HIGHER SECONDARY COURSE

STATISTICS

CLASS - XI

Government of Kerala
DEPARTMENT OF EDUCATION

State Council of Educational Research and Training (SCERT); Kerala
2016

THE NATIONAL ANTHEM

Jana-gana-mana adhinayaka, jaya he
Bharatha-bhagya-vidhata.
Punjab-Sindh-Gujarat-Maratha
Dravida-Utkala-Banga
Vindhya-Himachala-Yamuna-Ganga
Uchchala-Jaladhi-taranga
Tava subha name jage,
Tava subha asisa mage,
Gahe tava jaya gatha.
Jana-gana-mangala-dayaka jaya he
Bharatha-bhagya-vidhata.
Jaya he, jaya he, jaya he,
Jaya jaya jaya, jaya he!

PLEDGE

India is my country. All Indians are my brothers and sisters.

I love my country, and I am proud of its rich and varied heritage. I shall always strive to be worthy of it.

I shall give my parents, teachers and all elders respect, and treat everyone with courtesy.

To my country and my people, I pledge my devotion. In their well-being and prosperity alone lies my happiness.

Prepared by :
State Council of Educational Research and Training (SCERT)
Poojappura, Thiruvananthapuram 695012, Kerala

Website : www.scertkerala.gov.in *e-mail* : scertkerala@gmail.com
Phone : 0471 - 2341883, Fax : 0471 - 2341869
Typesetting and Layout : SCERT
To be printed in quality paper - 80gsm map litho (snow-white)

Foreword

I am pleased to introduce the new textbook in Statistics for Class XI. The book is prepared strictly in accordance with the yardsticks as suggested by the revised Higher Secondary School Curriculum, 2013. Maximum care is given to generate the targeted learning outcomes in the beginners of Statistics.

In order to kindle a genuine interest in the learners, the subject is presented with the help of cartoons, simple descriptions, problems, learning exercises and questions for self-evaluation. A keen learner can successfully do the exercises individually or in groups. However, the learners need not hesitate to seek the advice of their teachers if they find any exercise difficult to solve on their own. I hope this book will surely inspire the learners to take up higher studies in Statistics.

On behalf of the SCERT, I take this opportunity to thank the team of experts who have put in their valuable effort in realizing the textbook.

We look forward to your valuable criticisms, creative suggestions and advice for improvement.

Wish you all success!

Dr. P. A. Fathima
Director
SCERT; Kerala

Textbook Development Team

1. **MANOJ. K**
   HSS Panagad , Thrissur.
2. **ANWAR SHAMEEM. Z. A**
   GHSS Medical College
   Campus, Kozhikode.
3. **SAJISHKUMAR. M**
   MNKM HSS, Chittilamchery
   Palakkad.
4. **BIJU . G. V**
   GHSS Kulakkada, Kollam.
5. **MARY GEORGE**
   St. Raphael's CGHSS,
   Ollur Thrissur
6. **MOHAMMED ASLAM.K**
   PPM HSS Kottukkara
   Malappuram .
7. **SREESAN.M.B**
   Karimpuzha HSS, Thottara
   Palakkad.
8. **BIJI.K**
   NSS HSS, Prakkulam, Kollam.
9. **SAKKEER.M**
   Govt. VHSS Meppayyur
   Kozhikode.
10. **DEEPA KOSHY**
    PSVPM HSS, Ayravon
    Pathanamthitta.

Experts

P.K. VENUGOPAL,
Associate Professor, Department of Statistics
Sree Kerala Varma College, Thrissur.

Dr. K. K. HAMSA,
Associate Professor, Department of Statistics
Farook College, Kozhikode.

Artists

Hrishikesh K. B (LaTex)

Harikumar (HaKu - Cartoonist)

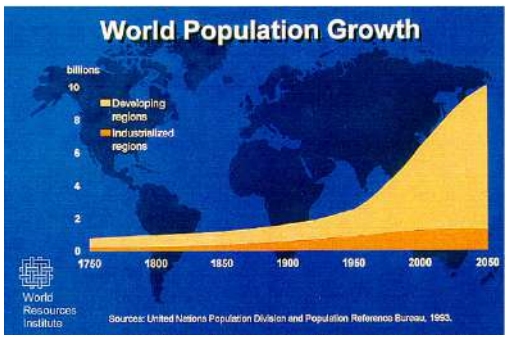
Academic Co-ordinator

Dr. Chandini. K. K
Head, Higher Secondary and Teacher Training

Contents

Introduction

Decision making is one of the highest forms of human activities. Every day we make decisions that may be personal, business related or of some other kind. Usually these decisions are made under conditions of uncertainty. Many times the situations or problems we face in the real world have no precise or definite solution. Statistical methods help us to make scientific and intelligent decisions in such situations. In recent years, the growth of statistics has made itself felt in almost every phase of human activity. Statistics no longer consists merely of collection of data and their presentation in charts and tables. It is now considered the science of inferences on observed data and the entire problem of making decisions in the face of uncertainty. This covers considerable ground since uncertainties are met when we flip a coin, when a dietician experiments with food additives, when an actuary determines life insurance premiums, when a quality control engineer accepts or rejects manufactured products, when a teacher compares the abilities of students, when an economist forecast trend, when a newspaper predicts an election result and so forth.



It would be presumptuous to say that statistics in its present state of development can handle all situations involving uncertainties, but the new techniques are constantly being developed and modern statistics can, provide the framework for taking at these situations in a logical and systematic fashion. The beginning of mathematics of statistics may be found in mid-eighteenth century studies in probability motivated by interest in game of chance. Thus the scholars began to apply probability theory to actuarial problems to some aspects of social

sciences. By this century it found application in all phases of human endeavour that in some way involve an element of uncertainty or risk.

Like almost all fields of study, statistics has two aspects, Theoretical and Applied. Theoretical or Mathematical Statistics deals with development, derivations and proof of statistical theorems, formulai, rules and laws. Applied statistics involves the application of those theorems, formulai, rules and laws to solve real world problems. Broadly speaking, applied statistics can be divided into two areas, Descriptive Statistics and Inferential Statistics. Descriptive statistics consists of methods for analysis of data and the area that deals with decision making procedure is referred to as inferential statistics.

1.1 History of Statistics

The word Statistics have been derived from Latin word "Status" or the Italian word "Statista". The meaning of these words is "Political State" or a Government. Shakespeare used a word Statist in his play Hamlet (1602). In the past, the statistics was used by rulers for official purposes. Even though application of Statistics was very limited, the rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of Government.

Sir Ronald Aylmer Fisher.

Gottfried Achenwall used the word 'statistik' at German University in 1749 which means political science of different countries. In 1771, W. Hooper (Englishman) used the word 'statistics' in his translation of *Elements of Universal Erudition* written by Baron B.F Bieford. In his book, statistics has been defined as the science that teaches us what is the political arrangement of all the

modern states of the known world. There is a big gap between the old statistics and the modern statistics, but old statistics is also used as a part of the present statistics.

During the 18th century English writers have used the word statistics in their works. A lot of work has been done in the end of the nineteenth century.

At the beginning of the 20th century, William S Gosset developed the methods for decision making based on a small set of data. During the 20th century several statisticians were active in developing new methods, theories and application of statistics. The advent of electronic computers is certainly a major factor in the development of modern statistics. **Sir Ronald Aylmer Fisher is known as father of modern statistics.**

1.2 Definition of Statistics

1. ″Statistics can be defined as the collection, presentation and interpretation of numerical data.″ - Croxton and Cowden.

2. ″Statistics are measurement, enumerations or estimates of natural or social phenomena, systematically arranged to exhibit their inner relation.″ -Conner.

3. ″The science of Statistics is essentially a branch of applied mathematics and can be regarded as a mathematics applied to observational data.″ - R.A Fisher.

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

This definition points out some essential characteristics of statistics. These

characteristics are:

(i) **Statistics are the aggregates of facts.** It means, a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

(ii) **Statistics are affected by a number of factors.** For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

(iii) **Statistics must be reasonably accurate.** Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

(iv) **Statistics must be collected in a systematic manner.** If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions. It is collected with a pre-determined purpose.

(v) **Statistics should be placed in relation to each other.** If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

1.3 Functions of Statistics

1. **Statistics simplifies complexity**

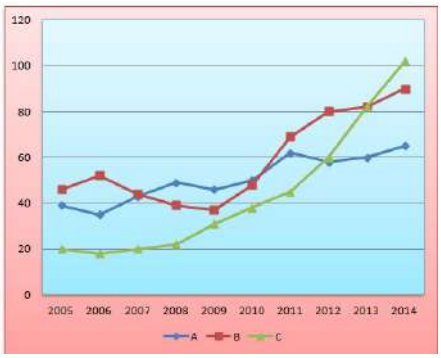
   The complex mass of data are made simple and understandable with the help of statistical methods.

2. **Statistics presents facts in a definite and precise form**

   Statistics presents statements of facts in a precise, quantitative and definite form.

3. **Statistics provides comparison**

Statistics provides a number of suitable methods of comparison between present and past values and hence able to predict future.



