}{p_b} \times \frac{p_a}{p_c} = 1.$$

4.
$$P_{ab} \times P_{bc} \times P_{cd} = P_{ad}. \tag{8.2.1}$$

Link and Chain Relatives

Consider a series of successive periods or points of time, denoted by $t_0,\ t_1,\ t_2, \ldots$ and let $p_0,\ p_1,\ p_2, \ldots$ denote the prices of a commodity at times $t_0,\ t_1,\ t_2, \ldots$ respectively. Then $P_{01} = \frac{p_1}{p_0}$ is the price relative of t_1 relating to the proceeding period t_0, $P_{12} = \frac{p_2}{p_1}$ is the price relative of t_2 relating to the just preceding period t_1 and so on. The series of price relatives, namely, $P_{01},\ P_{12},\ P_{23}, \ldots$ are called *link relatives*.

It follows from (8.2.1) that any price relative can be expressed as the product of link relatives. For example,

$$P_{13} = P_{12} \times P_{23}.$$

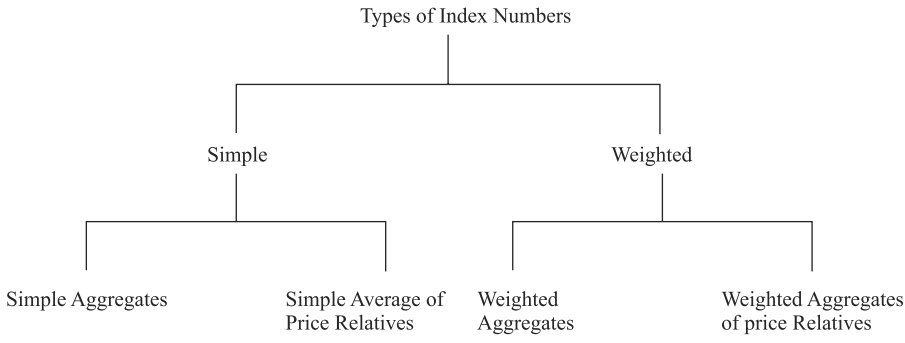
Similarly,

$$P_{0k} = P_{01} \times P_{12} \times \ldots \times P_{k-1k}.$$

That is, the price relative with respect to a fixed base can be obtained by successive multiplication of link relatives. It is therefore called *chain index* or *chain relative*.

8.3 CONSTRUCTION OF INDEX NUMBERS

There are two methods for construction of index numbers (1) by computing aggregate values and (2) by taking average of relatives. Each of these methods have weighted and unweighted formulae.



Simple Aggregates

Suppose that there are n commodities. The prices of each commodity in any given year are merely added together to give the index number for that year. A simple aggregate formula for index number is

$$P_{0k} = \frac{\Sigma p_k}{\Sigma p_0} \times 100,$$

where Σp_k is the sum of the prices of n commodities in the year 'k' and Σp_0 is the sum of the prices of these n commodities in the base year.

Simple Average of Price Relatives

In this method, the price relative of each item is calculated separately and then averaged. The formula for index number when the mean is used is

$$P_{0k} = \frac{1}{n}\Sigma\left(\frac{p_k}{p_0}\right) \times 100,$$

where n is the number of commodities.

When the geometric mean is used,

$$P_{0k} = \left(\prod \frac{p_k}{p_0}\right)^{1/n} \times 100$$

or

$$P_{ok} = \text{Antilog}\left[\frac{1}{n}\Sigma\text{log}\left(\frac{p_k}{p_0}\right)\right] \times 100.$$

Example 8.3.1. Construct index numbers for 2012 taking 2001 as base using simple aggregate and average of price relative methods.