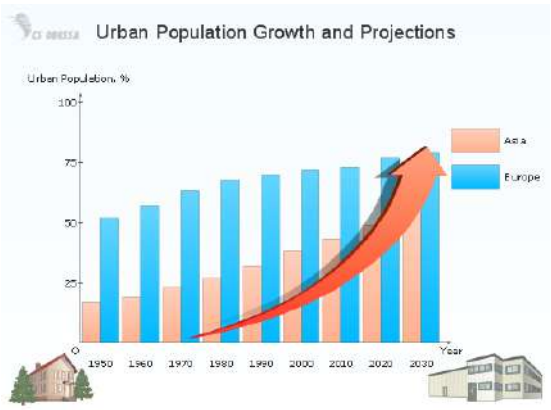
4. **Statistics enlarges human knowledge and experience**

   Statistics makes most of our vague and indefinite opinions, clear and definite.

5. **Statistics helps in formulating policies, testing of hypotheses and forecasting future events**

   Important policies, decision making and forecasting in business, economics, finance, industry, etc are taken on the basis of statistical methods.

1.4 Scope and importance of Statistics



1. **Statistics and Planning :** Statistics is an indispensable tool in planning the modern age. Because of the complexities and uncertainties, planning is essential for solving the complex problem in various walks of life.

2. **Statistics and Economics:** Statistical data and techniques of statistical analysis such as time series analysis and demand analysis are immensely useful in solving problems in Economics .

3. **Statistics and Industry:** In industry, Statistics is widely used in quality control. In production engineering, to find out whether the product is confirming to the specifications or not. Statistical tools, such as inspection plan, control chart, etc. are highly useful.

4. **Statistics and Mathematics:** Statistics is intimately related to Mathematics. Statistical techniques are the outcomes of wide applications of Mathematics.

5. **Statistics and Medical Science:** In medical science the statistical tools for collection, presentation and analysis of observed facts relating to causes and incidence of disease and the result of application of various drugs and medicines are of great importance.

6. **Statistics, Psychology and Education :** In Education and Psychology, Statistics has found wider applications such as, determining (or to determine) the reliability and validity of a test, measuring intelligence quotient, factor analysis, etc.

7. **Statistics and Management Studies :** Statistical analysis is frequently used in providing information for making decisions in the field of marketing, production, finance, banking, investment, purchase and accounting.

Activity

Prepare a report regarding the functions and importance of statistics in daily life by reading the features and reports in news papers and magazines.

1.5 Limitations of Statistics

(i) There are certain phenomena or concepts where Statistics cannot be used. For example, beauty, intelligence and courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

(ii) Statistics reveals the average behaviour, the normal or the general trend. Statistics does not study individual items but deals with aggregate. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet. There may be some points in between where its depth is far more than four feet.

(iii) Since statistics are collected for a particular purpose, such data may or may not be relevant or useful in other situations or cases. For example, secondary data (i.e., collected by a person) need not be useful for another person.

(iv) Statistics are not 100 per cent precise as in Mathematics. Those who use Statistics should be aware of this limitation

Misuse of Statistics

The misuse of Statistics is the main cause of discredit to this science and has led to public distrust in Statistics. The various reasons of misuse are:

(i) Sources of data not given.

(ii) Defective data.

(iii) Unrepresentative sample.

(iv) Inadequate sample.

(v) Unfair Comparisons.

(vi) Unwanted conclusions.

(vii) Inappropriate statistical tools.

1.6 Some applied areas of Statistics

Actuarial Science

Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in the insurance and the finance sectors. Actuaries are professionals who are qualified in this field. In many countries, actuaries must demonstrate their competence by passing a series of rigorous professional examinations. Actuarial science includes a number of interrelating subjects, including Probability, Mathematics, Statistics, Finance, Economics, Financial Economics, and Computer Programming. Historically, actuarial science used deterministic models in the construction of tables and premiums. The science has gone through revolutionary changes during the last 30 years due to the proliferation of high speed computers and the union of stochastic actuarial models with modern financial theory (Frees 1990).

Biostatistics

Biostatistics (sometimes referred to as Biometry or Biometrics) is the application of Statistics to a wide range of topics in Biology. The science of Biostatistics encompasses the design of biological experiments, especially in medicine, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results. A major branch is Medical Biostatistics, which is exclusively concerned with medicine and health.

Agricultural Statistics

The agricultural investigations are based on the application of statistical methods and procedures which are helpful in testing hypotheses using observed data, in making estimations of parameters and in predictions. The application of statistical principles and methods is necessary for effective practice in resolving various problems that arise in the many branches of agricultural activity. Because of the variability inherent in biological and agricultural data, knowledge of statistics is necessary for their understanding and interpretation. Numerous activities in agriculture are very different from each other, resulting in different branches of agricultural science like: field crop production, vegetable production, horticulture, fruit growing, plant protection, livestock, veterinary medicine, agricultural mechanization, water resources, agricultural economics, etc.

Activity

List out the various branches of statistics related to different disciplines.

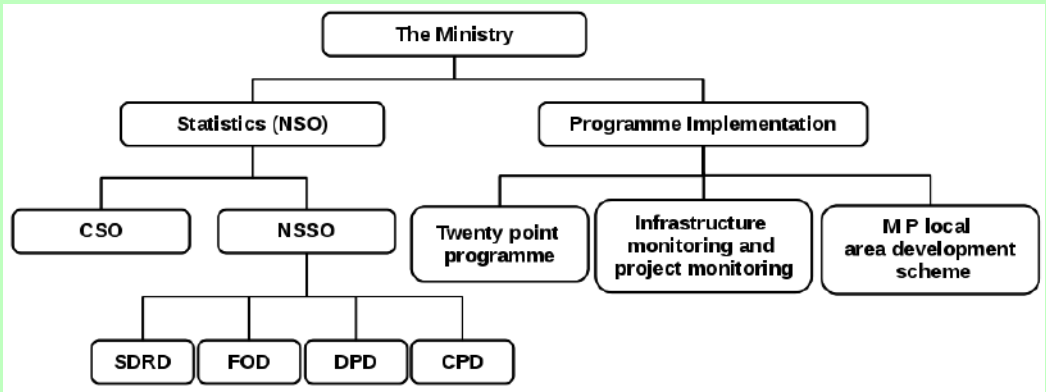
1.7 Official Statistics

Official Statistics are statistics published by government agencies or other public bodies such as international organizations. They provide quantitative or qualitative information on all major areas of citizens' lives. Official Statistics make information on economic and social development accessible to the public, allowing the impact of government policies to be assessed, thus improving accountability.

The Ministry of Statistics and Programme Implementation (MOSPI) came into

existence as an Independent Ministry in 1999 after the merging of the Department of Statistics and the Department of Programme Implementation.

The Ministry has two wings, Statistics and Programme Implementation.



- SDRD - Survey Design and Research Division
- FOD - Field Operations Division
- DPD - Data Processing Division
- CPD - Co-ordination and Publication Division

The Statistics Wing called the National Statistical Office(NSO) consists of the Central Statistical Office (CSO), the Computer Centre and the National Sample Survey Office (NSSO).

Central Statistical Office (CSO)

The Central Statistical Office which is one of the two wings of the National Statistical Organisation (NSO) is responsible for co-ordination of statistical activities in the country and for evolving and maintaining statistical standards. Its activities include compilation of National Accounts; conduct of Annual Survey of Industries and Economic Censuses, compilation of Index of Industrial Production as well as Consumer Price Indices. It also deals with various social statistics, training, international cooperation, Industrial Classification, etc.

The CSO is headed by a Director-General who is assisted by 5 Additional Director-Generals looking after the National Accounts Division, Social Statistics Division, Economic Statistics Division, Training Division and the Coordination and Publication Division.

CSO is located in the Sardar Patel Bhawan, Parliament Street, New Delhi. The Industrial Statistics Wing of CSO is located in Kolkata. The Computer Centre also under the CSO is located in R K Puram, New Delhi.

National Sample Survey Office (NSSO)

The National Sample Survey Organisation, now known as National Sample Survey Office, is an organization under the Ministry of Statistic of the Government of India. It is the largest organisation in India, conducting regular socio-economic surveys. It was established in 1950.

NSSO has four divisions:

1. Survey Design and Research Division (SDRD)
2. Field Operations Division (FOD)
3. Data Processing Division (DPD)
4. Co-ordination and Publication Division (CPD)

The Programme Implementation Wing has three Divisions, namely,

(i) Twenty Point Programme

(ii) Infrastructure Monitoring and Project Monitoring

(iii) Member of Parliament Local Area Development Scheme.

Besides these three wings, there is National Statistical Commission created through a Resolution of Government of India (MOSPI) and one autonomous Institute, viz., Indian Statistical Institute declared as an institute of National importance by an Act of Parliament.

Know your progress

Discuss the important statistical organizations (offices) in India.

Indian Statistical Institute (ISI)

Prof. P.C. Mahalanobis

Indian Statistical Institute (ISI), a unique institution devoted to the research, teaching and application of statistics, natural sciences and social sciences. Founded by Prof.Prasanta Chandra Mahalanobis in Kolkata on 17th December, 1931.He is known as father of Indian statistics. The Indian Statistical Institute publishes Sankhya, the Indian Journal of Statistics.

In recognition of the notable contributions made by Prof.P.C.Mahalanobis in the fields of economic planning and statistical development in the post independent era, the Govt. of India has decided to designate 29th June every year, coinciding with his birth anniversary, as Statistics Day in the category of special day to be celebrated at the national level. The Day is celebrated by holding seminars, discussions and competitions to highlight the importance of official statistics in national development.

Economics & Statistics Department

The Directorate of Economics & Statistics, Government of Kerala is the nodal agency of the State responsible for the systematic collection, compilation, analysis, interpretation and dissemination of statistics relating to various sectors of Kerala Economy.

The Directorate of Economics & Statistics is the nerve centre of the State statistical system. Director is the technical and administrative head of the Department. Being the statistics authority of the State the director functions as the authority for the collection, processing and dissemination of all statistical data relating to the State economy.

Besides the Directorate there are 14 District Offices, each headed by a Deputy Director with the exception of Wayanad. The Deputy Director in the District Offices is assisted by one District Officer, one or more Additional District Officers, one Price Supervisory Officer and one or two Research Officers. At taluk level, there is a Taluk Statistical Office, which is the lowest statistical unit in the State. There are at present 61 Taluk Statistical Offices, each under the control of a Taluk Statistical Officer.

Activity

Visit the nearest economics and statistics department and prepare a detailed report regarding their functions.

Let us sum up

Statistics are all around us. Without statistics we couldn't plan our budgets, pay our taxes, enjoy games to their fullest, evaluate classroom performance, etc. In this chapter we discussed the history ,importance,development,scope and some definitions of statistics . Statistics is applied in all walks of life.Various branches of statistics are explained here .We have seen the functions and roles of the ministry of statistics and programme implementation and the famous Indian Statistical Institute in Kolkatta. We also introduced the Department of Economics and Statistics of the state.

Learning outcomes

After transaction of this unit, the learner:-

- explains the history, definitions and scope of Statistics.
- recognises the importance of Statistics in various fields.
- compares different branches of Statistics.
- illustrates the functions of MOSPI, CSO, NSSO, ISI and Department of Economics and Statistics of Kerala.

Evaluation Items

1. ″ Statistics can be defined as the collection presentation and interpretation of numerical data″. This definition is given by:
   a) R.A Fisher **b)** Horace Secrist
   c) Croxton and Crowden **d)** Conner.

2. Who is known as father of Modern Statistics ?
   a) Conner **b)** R.A Fisher
   c) Mahalanobis **d)** Gosset

3. The discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries is called
   a) Bio statistics **b)** Agricultural statistics
   c) Actuarial Statistics **d)** Production Statistics

4. The Central Statistical Office is located in:
   a) Mumbai **b)** Kolkatta
   c) New Delhi **d)** Chennai

5. The largest organisation in India conducting regular socio-economic surveys is
   a) CSO **b)** NSSO
   c) ISI **d)** NASA

6. Indian Statistical Institute (ISI), is founded by:
   a) P.C. Mahalanobis **b**) R.A fisher
   c) Horace Secrist **d**) C. R.Rao

7. The Indian Statistical Institute (ISI) is situated in :
   a) Kolkatta **b**)Bangaluru
   c) Chennai **d**) Pune

8. National Statistics Day is celebrated on :
   a) June 1 **b**) june 29
   c) july 4 **d**) july 29

9. .....is known as the father of Indian Statistics.
   a) R.A Fisher **b**) S.P Gupta
   c) C.R. Rao **d**) P.C. Mahalanobis

10. The journal published by Indian Statistical Institute (ISI) is
    a) Statistica **b**) Sankhya
    c) Sample surveys **d**) Census

11. Name the nerve centre of Kerala state statistical system.

12. ″ In this century statistics found application in all phases of human endeavour″ . Comment on the statement.

13. How will you critically approach the definition of statistics given by Horace Secrist?

14. Examine the scope of statistics in various fields.

15. ″ The number of accidents is lower in foggy weather than on clear days.Hence it is safe to drive in fog.″ Do you agree with the statement ? Why ?

16. Explain the importance of Statistics in the following branches of study
    a) Acturial science **b**) Bio statistics **c**) Agricultural Statistics

17. Write short notes on the following :
    a)CSO **b**) NSSO **c**) ISI

18. What are the Divisions of NSSO?
19. Give some misuses of Statistics.
20. Write a short note on the Directorate of Economics and Statistics of Government of Kerala.

Answers:

1) c 2)b 3)c 4)c 5)b 6)a 7)a 8)b 9)d 10)b

Introduction

In chapter 1, we discussed Statistics as the study of collection, organization, analysis, interpretation and presentation of data. For studying statistics the first step is collection of data, which we will discuss in detail in this chapter.

2.1 Data Collection

In our day to day life we deal with different types of data collection situation. A teacher might collect information regarding the test score of a student, a journalist might collect information regarding the recent social issues, a politician collects information on how voters plan to vote in the upcoming election, etc. Data collection is the systematic gathering of **data** for a particular purpose from various sources.

Data is the plural of the term datum, which means any measurement, result, fact or observation which gives information. Statistical surveys are the most popular devices for obtaining the desired data.

Before dealing with statistical surveys, we have to familiarize with the following terms.

Statistical Investigation

Statistical investigation includes collection, classification, presentation, analysis and interpretation of data according to well defined procedures. The person authorized to make investigation is known as **Investigator**. In a statistical investigation the investigator formulates the problem, suggests the data collection methods, organises various steps in an appropriate way, analyses the data and interpret the result. Usually, the investigators depute some persons to collect the data from the field. These persons are known as **Enumerators.** The enumerator may not be aware of the investigation procedures completely.

His/her duty is to collect the data for the investigator. It is the duty of the investigator to train and supervise the work of the enumerator. The process of data collection by the enumerator is known as **Enumeration.**

Population and Sample

A **population** consists of all elements, individuals, items or objects whose characteristics are being studied. For a politician, while considering voters plan for the next election, all registered voters in the specified constituency determines the population. If data are collected from each and every unit of the population, the investigation is called **census**. Based on the number of objects in a population, we can classify the population as finite and infinite. A population is said to be **finite**, if the number of individuals involved in the population is finite. All students of Kerala University for the year 2013-14 constitute a finite population. A population which is not finite or extremely large is **infinite**. The population comprises of all people in the world above 18 years of age is considered as an infinite population.

If the population is infinite or is of extremely large size, it is not feasible or practicable to access the entire population for study. As a result, it is apt to take a representative part as a substitute for the entire population. This representative part of the population is known as **sample**. The method of collecting data from the sample is known as **sampling** or **sample survey**. Various sampling designs and their selections are discussed in the last chapter.

The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 BC. In addition, the roman Emperor Augustus (27 BC to 17 AD) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly. In India about 2000 years ago we had an efficient system of collecting administrative statistics, particularly,during the regime of Chandra Gupta Maurya (324 to 300 B.C.). The system of collecting data related to births and

deaths is mentioned in Kautilyas Arthshastra (around 300 B.C.). Ain-i-Akbari, written by Abul Fazl, gives us a detailed account of the administrative and statistical survey conducted during the reign of Emperor Akbar.

Statistical Survey

A survey is a process of collecting data either from the population or from sample units. The statistical survey may be either by Census method or by Sampling method. The purpose of conducting a sample survey is to collect information about population using sample.

Know your progress

1. Illustrate Data, Statistical Investigation and Statistical Survey.
2. Distinguish between Population and sample?
3. Explain Finite and Infinite Population with the help of examples

2.2 Variables

Consider a group of people in a locality. The members of the group are found to be varying in many factors like sex, age, eye colour, intelligence, height, weight, blood pressure etc. The factors which can vary from one object to another are called variables. Among these variables sex, eye colour and intelligence which cannot be numerically measured are called qualitative variables or attributes. A qualitative variable is one that can be identified by noting its presence or identified with different categories of the factor. The other variables height, weight, age and blood pressure which are numerically measured are called quantitative variables. A quantitative variable consists of numerical values. Depending on the values taken by a quantitative variable, it is further classified as discrete variable and continuous variable. If the variable takes specific values only, it is called discrete variable. The variable, number of children in a family, does not take values other than $0, 1, 2, 3$,etc. That is, there is a specified

break between the successive values. This is an example of a discrete variable. A continuous variable takes any value within the defined range of values. Between any two values of a continuous variable, an indefinitely large number of values may occur. Height, weight, time etc are examples of continuous variables. Depending on the type of variables involved, data may also be classified as discrete or continuous.

Know your progress

1. Write examples for variables.
2. Write examples for discrete and continuous variables.
3. Give examples other than those presented in this section of a qualitative variable, discrete quantitative and continuous quantitative variable

Levels of Measurement -

Nominal, Ordinal and Cardinal Data

S.S.Stevens (1906 -1973) described the data into different scales of measurement as nominal,ordinal and cardinal data. This classification is based on the data under consideration. A nominal scale of measurement is used to name categories such as gender, nation, etc. For example the categorization like male, female is a nominal data. While filling Higher Secondary single window application form for admission we give different codes to represent data such as Thiruvanathapuram 01, Kasaragod 14 etc. This is also a nominal data. Here the number is merely a label, does not have a quantitative significance. The only effect of such a measurement is that we can count how many objects fall in each category like 10000 belongs to Thiruvanathapuram,

S.S.Stevens

Stevens classify the cardinal data into two other classes as interval and ratio scale. This classification is generally known as Stevens Taxonomy

9700 belongs to Kasaragod, etc. In the ordinal scale of measurement, we can put an order to the data according to the relation among the values of the variables. While considering the educational qualification of a group of people, we can categorise them as Secondary, Higher Secondary, Graduate, Postgraduate, etc. Here we can rank Secondary -1, Higher Secondary -2, Graduate -3, Postgraduate -4, etc. Here code 3 is surely higher than code 1 as a graduate is at a higher level than a secondary student based on educational qualifications. That is in ordinal scale, there is a specific order or rank for the codes given to each category. The data regarding a quantitative variable is a cardinal data. The height of students in a class, monthly salary of school teachers, marks of students in a class etc. are examples of cardinal data.