Commodities	Price in 2001(Rs)	Prices in 2012(Rs)
A	70	95
B	40	70
C	120	180
D	30	45
Total	260	390

We shall denote the base year and current year prices as p_0 and p_1 respectively. Then a simple aggregate index number is

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{390}{260} \times 100 = 150.$$

This means that the prices have increased in the year 2012 to the extent of 150% when compared with the prices of 2001. The price relatives computed are given in the following table:

Commodities	Price relatives $P = \frac{p_1}{p_0} \times 100$
A	$\frac{95}{70} \times 100 = 135.7$
B	$\frac{70}{40} \times 100 = 175$
C	$\frac{180}{120} \times 100 = 150$
D	$\frac{45}{30} \times 100 = 150$
Total	510.7

Price index by using simple aggregate average of price relative method is

$$P_{01} = \frac{1}{n}\Sigma\left(\frac{p_1}{p_0}\right) = \frac{510.7}{4} = 127.675.$$

When geometric mean is used,

$$P_{01} = \left(\prod \frac{p_1}{p_0}\right)^{1/n} \times 100 = (5.3432)^{1/4} \times 100 = 152.04.$$

Simple aggregate index add together the prices of all items included in the 'basket of goods', ignoring the quantity bought. It gives equal importance to all items. Hence it is apparent that the commodity having higher price exerts more influence on a simple aggregate index. It fails to consider the actual importance of the different commodities and that the relative influence of different items is determined by factors quite irrelevant to the purpose of the index. A highly important item, say fuel per kilogram is relatively cheap whereas items such as diamonds, are very costly per kilogram. Hence it is not reasonable to express the prices of commodities per kilogram. Furthermore, some items, such as electric power cannot be expressed to a kilogram basis. Still another solution is to take as the unit of quotation as the quantity that can be purchased for one rupee in the base year. But this is also illogical because it would be very unusual to spent the same amount of money on each commodity every year. The relative importance of various commodities is not taken into account in the index as it is unweighted. Thus it is necessary to consider a weighted aggregate rather than a simple aggregate of prices of commodities.

Weighted Aggregates

According to this method, price themselves are weighted by quantities. Thus physical quantities are used as weights. Let p_0 and q_0 respectively denote the base period prices and quantities and p_1 and q_1 be the respective end period prices and quantities. There are various methods of assigning weights and thus various formulae have been formed for the construction of index numbers. Some of the important formulae are given below:

$$\text{Laspeyre's index number, } L = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100.$$

$$\text{Paasche's index number, } P = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100.$$

Dorbish and Bowley have suggested the arithmetic mean of the Laspeyre's price index and Paasche's price index.

$$\text{Dorbish and Bowley index number} = \frac{L+P}{2} \times 100.$$

Professor Irwing Fisher has suggested a compromise between Laspeyre's and Paasche's formula by taking geometric mean of them.

$$\text{Fisher's index number} = \sqrt{L \times P} \times 100.$$

Laspeyre's index number can be misleading in telling us what is actually going on. For example, the fluctuations in the tastes and habits of people might have a considerable impact on an index. Suppose that skirts were considered as a separate item in women's clothing manufacturer's index. The greatly increased relative popularity of trousers would dramatically affect the quantities sold and any index which used base year quantities from some time back would be misleading. However, Paasche's index avoids this particular problem. In Laspeyre's method, the base year quantities are taken as weights and hence this index number has an upward bias. This is due to the fact that when prices increase there is a tendency to reduce the consumption of higher priced goods. Obviously, Paasche's index number has a downward bias. Dorbish and Bowley index number and Fisher's index number have taken into account both current year as well as base year prices and quantities. Hence it is free from bias, upward as well as downward.

Example 8.3.2. From the following data construct price index using (i) Laspeyre's Method, (ii) Paasche's Method, (iii) Dorbish and Bowley Method and (iv) Fisher's index number.

Commodity	Base year(2005)		Current year(2012)	
	Price	Quantity	Price	Quantity
A	15	4	30	3
B	40	7	65	8
C	25	10	40	8
D	15	6	20	10
E	75	2	90	2

To compute the required index numbers let us prepare the table given below:

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1	p_1q_0	p_0q_1
A	15	4	30	3	60	90	120	45
B	40	7	65	8	280	520	455	320
C	25	10	40	8	250	320	400	200
D	15	6	20	10	90	200	120	150
E	75	2	90	2	150	180	180	150
Total					830	1310	1275	865