Know your progress

1. Compare Nominal, Ordinal, Cardinal data.
2. Give some examples Nominal, Ordinal and Cardinal data.

2.3 Types of Data

Primary and Secondary Data

Data means the raw facts and figures that have been collected. Data can be gathered by looking through existing sources, conducting experiments or by conducting surveys. Based on these sources of collections, statistical data may be classified as primary or secondary. Primary data are those which have to be collected by the investigator for the first time for his/her own purpose. It is fresh in nature. In the words of Wessel data originally collected in the process of investigation are known as primary data. It is collected by using appropriate survey techniques. Data obtained from existing sources which may be published or unpublished are known as secondary data. Secondary data may not be in the required form. These data are obtained by other persons and are being used now at second hand. According to M. M. Blair, "secondary data

those which are already in existence, and which have been collected for some other purpose than the answering of question in hand"

Comparison between Primary and Secondary data

Primary Data	Secondary Data
It is original in nature	It is not original in nature
It is in the form of raw materials	It is in the form of finished products
Collection involves more money and time	Less time and money are needed
Trained persons are required for data collection	The investigator should be vigilant while collecting secondary data
Primary data, after use become secondary data	Secondary data cannot be converted to primary data

2.4 Questionnaire and Schedule

Questionnaires and schedules are series of questions arranged in a logical order so as to collect information for a specified purpose. The purpose may be single or multiple.

Questionnaire

A questionnaire is usually mailed by post or by email to selected informants. The informants are allowed a specified time to fill up the questionnaire and have to return to the investigator. Here the quality of the obtained data depends on the quality of the questions and the honesty of the informants. As the informants are to fill up the data, they should be literate. This method is suitable in cases where the informants are widely scattered. One of the main disadvantages of this method is that the chance of getting incomplete information is large.

Schedule

If the group of informants are not widely scattered, or if they are not literate, the enumerator himself/herself can personally approach the informants with the set of question and collect information. These questions may not be in detailed manner as questionnaire. It may not contain explanatory foot notes or explanations of terms used. These set of questions used for data collection is termed as schedule. In some cases questionnaire itself can be used as a schedule.

Comparison between Questionnaire and Schedule

Questionnaire	Schedule
It is often sent by post	Enumerators carry the schedule personally to the informant
Answers are filled by the respondents	Answers are filled by the enumerators
Informants are to be literate	Informants need not be literate
Success depends on the quality of questions and sincerity of the informant	Success depends on the honesty and competence of the enumerator
Chance of getting incomplete information is more	Chance of getting incomplete information is less as enumerators explain the questions

Requisites of good questionnaire

While preparing a questionnaire we have to keep in mind the following points.

1. Questions should be capable of generating all required information
2. The language and wording of the questions should be convenient to the informant

3. Question should not contain technical terms and words with uncommon meaning, such questions leads to different information from different informants
4. Yes or No questions or multiple answer choice questions should be preferred.
5. Personal questions should be avoided
6. Necessary foot notes should be provided
7. Usually the number of questions should be 20 to 25
8. Questions should have a logical order
9. Questions should be self explanatory
10. Questionnaire should be attractive so as to impress the informant
11. Questions should be unambiguous

Once the questionnaire is ready, it is advisable to conduct a pre-test with the questionnaire for a small group. This is known as pilot survey. It helps the investigator to measure the worthiness and reliability of these questions.

Drafting of Questionnaire

A sample questionnaire is prepared below for studying the socio economic status of people residing in a village.

QUESTIONNAIRE TO COLLECT DATA ABOUT THE SOCIO-ECONOMIC STATUS OF PEOPLE

1. Name :
2. Address:
3. Members in the family : No. of Males ☐ No. of Females ☐
4. Age of family members : (write number of members in each category in Corresponding boxes)
   a) below 10 ☐ b)10-20 ☐ c)20-50 ☐ d) 50 and above ☐
5. Residing in :
   a) Own house ☐ b) Rental ☐
6. Type of House :
   a) Temporary ☐ b) Structured ☐
7. Toilet facility :
   a) Proper ☐ b) Improper ☐
8. Water Facility :
   a) own well ☐ b)water provided by panchayat ☐ c) other Sources ☐
9. Electrified Home :
   Yes ☐ No ☐
10. Details of electronic items in house : (write number of items in each boxes)
    a) Bulbs ☐ b)Refrigerator ☐ c) Fan ☐ d) Television ☐
    e) Tubes ☐ f) Mixer grinder ☐ g) Others ☐
11. Mode of cooking :
    a) Wood ☐ b) Kerosene ☐ c) Gas ☐ d) Electricity ☐

12 Occupation :
a)Govt. service ☐ b) Non Govt. Service ☐ c)Own business ☐
d)Agriculture ☐ e) Others ☐

13 Monthly income of family (in rupees) :
a)Below 8000 ☐ b) 8000-20000 ☐
c) 20000-50000 ☐ d) above 50000 ☐

14 The amount you monthly spend for education of children
a)below 500 ☐ b)500-2000 ☐
c) 2000-4000 ☐ d) above 4000 ☐

15 Do you own a private vehicle? :
Yes ☐ No ☐

16 If Yes,give the number of vehicles in each category. :
a)Two wheeler ☐ b) Three wheeler ☐
c)Car ☐ d) Others ☐

17 The approximate monthly expenditure :
a) Below 5000 ☐ b)5000-15000 ☐
c) 15000-20000 ☐ d) above 20000 ☐

18 Would you prefer an outing occasionally with your family?
Yes ☐ No ☐

19 If Yes what will be your budget for a single trip?
a) below 2000 ☐ b) 2000-5000 ☐ c)above 5000 ☐

20 Are you able to properly maintain your standard of living with your actual income?
Yes ☐ No ☐

21 Any other information regarding your family (Give details in one or two sentences)

(** This data will be used for study purpose only)

2.5 Methods of Primary Data Collection

Direct Personal Interview

If the field of investigation is small, the investigator or enumerator can access all the informants personally and conduct spot enquiry. This method is called direct personal interview. The success of this method depends on the efficiency of the enumerator. The enumerator should be tactful to get all the required information. In this method the enumerator can collect all the supplementary information required for interpretation of data

Indirect Oral Investigation

Consider a situation in which the investigator wants to collect data about a resident in a city. In this case, the investigator may approach a third party, called witness, who is capable of giving sufficient information about the resident. This is a case of indirect oral investigation. Indirect oral investigation is applicable in cases where the informant is reluctant to give information or when the informant is not available. The disadvantage of this method is that the reliability of the information heavily depends on the quality and honesty of the witness or intermediate person.

Direct Observation

This method is widely used by mass media for collecting information or by journalists. Consider a situation in which the investigator wants to report the current situation in an area due to heavy rain and flood. He does not have a predetermined set of questions to collect data. The investigator collects data from what he observes. This may not be in a well defined manner. The investigator must be well equipped so as to collect maximum information from the place. The quality of the information depends mainly on the honesty of the investigator to report it to the maximum extend.

Telephone interview

In some cases the informant may be reluctant to give answer in a face to face personal interview. In such cases it is better to select another method for data collection. Telephone interview is one such method. In this case the investigator collects data from the informant indirectly but personally. This is less time consuming and cheaper than direct personal interview. The disadvantage is that it will not worked in some rural areas were telephonic connection is very low.

Mailed Questionnaires and Schedules

Questionnaires and schedules are one of the most popular methods for collecting primary data. As the questionnaires are usually mailed to the respondents, it is known as mailed questionnaire method. The only difference between the questionnaire and schedule is that in questionnaires the answers are filled by the respondents themselves but in schedule, the answers are filled by the enumerators.

Focus Group Discussion

A Focus Group Discussion (FGD) is a small group discussion guided by a trained leader. It is used to collect more opinions about a specified topic in order to take better decisions in future plans. For example, Mrs. Rema wants to start a pre-school in an area. Her aim is to provide childcare as much as possible. For that she invites the parents who have of children under four years from that locality and arranges Focus Group Discussion. The parents have their own ideas about childcare and other locally adopted programmes. Their suggestions will help Mrs. Rema to start the institution in a well equipped manner. Here Mrs.Rema plays the role of the investigator. She collects information through focus group discussion for her specified purpose.

2.6 Sources of Secondary Data

Any published or unpublished data which are reliable for the current situation is a source of secondary data. While collecting secondary data the investigator must be aware of the following points.

- The geographical area of the collected data.
- The time at which the data was collected.
- The terms and definitions involved in the data.
- The person who collected the data and the purpose for which they are collected.

Some sources of secondary data are listed below

- Government publications.
- Office records in panchayats, municipalities etc.
- Survey reports of various research organizations.
- Survey reports in Journals, Newspapers and other publications.
- Websites.

Let us sum up

For studying statistics, the first step is collection of data. It is systematic gathering of informations. Statistical surveys are tools of data collection. Data means information regarding a variable. The variable may be qualitative and quantitative or it may be discrete or continuous. The two survey techniques are Census and sampling. Depending on the source of information, data can be classified as primary or secondary. The important primary data collection methods are direct personal investigation, indirect oral investigation, direct observation, telephone interview, mailed questionnaire or schedule sent through enumerators and focus group discussion. Any published or unpublished data which are collected by a third party, and now used by the investigator for his purpose is a secondary data. Better trained persons are required for converting the secondary data to the required form.

Learning outcomes

After transaction of this unit, the learner:-

- differentiates population and sample.
- recognises investigator, investigation, enumerator and enumeration.
- classifies variables and constants.
- distinguishes qualitative variables and quantitative variables.
- differentiates discrete and continuous variables.
- compares primary and secondary data.
- identifies questionnaire and schedule.
- constructs/drafts questionnaire.
- explains different methods of data collection.
- recognises the sources of secondary data.

Evaluation Items

1. Data that can be classified according to colour. They are measured on .....scale
   a) Nominal b) Ordinal c) Cardinal

2. The group of all subjects under study is called .....
3. The representative part of a population is called .....
4. The number of days of absence of a worker per year is .....type of data
   a) Nominal b) Qualitative c) Discrete d) Continuous
5. The pre-test with the questionnaire before conducting a survey is called .....
6. The blind population of India constitute:
   a) a hypothetical population b) a sample c) an infinite population d) a finite population
7. Which of the following represents data
   a) a single value b) only two values in a set c) a group of values in a set d) all the above
8. Statistics deal with
   a) qualitative information b) quantitative information c) both (a) and (b) d) none of (a) and (b)
9. Data taken from a publication Agricultural situation in India will be considered as
   a) primary data b) secondary data
10. Mailed questionnaire method of enquiry can be adopted if respondents
    a) live in cities b) have a high income c) are literate
11. A study based on complete enumeration is known as:
    a) sample survey b) pilot survey c) none of the above
12. Statistical data are collected for.
    a) no purpose b) a given purpose c) any purpose
13. A statistical population may consists of
    a) an infinite number of items b) a finite number of items c) either of (a) and (b) d) none of (a) and (b)

14. Compare Primary and Secondary data
15. Distinguish between a questionnaire and a schedule.
16. What kind of data you receive when you are told about
    (a) blood type b) house hold c)heights of waterfall
17. What are points to be remembered while drafting a questionnaire?
18. What are the points to be remembered while collecting secondary data?
19. Explain the various primary data collection methods.
20. What are the important sources of secondary data?
21. Find the discrete data and continuous data from the following list
    a. Number of shares sold each day in a stock market
    b. Temperature recorded every half an hour at a weather bureau
    c. Life time of television tubes reduced by a company
    d. Yearly income of employees in a company
    e. The age of an individual
    f. Number of petals a flower has.
22. Categorize the data obtained in the following situations as quantitative and qualitative
    a. Political preference of a group of a people
    b. Family size (Number of members of a family) of hundred families in a township
    c. IQ score of plus one students undergoing state syllabus in Thiruvananthapuram district d. Academic qualification of a group of unemployed youth in a city
23. Which of the following constitute finite or infinite population
    a. Population consisting of odd integers.
    b. Weight of 200 new born babies in a hospital.
    c. Height of fifteen year old children in a school.
    d. Number of head and tail when a coin is tossed.

24. Categorize the following under cardinal, nominal or ordinal
    a. Telephone Numbers
    b. Roll numbers given to students in a class
    c. Ranks given to a class after a test
    d. Respondents attitude towards a newly designed project in an institution on a five point scale such as 1=strongly opposed, 2= may be opposed, 3= not strongly favoured, 4= may be favoured and 5= very strongly favoured
    e. The quantity of water in a reservoir measured in every half an hour.
    f. The price of furniture exhibited in a shop.

25. It is proposed to conduct a survey to obtain information on the study habits of Higher Secondary students in Kannur district and also the facilities available to them. Prepare a questionnaire for this purpose

26. A survey is to be carried out amongst school children to study how they spent time after school hours. Prepare a questionnaire for that purpose

27. Indicate whether the following statements are true or false. If false correct the statements
    a. Secondary data are generally used in those cases where the primary data do not provide an adequate basis for analysis.
    b. Secondary data does not need much scrutiny and should be accepted at its face value
    c. The task of editing secondary data is a highly specialised one
    d. The questionnaire requires a pre-testing before putting into practice

28. Which type of study do you prefer in the following cases? (Census or Sampling). Give reason
    a. The effect of a medicine
    b. To study about the wage distribution of 250 employees in a company
    c. A study on the roll of media in the marketing of a face cream
    d. A study of a patients heart beat who is admitted to ICCU of a hospital
    e. A study on the number of petals of a flower of a special kind

f. The life span of an electric bulb.

29. Which of the primary data collection method do you suggest in the following situations?. Give reasons
a. You are appointed as marketing manager of a company. The company introduces a washing machine with many options. You are asked by your employer to prepare a datasheet regarding the opinion of your customers about the new equipment
b. To prepare a report for a media on Nehru Trophy Vallamkali in the current year.
c. To introduce shift sessions in an institution

30. As a reporter of a certain media, you got an opportunity to interview an IAS topper. Which type of data collection method will you use?. List out other primary data collection methods.

Answers:

1)a 2)population 3)sample 4)discrete 5)pilot survey 6)d 7)c
8)c 9)b 10)c 11)c 12)b 13)c

22)(a) qualitative (b) quantitative (c) quantitative (d)qualitative

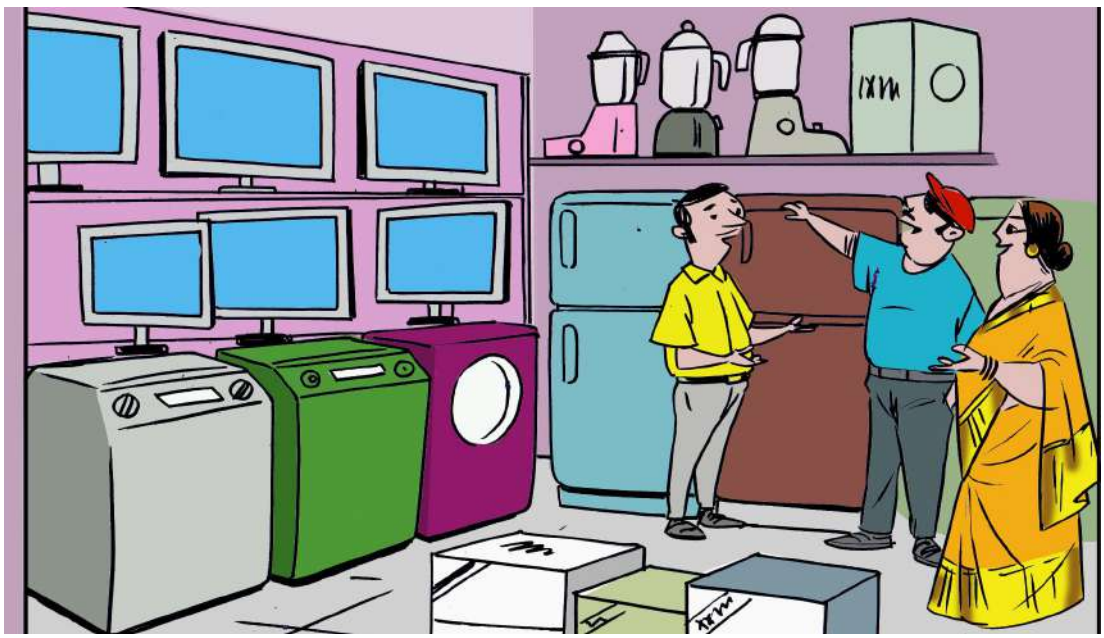
23)(a) infinite (b) finite (c) finite (d) finite 24(a)nominai
(b) ordinal (c)ordinal (d) nominal (e) cardinal (f) cardinal
27 (a) false (b) true (c)true (d) true

3 Classification and Tabulation

Introduction

In the previous chapter we learnt how the data is collected from the source. The collected data is usually contained in schedules or questionnaires. It is called raw data. Arriving at a conclusion from the raw data is a difficult task, because they are always in an unorganised form. Therefore a proper organisation and presentation of data is required for the systematic and comprehensive statistical analysis.

From a hypermarket we can easily select the required items, because of the items of same kind are arranged together. Otherwise it would be very difficult for the customer to search where the required items are stored. Likewise the raw data should be arranged in an organised manner and systematically to simplify the further statistical procedures. This process is called classification of data.



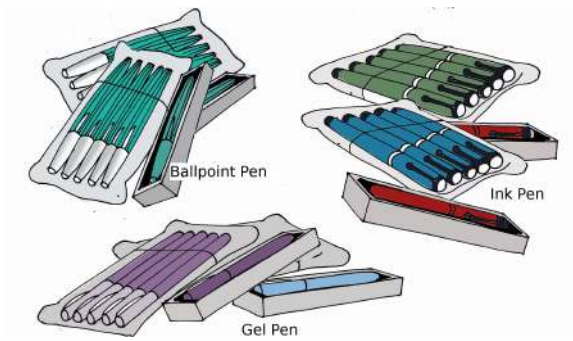
The process of grouping the data according to some characteristics is called Classification of Data.

3.1 Types of Classification

In some electronics hypermarkets we can see TVs, Refrigerators, Home Theatres, Washing Machines, Air Conditioners etc, arranged in separate section. In some other hypermarkets the item of same kind may be arranged in such a way that separate sections are allotted to each brand. In both cases the items are arranged according to some criterion or different types of classifications. In the first hypermarket the classification is according to the type of the item, whereas in the second hypermarket the classification is according to the brand of the product. Similarly the data is also classified in different ways.

Consider you are appointed as a salesman at a wholesale shop which deals in different kinds of pen. The shop manager shows a box containing packets of different types of pens, some packets contain a single pen and some others five. There are Ball Pens, Gel Pens and Ink Pens. Pens are made in various countries viz India, China, Thailand and Japan. The packing dates are also varied from February to June. You are asked to group same kind of pen together. How will you do the task?