We have,

$$\Sigma p_0q_0 = 830,\ \Sigma p_1q_1 = 1310,\ \Sigma p_1q_0 = 1275,\ \Sigma p_0q_1 = 865.$$

$$\begin{aligned}
\text{Laspeyre's index number, } L &= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 \\
&= \frac{1275}{830} \times 100 \\
&= 153.615. \\
\text{Paasche's index number, } P &= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 \\
&= \frac{1310}{865} \\
&= 151.445. \\
\text{Dorbish and Bowley index number} &= \frac{L+P}{2} \\
&= \frac{153.61+151.45}{2} \\
&= 152.530.
\end{aligned}$$

$$\begin{aligned}
\text{Fisher's index number} &= \sqrt{L \times P} \\
&= \sqrt{153.61 \times 151.45} \\
&= 152.526.
\end{aligned}$$

Weighted Aggregates of Price Relatives

Quantity weights are appropriate when we deal with the prices themselves and value weights are appropriate when we deal with price relatives. If we use money value of commodities consumed or produced in the base year as weights, we get

$$P_{01} = \frac{\Sigma p_1 q_0 (\frac{p_1}{p_0})}{\Sigma p_0 q_0} \times 100 = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = L, \text{ Laspeyre's index number.}$$

If,however, we use the money values of commodities consumed in the current year as weights, then

$$P_{01} = \frac{\Sigma p_0 q_1 (\frac{p_1}{p_0})}{\Sigma p_0 q_1} \times 100 = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = L, \text{ Paasche's index number.}$$

Thus the quantity weights applied to prices and value weights applied to price relatives lead to identical results.

8.3.1 Consumer Price Index

A *consumer price index*(CPI) measures changes in the price level of consumer goods and services purchased by households during the current period as compared to the base period. The CPI is a statistical estimate constructed using the prices of a sample of representative items whose prices are collected periodically. The 'basket' of goods for calculation of CPI is developed from detailed expenditure information provided by families and individuals on what they actually bought. The expenditure items are classified into several categories such as food, clothing, housing, health care, education etc. The CPI does not include investment items, such as stocks, bonds, real estate and

life insurance as these items relate to savings and not to consumption expenses. Sub-indices are computed for different categories of goods and services. These sub-indices being combined to produce the overall index with weights reflecting their shares in the total of the consumer expenditures covered by the index. It is one of the several price indices calculated by most national statistical agencies.

The annual percentage change in a CPI is used as a measure of inflation. Inflation has been defined as a process of continuously rising of prices or equivalently, of a continuously falling value of money. A CPI can be used to index (that is, adjust for the effect of inflation) the real value of wages or salaries or pensions for regulating prices and for deflating monetary magnitudes to show changes in real values. The CPI is generally the best measure for adjusting payments to employees when the intent is to allow employees to purchase at today's prices, a basket of goods and services equivalent to one that they could purchase in an earlier period.

Is the Consumer Price Index (CPI) a cost-of-living index?
The CPI frequently is called a cost-of-living index, but it differs in important ways from a complete cost-of-living measure. A *cost-of-living index* would measure changes over time in the amount that consumers need to spend to reach a certain 'utility level' or 'standard of living'. Both the CPI and a cost-of-living index would reflect changes in the prices of goods and services, such as food and clothing, that are directly purchased from the market; but a complete cost-of-living index would go beyond this. It also takes into account changes in other governmental or environmental factors that affect consumers' well-being. It is very difficult to determine the proper treatment of public goods, such as safety and education, and other broad concerns such as health, water quality, and crime that would comprise a complete cost-of-living framework. The CPI is considered to be an upper bound to cost-of-living index. The cost of living index number refers to a particular population group such as middle class people, low salaried workers living in a certain well-defined geographical region such as a city or an industrial area.

The index is usually computed monthly, or quarterly in some countries, as a weighted average of sub-indices for different components of consumer expenditure, such as food, housing, clothing. In India, Ministry of Labour is responsible for the publication of important economic indicators like Consumer Price Index Numbers for industrial, agricultural and rural labourers; wage rate indices etc.

Two basic types of data are needed to construct the CPI- price data and weighting data. The price data are collected for a sample of goods and services from a sample of sales outlets in a sample of locations for a sample of times. The weighting data are estimates of the shares of the different types of expenditure in the total expenditure covered by the index. These weights are usually based upon expenditure data obtained from expenditure surveys for a sample of households or upon estimates of the composition of consumption expenditure. The consumer price index is based upon the Laspeyres method. As we have seen, the quantities of commodities consumed by a particular group in the base year are the weights.

$$\text{Consumer price index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \Sigma \left(\frac{p_0 q_0}{\Sigma p_0 q_0} \right) \frac{p_1}{p_0} \times 100.$$

The second formula of CPI shows that it is also the weighted average of the price relatives with weights which are the amounts of expenditure on various items in the base year. When the first formula is used it is usually termed as *aggregate expenditure method* and when the second formula is used it is called the weighted average price relative method or *family budget method*.

Retail prices of items and services used in the computation of CPI have to be collected both for the base period and current period. The procedure of computation of CPI is illustrated in the next example. Note that for purposes of illustration the number of consumption groups and also the number of items in each group have been reduced in this example.

Example 8.3.3. Computation of consumer price index number by the aggregate expenditure method is given below:

Article	Unit for quotations	Base year quantitie(q_0)	Base year price(p_0)	Current year price(p_1)	p_0q_0	p_1q_0
Food:						
Rice	Kilogram	400	20	35	8000	14000
Wheat	Kilogram	200	15	25	3000	5000
Dal	Kilogram	80	23	45	1840	3600
Sugar	Kilogram	40	18	38	720	1520
Oil	Kilogram	20	30	75	600	1500
Milk	Litre	360	15	32	5400	11520
Vegetables	Kilogram	300	12	25	3600	7500
Fuel and Electricity:						
LPG	Cylider	12	200	750	2400	9000
Firewood	Ton	4	700	2000	2800	8000
Electricity	kWh	1800	1	2	1800	3600
Clothing:	-	-	-	-	5000	15000
House rent:	-	-	-	-	18000	30000
				TOTAL	53160	110240

The consumer price index number for the current year with respect to the base year is

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{110240}{53160} \times 100 = 207.374.$$

8.3.2 Wholesale Price Index

In a dynamic world, prices do not remain constant. Inflation rate calculated on the basis of the movement of the wholesale price index(WPI) is an important measure to monitor the dynamic movement of prices. As WPI captures price movements in a most comprehensive way, it is widely used by governments, banks, industries and business circles. Important monetary and fiscal policy changes are often linked to WPI movements. Similarly, the movement of WPI serves as an important determinant, in formulation of trade, fiscal and other economic policies by the Government of India.

In India, the Office of the Economic Adviser in the Department of Industrial Policy and Promotion under Ministry of Commerce and Industry is responsible for compiling WPI and releasing it. Ever since the introduction of the WPI on a regular basis, six revisions have taken place introducing the new base years, viz., 1948- 49, 1952- 53, 1961-

62, 1970- 71, 1981- 82 and 1993- 94. Latest revision of WPI has been done by shifting base year from 1993- 94 to 2004-05. At present, a representative commodity basket comprising 676 items has been selected and weighting diagram has been derived for the new series consistent with the structure of the economy. The number of quotations selected for collecting price data for the 676 items is 5482.

Characteristics of Index Numbers

From the above discussion the characteristics of index numbers can be classified as follows:

1. Index numbers are specialized averages.
   Index numbers help in comparing the changes in variables which are in different units.

2. Index numbers are expressed in percentages.
   Index numbers are expressed in terms of percentages so as to show the extent of changes.

3. Index numbers measure changes not capable of direct measurement.
   Where it is difficult to measure the variation in the effects of a group of variables directly, but relative variation can be measured with the help of index numbers.

4. Index number is for comparison.
   Index numbers compare changes taking place over time or between places.

General Principles in the Construction of Index Numbers

The following general principles are to be carefully adopted in the construction of index numbers.

1. Purpose or objective
   Every index number has got its own uses. The nature of the information to be collected and the method to be followed depend mainly on the purpose and the scope of the index number. Before collecting data, it is important to decide what we are trying to measure, and also how we intend to use our measures.

2. Selection of items to be included
   The number of commodities that should be included in the index number depends largely on the purpose of the index number. The commodities selected should be representative of the tastes, habits, customs and necessities of the people to whom the index number relates. In short, the items to be included should match with the purpose.