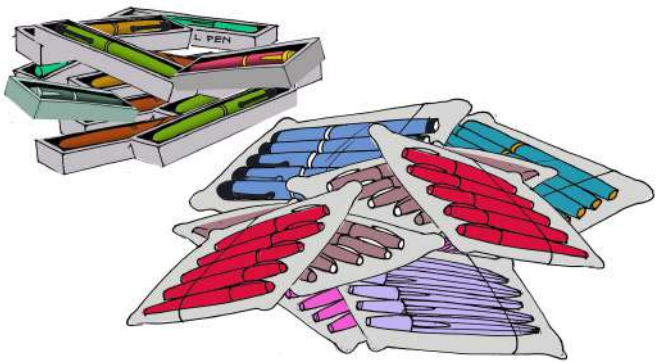
You can classify the pens according to their characteristic, such as ball point, gel or ink pens as shown in the picture. In other word you classified the pens according to their quality. This kind of classification is called Qualitative Classification.



Qualitative classification of pens

Classification based on the characteristics like Sex, Colour, Literacy, Religion,etc..., which cannot be numerically measured is called Qualitative Classification.

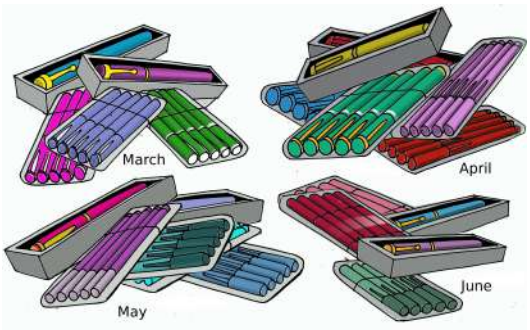
You can classify the pens with respect to the quantity of pens in the packet. If the pens are classified according to the quantity of pens contained in the packet, then it is known as Quantitative Classification.



Quantitative classification of pens

Classification based on the characteristics like Height, Weight, Thickness, Count, Area, Volume, etc.. which can be numerically measured is called Quantitative Classification

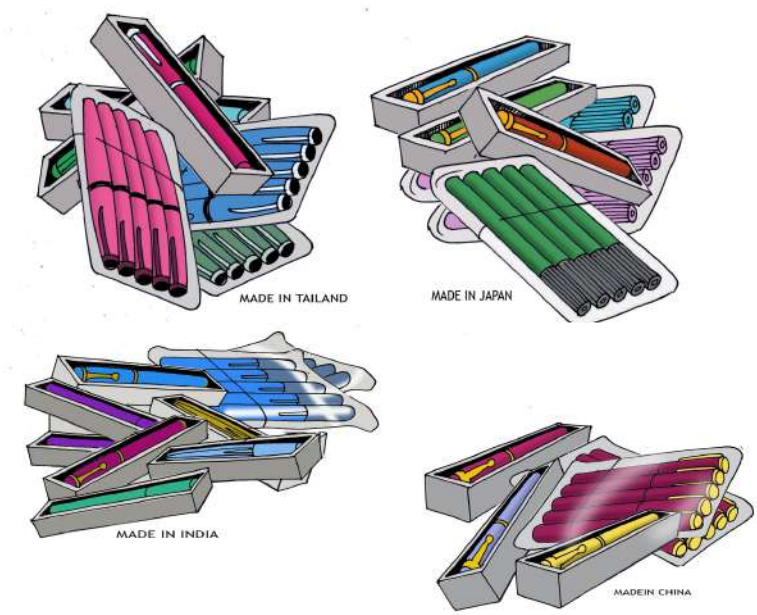
The pens may be classified according to the month of packing also. That is the pens packed in February, March, April, May and June are arranged separately. Then it is termed as Chronological Classification.



Chronological classification of pens

Classification based on some units of time like year, month, week, hour,etc, is called Chronological Classification

If you classify the pen according to the country where they are made, it is called Geographical Classification.



Geographical classification of pens

Classification based on place like Continents, Country, State, District, Village, etc is called Geographical Classification.

3.2 Tabulation of Data

We can present the classified items in different ways. Some of the methods are textual presentation, tabular presentation, diagrammatic presentation and graphical presentation.

In textual presentation, data is presented within the text. For example, you can present the data regarding the pens as shown below.

The box contains 260 packets of pens. 136 packets carry only one pen per pack.

124 packets contain 5 pens each. 101 packets of pen are made in India, 106 packets in China, 31 packets in Thailand and 22 packets in Japan. 81 packets contain ball pointed pens, 86 packets contain gel pens and 93 packets contain ink pens. 53 packets of pen are packed in February 50, 47, 47 and 63 packets were packed in March, April, May and June respectively. The box contains a total of 756 pens.

But it is not easy and suitable to present the data in this method when the quantity of data is too large or minute level classification is required. We cannot represent the data like, the number of packets containing 5 ink pens made in India and packed on June is 5, number of packets containing one ball pointed pen made in Thailand and packed on April is 2, etc.

Such details can be easily represented by a statistical table. The method of representing data with the help of a statistical table is called **Tabulation of Data.** In another words Tabulation is the way of systematic summarisation and presentation of information contained in the given data, in rows and columns. Tabular representation of data facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation. In fact tabulation is the final stage in collection and compilation of data and forms a gateway for further statistical analysis and interpretations as well as making the data suitable for further diagrammatic and graphic representation.

Type of Pen	Number of Packets
Ball Point	81
Gel Pen	86
Ink Pen	93
Total	**260**

Table 3.1: Qualitative Classification of data

The above table (Table - 3.1) represents the qualitative classification of data.Similarly the number of packets of pens can be presented in different tables according to different types of classification. See the following tables, Table - 3.2, Table -

3.3 and Table - 3.4.

Number of Pens per Packets	Number of Packets
1	136
5	124
Total	260

Table 3.2: Quantitative Classification of data

Made in	Number of Packets
India	101
China	106
Thailand	31
Japan	22
Total	**260**

Table 3.3: Geographical Classification of data

Month of Packing	Number of Packets
February	53
March	50
April	47
May	47
June	63
Total	**260**

Table 3.4: Chronological Classification of data

3.3 Objectives of Classification and Tabulation

Following are the objectives of classification and tabulation of data.

1. To simplify the complex data.
2. To facilitate comparison.
3. To facilitate statistical analysis.
4. To save time, space and energy.
5. To clarify similarity and dissimilarity.
6. To organise data logically and scientifically.
7. To grasp the information.

Activity

Collect tables published in journals, news papers and websites and identify the type of classifications applied in it.

3.4 One Way and Two Way Classification

In Table- 3.1, the classification is based on the characteristic, type of the pen. In Table - 3.2, Table - 3.3 and Table - 3.4 the classifications are based on number of pens per packet, month of packing and country respectively. In all these cases only one characteristic is considered for classification. So these classifications are called One-Way Classification of Data. The tables used to represent the one way classifications are called One Way Tables. If we consider two characteristics at a time for classification of data, it is termed as Two Way Classification of Data. The Table representing Two Way Classification is called **Two Way Table.** The following tables, Table - 3.5 and Table - 3.6 are examples of two way tables.

Month	Country				Total
	India	China	Thailand	Japan	
February	18	23	5	7	53
March	18	26	4	2	50
April	20	18	5	4	47
May	21	16	6	4	47
June	24	23	11	5	63
Total	101	106	31	22	260

Table 3.5: Two way table representing country and month of packing

No of Pens per Packet	Type of pen			Total
	Ball Point	Gel pen	Ink Pen	
ONE	38	49	49	136
FIVE	43	37	44	124
Total	81	86	93	260

Table 3.6: Two Way Table representing type of pen and number of pens per packet

Activity

Construct the skeleton of two way tables regarding the above problem. one example is given below.

Month	No. of Pens		Total
	ONE	FIVE	
February			
March			
April			
May			
June			
Total			

Similarly we can represent the data by considering more than two characteristics or attributes together for classification and tabulation of data. Table 3.7 is an example for such a classification.

Made in		India				China				Thailand				Japan				Total			
Packed on	Type	Ball Point	Jell Pen	Ink Pen	Total	Ball Point	Jell Pen	Ink Pen	Total	Ball Point	Jell Pen	Ink Pen	Total	Ball Point	Jell Pen	Ink Pen	Total	Ball Point	Jell Pen	Ink Pen	Total
Feb	1/Pack	2	3	3	**8**	3	5	7	**15**	0	0	1	**1**	0	3	1	**4**	5	11	12	**28**
	5/Pack	5	4	1	**10**	4	2	2	**8**	1	1	2	**4**	1	1	1	**3**	11	8	6	**25**
	Total	**7**	**7**	**4**	**18**	**7**	**7**	**9**	**23**	**1**	**1**	**3**	**5**	**1**	**4**	**2**	**7**	**16**	**19**	**18**	**53**
Mar	1/Pack	3	4	3	**10**	3	4	5	**12**	0	0	1	**1**	0	0	0	**0**	6	8	9	**23**
	5/Pack	2	2	4	**8**	4	4	6	**14**	0	1	2	**3**	1	0	1	**2**	7	7	13	**27**
	Total	**5**	**6**	**7**	**18**	**7**	**8**	**11**	**26**	**0**	**1**	**3**	**4**	**1**	**0**	**1**	**2**	**13**	**15**	**22**	**50**
Apr	1/Pack	3	4	5	**12**	2	2	5	**9**	2	0	0	**2**	2	0	0	**2**	9	6	10	**25**
	5/Pack	1	2	5	**8**	4	3	2	**9**	1	0	2	**3**	1	0	1	**2**	7	5	10	**22**
	Total	**4**	**6**	**10**	**20**	**6**	**5**	**7**	**18**	**3**	**0**	**2**	**5**	**3**	**0**	**1**	**4**	**16**	**11**	**20**	**47**
May	1/Pack	3	5	4	**12**	2	3	2	**7**	1	1	0	**2**	1	1	0	**2**	7	10	6	**23**
	5/Pack	3	3	3	**9**	3	4	2	**9**	2	0	2	**4**	2	0	0	**2**	10	7	7	**24**
	Total	**6**	**8**	**7**	**21**	**5**	**7**	**4**	**16**	**3**	**1**	**2**	**6**	**3**	**1**	**0**	**4**	**17**	**17**	**13**	**47**
Jun	1/Pack	5	4	5	**14**	5	6	4	**15**	1	3	2	**6**	0	1	1	**2**	11	14	12	**37**
	5/Pack	2	3	5	**10**	3	4	1	**8**	2	2	1	**5**	1	1	1	**3**	8	10	8	**26**
	Total	**7**	**7**	**10**	**24**	**8**	**10**	**5**	**23**	**3**	**5**	**3**	**11**	**1**	**2**	**2**	**5**	**19**	**24**	**20**	**63**
Total	1/Pack	16	20	20	**56**	15	20	23	**58**	4	4	4	**12**	3	5	2	**10**	38	49	49	**136**
	5/Pack	13	14	18	**45**	18	17	13	**48**	6	4	9	**19**	6	2	4	**12**	43	37	44	**124**
	Total	**29**	**34**	**38**	**101**	**33**	**37**	**36**	**106**	**10**	**8**	**13**	**31**	**9**	**7**	**6**	**22**	**81**	**86**	**93**	**260**

Table 3.7: Detailed classification of pens

3.5 Parts of a Table

The structure and parts of a table is depended on the nature of the data and purpose of our study. However in general a statistical table is divided into eight parts, which are explained below.

a) **Table Number**

Each table should have a table number to identify the table and for further references. The table number is usually given on the top of the table or on the left hand side along with the title of the table. Sometimes table number is given in the bottom centre of the table.

b) **Title of the table**

The title of a table is generally given at the top or bottom of the table in the centre, along with the table number or just below the table number. A good title will contain a brief statement about the nature, geographical

region, time span, etc. to which the data relate. This gives fairly good information about the content.

c) **Captions**

Caption refers to the headings of the vertical columns. A caption generally has a main heading and a number of small sub headings. A caption should be written in unambiguous terms and placed at the middle of the column. The unit of measurement may also be mentioned along with the caption.

d) **Stubs**

Stubs refer to the headings of the horizontal rows and they are usually written in the left hand side of the rows. The number of stubs is depending upon the nature of the data.

e) **Body**

The most vital part of the table is body of the table. It contains statistical data arranged according to captions and stubs. Usually the content of the body is numerical information in the cells, mostly frequencies or observations. The column totals, row totals and grand total should also be mentioned in the body of the table.

f) **Head Note**

Head notes are usually given on the right top corner of the table, just below the title. It refers to the data contained in the body of the table which is not mentioned in the title. For example the units of measurements are generally written as head notes like in lakhs or in kilogram, etc. Generally it is put in a bracket.

g) **Foot Note**

Foot notes are given below the table and are meant to clarify anything which is not clear from the title, Caption, Stubs, etc.

h) **Source Note**

The source of the data given in the table is to be disclosed as a rule. This enables one to verify all facts about the data. It is given below the foot note.

All the parts may not be compulsory for a table.

Table Number:

Title:

Head Note:

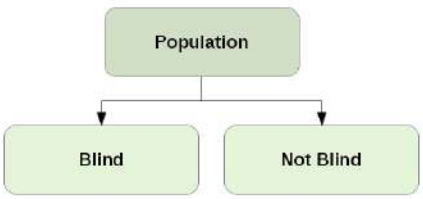
Stubs	Caption
	B_1 B_2 ..... B_n
A_1 A_2 .. .. .. A_m	BODY

Foot Note:

Source Note:

3.6 Classification according to Attributes

In qualitative analysis, data is classified on the basis of descriptive characteristics or on the basis of attributes like sex, literacy, religion, education etc. which cannot be numerically measured. We can only find out the presence or absence of a particular attribute in an individual. If we are dealing with the problem of blindness, we can only find out whether the individual is blind or not blind. We cannot measure the blindness. If we are dealing with the problem of deafness, we can only find out whether the individual is deaf or not. We cannot measure the deafness. In such cases the attribute divides the population into two parts. One in which the attribute is present and the other in which the attribute is not present. In the case of blindness there are only two classes, that is, those who are blind and those who are not blind. These classes are mutually exclusive (disjoint), so that those who are blind cannot come into the category of those who are not blind. Such a classification is called **Dichotomy**

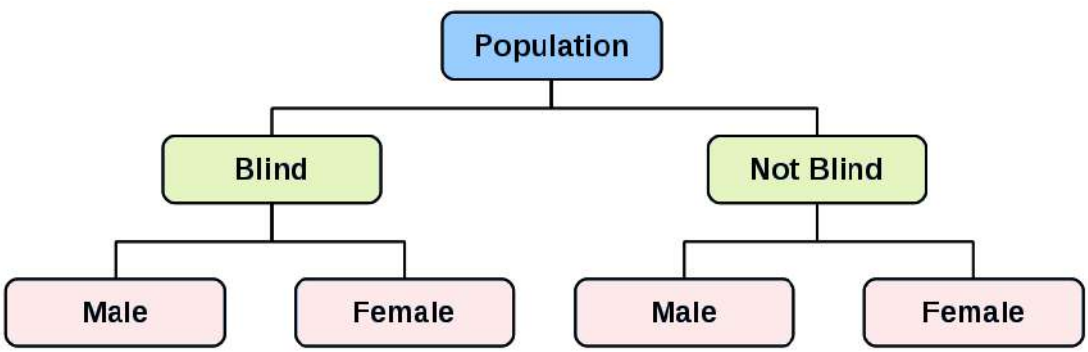


or **Two-fold Classification.**

The following table is an example of two-fold classification

Course	Number of students		
	Passed	Failed	Total
Humanities	29	18	47
Commerce	83	16	99
Science	96	4	100
Total	208	38	246

If we consider another attribute male or female together with the above attribute blindness there will be four classes. Males who are blind, males who are not blind, females who are blind and females who are not blind. The classification can be further extended if we have a third attribute, say nationality. Then we will get $2^3 = 8$ mutually exclusive classes. Such classification in which more than one attribute is taken into account is called **Manifold Classification.**



The following table is an example of manifold classification

Course		Number of students		
		Passed	Failed	Total
Humanities	Boys	11	16	47
	Girls	18	2	
	Total	29	18	
Commerce	Boys	32	11	99
	Girls	51	5	
	Total	83	16	
Science	Boys	42	3	100
	Girls	54	1	
	Total	96	4	
Total	Boys	85	30	246
	Girls	123	8	
	Total	208	38	

Illustration 3.1

In a sample study about coffee habits in two towns, the following information was received.

Town X: Females were 40%, Total coffee drinkers were 45% and male non coffee drinkers were 20%.

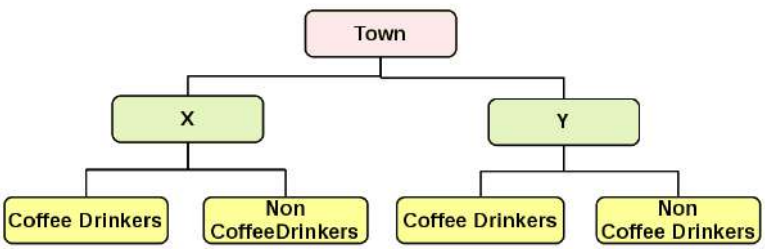
Town Y: Males were 55%, male non coffee drinkers were 30% and female non coffee drinkers were 15%.

a) Draw the classification sketch of the above data

b) Present the above data in tabular form and complete the table.

Solution:

a).



b).

Persons	*Town X*			*Town Y*		
	Male	*Female*	*Total*	*Male*	*Female*	*Total*
Coffee Drinkers	*40*	*5*	*45*	*25*	*30*	*55*
Non Coffee Drinkers	*20*	*35*	*55*	*30*	*15*	*45*
Total	*60*	*40*	*100*	*55*	*45*	*100*

Frequency

The number of repetitions of a particular observation in a series is called **Frequency of the observation.** For example if in a class 15 students scored 40 marks, then the frequency of the mark 40 is 15.

3.7 Frequency Tables

The series of observations in which items are listed individually is called **Raw data** or **Individual series.** Usually it will be in an ungrouped format. The number of family members of 60 students in a class is simply listed below. It is an example of individual series or ungrouped data.

4	3	3	2	3	4	5	7	10	6	3	9	6	3	4	5	5	5	8	7
8	6	9	7	2	7	6	5	11	4	4	3	5	6	6	7	6	5	7	6
4	5	6	10	8	6	5	4	5	6	4	5	5	6	5	4	6	7	5	8

The above series is arranged in the order of serial numbers. It means no organisation is made on the data. The first step in descriptive statistics is organisation of data. The above data can be arranged in ascending or descending order of the observations as you learnt in your lower classes. Try it yourself.