3. Data collection and sources of data
   A sufficiently large sample of relevant items must be selected to obtain reliable index numbers. When selecting the sources of data for index numbers we may rely on persons or agencies who possess the basic information needed. Thus, if retail food price changes are being measured, quotations should be from super-markets, public distribution stores, and any other important outlets. Care should be taken to ascertain how the data are collected. Commodities (items) are ordinarily selected from the various component groups so that it is a representative of the population.

4. Selection of base period
   A base year is one with reference to which price changes in the current year are expressed. We cannot say whether price level in the current year is increased or decreased unless the price level of the current year is compared with the price level of the base year. Therefore the base year must be carefully selected. We want a base where prices (or quantities) were not unnaturally high or low. It must be a "normal year" and should be free from all kinds of abnormalities like war, earthquakes, bad weather etc. Also, people are not happy with a base period which is too far in the past. Furthermore, tastes and availability can change a great deal over time so that such an index could be seriously misleading. One way to avoid these problems is to use chain-base index numbers. A chain-base index number is particularly suited for period by period comparisons, but a fixed-base index number makes it easier to compare the movement of prices over time.

5. Choice of suitable weights and formula
   In some cases the relative importance of all items are not equal so that we need a set of weights. In the construction of consumer price indices the weights may be either quantity consumed or the price of the commodity. After determining the weights we have to decide which formula is to be used for constructing the index? Index number computation involves averaging of values. AS we have studied earlier each average has its own merits and demerits. Hence a formula with a proper choice of average has to be used for computation of an index number.

8.3.3 Changing the Base Period or Splicing

It is usual to update the base period when any significant change which makes comparison with earlier figures meaningless. In some situations, different series of the same index may be chained or spliced and present on a uniform base.

In order to change from one index series to another we need values for both indices in one period. The ratio of these two values forms the basis of any conversion between them.

$$\text{New IN} = \text{Old IN} \times \frac{\text{New IN}}{\text{Old IN}}.$$

Example 8.3.4. The expenditure for cultivation of paddy in a hectare of paddy field is index linked and is described by the following indices:

Year	1	2	3	4	5	6	7	8
Index 1	100	138	162	196	220			
Index 2					100	125	140	165

Complete each index series. If the expenditure for Year 3 is given to be ₹4860, calculate the expenditures for the other years.

$$\text{Index 1 for Year 6} = \text{Index 2} \times \frac{220}{100} = 125 \times \frac{220}{100} = 275.$$

$$\text{Index 2 for Year 1} = \text{Index 1} \times \frac{100}{220} = 100 \times \frac{100}{220} = 45.45.$$

Similarly the remaining indices can be computed from the known year using the ratio of the relevant index numbers from both series.

Year	1	2	3	4	5	6	7	8
Index 1	100	138	162	196	220	275	308	363
Index 2	45.5	62.7	73.6	89.1	100	125	140	165
Expenditure(Rs)	3000	4140	4860	5880	6600	8250	9240	10890

If the cultivation expenditure is ₹4860 in Year 3, then it is for:

$$\text{Year 1} = 4860 \times \frac{100}{162} = 3000 = 4860 \times \frac{45.45}{73.64}.$$

$$\text{Year 2} = 4860 \times \frac{138}{162} = 3000 = 4140 \times \frac{62.73}{73.64}.$$

8.3.4 Tests of Index Numbers

We have seen several formulae for computing index numbers and the problem is that of selecting the most appropriate one in a given situation. If all prices moved in the same direction and changed at the same ratio, then it would make no difference what formula you were chosen. The following tests of consistent behaviour are suggested for choosing an appropriate index number.

Unit Test

The unit test requires that the formula for constructing an index should be independent of the units in which prices and quantities are quoted. Except for the simple aggregate index all other formulae discussed earlier satisfy this test.

Time Reversal Test

It is a test to determine whether a given method will work in both ways,, namely, forward and backward. The test is that the formula for calculating the index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as the base. In other words, when the data for any two years are treated by the same method, but with the bases reversed the two index numbers obtained should be reciprocals of each other so that their product is unity. Symbolically, the following relation should be satisfied.

$$P_{01} \times P_{10} = 1,$$

where P_{01} is the index for time 1 on time 0 as the base and P_{10} the index for time 0 on time 1 as the base. If the product is not unity, then there is a time bias in the method.

Factor Reversal Test

It holds that the product of a price index and the quantity index should be equal to the corresponding values index. In other words, the test is that the change in price multiplied by the change in quantity should be equal to the total change in value. The total value of a given commodity in a given year and the product of the quantity and the price per unit (total value $=p \times q$). The ratio of the total value in one year to the total value in the preceding year is p_1q_1/p_0q_0. If P_{01} represents the change in price in the current year and Q_{01} the change in quantity in the current year. Then

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}.$$

The factor reversal test is satisfied only by the Fishers index number.
Proof:

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}.$$

Now, changing p to q and q to p, we have

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_o p_1}}.$$

Thus,

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}.$$

It is Irving Fisher who suggested both time reversal test and factor reversal test. Fisher's index number satisfies both time reversal test and factor reversal test and hence it is known as an *ideal index number*.

8.4 USES OF INDEX NUMBERS

Index numbers are today one of the most widely used statistical devices. They are particularly useful in measuring relative changes. The following are some of the important uses of index numbers.

1. Index numbers measure the relative changes and can be used as a tool for comparison.
   Index numbers give better idea of changes in levels of prices, production, and sales etc. The index number reduces the changes of price level into more useful and comparable form. Index numbers are easy to compare as they are expressed in percentages. They are not restricted to the price phenomenon alone. Any aspect, for example, intelligence quotient(I.Q.) is also an index number. A frequently used index for determining obesity of persons is body mass index(BMI).

   $$\text{B.M.I.} = \frac{\text{Weight in kilogram}}{\text{Height in meter}^2}.$$

   Thus various kinds of index numbers serve different users. Various index numbers computed for different purposes are of immense value in dealing with different economic problems.

2. They measure the pulse of the economy.
   Index number of general price level is a measure of the purchasing power of money. Suppose the price of rice rises from ₹1200 per quintal in 2000 to ₹3600 per quintal in 2012 it means that in 2012 one can buy only third of rice if he spends the same amount which he was spending on rice in 2000. Thus the value (or purchasing power) of a rupee is simply the reciprocal of an appropriate price index written as a proportion. If prices increase by 25 per cent the price index is 125 and what a rupee will buy is only 100/125 or 4/5 or what it used to buy. In other words the purchasing power of rupee is 4/5 of what it was or 80

percent. This is the same as saying that the *purchasing power of money* is the reciprocal of the price index. The general expression may be given thus.

$$\text{Purchasing power of money} = \frac{1}{\text{price index}}.$$

3. They are used for deflating.
   Since the value of money goes down with rising prices the salaried people are interested not so much in money wages but they are interested in real wages or how much their income or wage will buy. For calculating real wages we can multiply money wages by a quantity measuring the purchasing power of the rupee or better we divide the cash wages by an appropriate price index. This process is referred to as *deflating*. By deflating we mean making allowances for the effect of changing price levels. A rise in price level means a reduction in the purchasing power of money. In principle it appears to be very simple but in practice the main difficulty consists in finding appropriate index to deflate a given set of values or appropriate deflators. Generally we use consumer price index as the deflator. The process of deflating can be expressed in the form of a formula as follows:

$$\text{Real wage or income} = \frac{\text{Money wage}}{\text{Consumer Price Index}} \times 100.$$

   In all fields of economy the wage adjustment are done with the study of consumer price index numbers. Dearness allowances of the employees are determined on the basis of consumer price index number. The stability of price or their inflating or deflating condition can well be observed with the help of indices. Hence they are called *economic barometers*.

4. Index number help to compare the standard of living of different classes of people.

5. They help in formulating polices.

6. Index number serves as a guide to make the relevant decisions for example, investment index number like NSE, BSE, etc are of great help to those interested in the stock market.

Example 8.4.1. The following table gives the annual income of a teacher and the price index number during 2007-2012. Prepare index number to show the changes in the real income of the teacher:

Year	2007	2008	2009	2010	2011	2012
Income	37000	42000	50000	55000	65000	72000
Price index number	100	120	145	170	210	260

Index number showing changes in the real income of the teacher is shown in the following table.

Year	Income(Rs)	Price I.N.	Real income	Real income I.N.
2007	37000	100	37000/100× 100=37000.0	100.0
2008	42000	120	42000/120× 100=35000.0	94.6
2009	50000	145	50000/145× 100=34482.7	93.2
2010	55000	170	55000/170× 100=32352.9	87.4
2011	65000	210	65000/210× 100=30952.4	83.7
2012	72000	260	72000/260× 100=27692.3	74.8

8.5 EXERCISES

1. If the cost price of 20 articles is equal to the selling price of 16 articles, what is the percentage of profit or loss that the merchant makes?
   (a) 20% profit (b) 25% loss (c)25% profit (d) 33.33% loss.

2. If the price of petrol increases by 25%, by how much must a user cut down his consumption so that his expenditure on petrol remains constant?
   (a) 25% (b) 16.67% (c) 20% (d) 33.33%.

3. The proportion of milk and water in 3 samples is 2:1, 3:2 and 5:3. A mixture comprising of equal quantities of all three samples is made. The proportion of milk and water in the mixture is:
   (a) 2:1 (b)5:1 (c)99:61 (d)227:133.

4. You're selling second-hand books and a buyer accidentally pays you the extra 20% VAT above the sale price when they didn't have to, making the total paid ₹60. They ask you to transfer 20% of the total they paid back to them. What percentage should you actually give back?
   (a) 20% (b) 16.7% (c) 15% (d) none of these.

5. A report claims that drinking coffee increases your risk of a certain type of mouth cancer by 50% (it currently affects 1 in 100,000 non-coffee-drinking people). If a million people all started drinking coffee, how many extra cases of that mouth cancer would you expect?
   (a) 5 (b) 50 (c) 500 (d) 10.

6. You are choosing between two universities by looking at their pass rates in different subjects. University A passed 98 of its 120 English students (81.7%) and 56 of its 78 mathematics students (71.8%). University B passed 1,367 of its 1,700 English students (80.4%) and 46 of its 87 maths students (52.9%). Which university has the better percentage of total student passes?
   (a) University A (b) University B (c) They have identical pass rates(d) data is insufficient.

7. Index number is a:
   (a) measure of relative changes (b)a special type of average
   (c) a percentage relative (d) all of these.

8. If the index number for 2010 to the base 2000 is 250, the index number for 2000 to the base 2010 is:
   (a) 4 (b) 40 (c) 400 (d)none of these.

9. Laspeyre's index number possesses:
(a) downward bias (b) no bias (c) upward bias (d) any of these.

10. If the CPI for 2012 is 800, then the purchasing power of a rupee is:
(a) 0.125 paise (b) 12.5 paise (c) 8 paise (d) none of these.

11. Suppose a family spends on food, housing and clothing in the ratio 5:3:2. If there is a rise in prices of these heads by 40,30 and 20 percent respectively, then the family budget will be increased by:
(a) 33% (b) 30% (c) 25% (d) none of these.

12. If Laspeyre's price index is 225 and Paasche's price index is 196, then Fisher's index is:
(a) 205 (b) 210.5 (c) 210 (d) 215.

13. The CPI in 2013 increases by 80% as compared to the base 2006. A person getting salary ₹30,000 in 2006 should now get a salary of:
(a) 24,000 (b) 54,000 (c) 40,000 (d) none of these.

14. The CPI for 2011 and 2012 to the base 2006 are 160 and 200 respectively. The CPI for 2012 to the base 2011 is:
(a) 125 (b) 80 (c) 130 (d) none of these.

15. The index number for 2012 to the base 2010 is 125 and for 2010 to the base 2012 is 80. The given indices satisfy:
(a) time reversal test (b) circular test (c) factor reversal test (d) all of these.

16. Fisher's ideal formula of index number does not satisfy:
(a) time reversal test (b) circular test (c) factor reversal test (d) unit test.

17. The error(s) involved in the construction of index numbers is(are):
(a)sampling error (b) error in data collection
(c) wrong formula selection (d) all of these.

18. The salary of a person in the base year is ₹2000 per month and in the current year ₹5000. If the CPI is 325 then the allowance required to maintain the same standard of living is:
(a) 3000 (b)3500 (c)1500 (d) none of these.

19. The ratio of boys to girls in a class is 5 : 3. The class has 16 more boys than girls. How many girls are there in the class?

20. The daily wages of two persons are in the ratio of 4:7. If each receives an increase of ₹25 in the wage, the ratio is altered to 3: 5. Find their respective wages.

21. 60 liters of diesel is required to travel 600 km using a 800 cc engine. If the volume of diesel required to cover a distance varies directly as the capacity of the engine, then how many liters of diesel is required to travel 800 km using 1200 cc engine?

22. A, B and C play cricket. A's runs are to B's runs and B's runs are to C's as 3:2. They get altogether 342 runs. How many runs did A make?

23. A, B and C enter into a partnership by investing ₹3600, ₹4400 and ₹2800. A is a working partner and gets one-fourth of the profit for his services and the remaining profit is divided amongst the three in the rate of their investments. What is the amount of profit that B gets if A gets a total of ₹8000?

24. Define an index number.

25. Distinguish between a price, quantity and value index number. Give examples to show the situations in which each can be used appropriately.

26. Describe the purpose of weighting in the computation of index numbers.

27. State briefly, giving reasons, the types of index numbers you consider most suitable in each of the following cases:

    (i) Quarterly production of rubber;

    (ii) Weekly sales in a shop;

    (iii) Wholesale prices of cereals;

    (iv) Construction costs;

    (v) Import and export of automobiles.

28. What is a general purpose index number and a specific purpose index number? Give examples of each type and describe their uses.

29. Which is the appropriate average for computing the price index number with the following data? Give reasons for your answer.

Commodity	Unit	1990	2012
Rice	Kilogram	225	850
Wheat	Kilogram	180	710
Fire wood	Cubic meter	150	600
LPG	Cylinder	100	790
Cloth	Meter	150	730

30. Compute index numbers of Paasche, Laspeyre, Fisher and Marshall-Edgeworth types for the following data.

Commodity	Base year		Current year	
	Quantity	Price	Quantity	Price
A	25	14	30	26
B	30	7	25	12
C	35	10	40	18
D	15	6	15	10
E	15	20	10	45
F	10	8	15	8

31. From the fixed base index numbers given below, prepare chain base index numbers.

Years	2004	2005	2006	2007	2008	2009	2010
Index numbers	96	94	99	105	102	108	114

32. From the chain base index numbers given below, prepare fixed base index numbers.

Years	2004	2005	2006	2007	2008	2009	2010
Index numbers	95	107	115	121	125	120	130

33. A budget enquiry conducted among the workers of an industrial unit gave the following information.

Monthly expenses	Food 30%	Fuel and lighting 20%	Clothing 15%	Rent 15%	Miscellaneous 20%
Prices (2004)	2000	300	350	500	150
Prices (2012)	6500	2100	1200	2000	2500

What changes in cost of living figures of 2012 as compared to that of 2004 are seen?

34. Following are the wholesale price index numbers from 1995 to 2000 to base 1990.

Year	1995	1996	1997	1998	1999	2000
I.N.	175	170	180	192	203	215

Find the wholesale price indices to the base 1997.

35. The following table gives the index numbers of wholesale prices from 1997 to 2000 to the base 1990 in the first row and from 2001 to 2004 to the base 2000. Combine the two series of indices into one.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Indices 1	180	192	203	215				
Indices 2					107	118	136	147

36. The following are the group index numbers and the percentage of expenditure for each group of an average working class family budget. Construct a suitable index number assigning suitable weights.

Groups	Weights	Index number
Food	45	160
Electricity and fuel	10	270
Travel	10	150
House rent	10	130
Clothing	10	120
Miscellaneous	15	75

37. Show that Marshall-Edgeworth price index number satisfies circular test if the quantities consumed in all years are equal. Does this index number satisfy time reversal test?

38. Explain the uses of consumer price index numbers.

39. Describe the procedure in the construction of cost of of living index number for your city.