The arrangement of data in ascending or descending order is more helpful for a statistical study than the individual series as it is. Though it is not suitable for our statistical analysis and graphical presentation. Thus we can arrange the observations and its frequencies in a table to make the data more compact and

useful. Such tables are called frequency tables. The frequency table prepared based on the previous data of the number of family members is shown below.

Number of family members	Tally Mark	Frequency
2	II	2
3	𝍸 I	6
4	𝍸 IIII	9
5	𝍸 𝍸 IIII	14
6	𝍸 𝍸 III	13
7	𝍸 II	7
8	IIII	4
9	II	2
10	II	2
11	I	1
Total		60

Table 3.8:

Here the number of family members is a discrete variable and exact measurements of units are clearly shown in the first column. This type of frequency table is called **Discrete Series** , **Discrete Frequency Table** or **Frequency Array.**

A Discrete Frequency Table is that series in which data are presented in a way that exact measurements of units are clearly shown.

Know your progress

The marks obtained by 48 students of a class in a class test on statistics with maximum score 10 are given below. Construct a frequency table.

5	10	5	6	6	6	7	7	2	7	8	8	7	8	10	6	8	7
3	5	5	5	5	7	8	5	2	6	8	5	1	7	6	9	6	7
4	8	4	6	4	3	6	7	9	2	7	10						

When the body weight of 60 students is collected, it will be very difficult to tabulate as before, since the body weight is a continuous variable and may take infinitely many possible values. Following are the body weights of the students studying in graduate courses. The measures are in kilograms .

51.1, 53, 66.6, 64.2, 63.9, 54.1, 69.7, 56.4, 52.9, 52.3, 59.5, 57.1, 45.5, 73.4, 41.4, 57.5, 62.4, 43.4, 38.4, 69.5, 48.5, 48.9, 45.2, 69.7, 36.4, 36.4, 58.8, 64.5, 66.4, 47.8, 54.7, 49, 49.9, 46.2, 53.2, 56.2, 32.8, 59.4, 31.8, 53.1, 50.2, 57.3, 30.1, 46.1, 41.8, 60.5, 40.2, 45.4, 47.1, 44.1, 52.6, 64.1, 38.3, 79.2, 48.5, 53.4, 51.6, 51.3, 63.5, 37.8 Clearly the body weight varies from 30.7 to 78.3. Here we cannot tabulate by taking the exact values in the first column and its corresponding frequencies in the second column. Instead of that we can find the frequency of observations falling in an interval of weights like 30-35, 35-40, 40-45, etc. It is shown in the table.

Weight	Tally Mark	Frequency
30-35	III	3
35-40	𝍸	5
40-45	𝍸	5
45-50	𝍸 𝍸 II	12
50-55	𝍸 𝍸 III	13
55-60	𝍸 III	8
60-65	𝍸 II	7
65-70	𝍸	5
70-75	I	1
75-80	I	1
Total		60

Table 3.9:

Here the intervals 30-35, 35-40, 40-45, etc. are usually called classes. 30 is the lower limit and 35 is the upper limit of the class 30-35. The difference between the lower limit of a class and lower limit of the next class or difference between the upper limit of a class and upper limit of the previous class is called **Class width.** Here 35-30=5 is the class width of first class. Frequency tables with classes and corresponding frequencies are known as **Continuous Frequency Distribution** or simply **Frequency Distribution.** Continuous Frequency Distribution are used for both discrete and continuous data by assuming the data is continuous.

Open Ended Classes:

If the lower limit of the first class or upper limit of the last class is not specified, then it is called open ended class. Sometimes we may not be able to determine the lower limit of the first class or upper limit of the last class or both. In such situations we can use open ended classes.

Illustration 3.2

Q) The marks obtained by 50 students in an examination with maximum score 100 are given below. Construct a frequency table to the data.

32	*45*	*67*	*44*	*25*	*53*	*54*	*37*	*42*	*48*	*46*	*59*	*72*
36	*46*	*58*	*65*	*68*	*78*	*57*	*68*	*43*	*23*	*45*	*58*	*12*
71	*36*	*55*	*61*	*58*	*23*	*34*	*25*	*17*	*61*	*70*	*56*	*43*
59	*40*	*14*	*39*	*60*	*41*	*56*	*65*	*33*	*29*	*58*		

Here the marks obtained by the students is a discrete data. The minimum and maximum marks are 12 and 78 respectively. So representing individual marks and its frequencies is not practical in this situation. So we consider the data is continuous and constructs a continuous frequency distribution as shown below.

Mark	Tally Mark	Frequency
10-19	𝍸	3
20-29	𝍸	5
30-39	𝍸 II	7
40-49	𝍸 𝍸 I	11
50-59	𝍸 𝍸 II	12
60-69	𝍸 III	8
70-79	IIII	4
Total		50

Table 3.10:

Know your progress

The temperature in a village is observed at randomly selected days during a year. Tabulate the data

33.98, 29.07, 29.90, 34.15, 28.98, 36.02, 32.77, 33.11, 27.55, 34.77, 39.76, 30.65, 29.15, 27.82, 33.67, 34.31, 33.18, 34.65, 34.76, 24.03, 27.14, 37.43, 23.88, 34.27, 27.25, 25.57, 35.86, 34.98, 36.75, 30.28, 33.10, 25.67, 29.97, 34.17, 27.44, 27.61, 35.47, 29.28, 28.56, 27.46, 29.12, 34.69, 25.07, 34.05, 35.00, 30.28, 37.65, 35.08, 33.03, 33.84, 34.65, 33.86, 36.76, 26.04, 35.20, 29.89, 27.14, 29.54, 32.25, 28.69, 32.23, 25.41, 27.27, 27.44, 37.76, 29.32, 31.11, 27.02, 36.93, 32.54, 37.31, 34.87, 31.61, 37.01, 23.08, 34.21, 31.48, 37.75, 24.75, 24.19, 33.69, 32.86, 28.46, 27.77, 37.14, 30.24, 24.45, 31.46, 29.80, 40.42

Inclusive and Exclusive Classes

An inclusive class includes all items between the lower limit and upper limit, including the limits. But the upper limits are not included in the exclusive classes. In an exclusive series the upper limit of one class is the lower limit of the next class. But in an inclusive series the upper limit of a class does not repeat itself as the lower limit of the next class. Usually inclusive classes are used only for the discrete data. Exclusive classes can be used for both discrete and continuous data. Table - 3.9 is an example of exclusive classes and Table - 10 is an example of inclusive classes.

An inclusive class can be converted to exclusive class by modifying its class limits. Let us convert the class intervals given in Table - 10 to exclusive classes as shown in the table.

Mark	9.5 - 19.5	19.5 - 29.5	29.5 - 39.5	39.5 - 49.5	49.5 - 59.5	59.5 - 69.5	69.5 - 79.5	Total
Frequency	3	5	7	11	12	8	4	50

Table 3.11:

Unequal Class Intervals:
The class widths need not be equal for all the classes. But it is better to use equal class intervals.

Note:-

The total number of observations in an individual series is usually represented by n and the total frequency of a frequency table is usually represented by N.

Relative Frequency Tables

The ratio of frequency to the total frequency is called **Relative Frequency.**

$$\text{Relative Frequecy} = \frac{frequency}{\text{Total frequency}} = \frac{f}{N}$$

The table representing relative frequency of observations (or classes) is called **Relative frequency tables.** The relative frequency table prepared from Table - 3.8 is shown below.

Number of family members	Frequency	Relative Frequency
2	2	2/60=0.03
3	6	6/60=0.10
4	9	9/60=0.15
5	14	14/60=0.23
6	13	13/60=0.22
7	7	7/60=0.12
8	4	4/60=0.07
9	2	2/60=0.03
10	2	2/60=0.03
11	1	1/60=0.02
Total	60	60/60=1

Table 3.12:

Know your progress

- Prepare a relative frequency table from Table 3.9
- A special diet is practiced for three months by 35 working women in an IT company to reduce their body weight. The loss of weight (in Kgs)occured is given below. Prepare a relative frequency table to the data.

4	2	3	2	6	4	3	2	7	3
4	4	3	3	2	1	5	2	1	5
2	3	2	5	4	0	3	2	1	4
3	5	1	1	3					

Percentage Frequency Tables

In a percentage frequency table the percentage of the total frequency corresponding to the observations(or classes) are shown. The percentage frequencies are obtained by the following relationship

$$\text{Percentage Frequency} = \frac{frequency}{\text{Total frequency}} \times 100 = \frac{f}{N} \times 100$$

Note:

$$\text{Percentage Frequency} = \text{Relative Frequency} \times 100$$

We can prepare a percentage frequency table from Table -3.9 as shown below.

Weight	Frequency	Percentage Frequency
30 - 35	3	$\frac{3}{60} \times 100 = 5.0$
35 - 40	5	$\frac{5}{60} \times 100 = 8.33$
40 - 45	5	$\frac{5}{60} \times 100 = 8.33$
45 - 50	12	$\frac{12}{60} \times 100 = 20.0$
50 - 55	13	$\frac{13}{60} \times 100 = 21.67$
55 - 60	8	$\frac{18}{60} \times 100 = 13.33$
60 - 65	7	$\frac{17}{60} \times 100 = 11.67$
65 - 70	5	$\frac{3}{60} \times 100 = 5.0$
70 - 75	1	$\frac{1}{60} \times 100 = 1.67$
75 - 80	1	$\frac{1}{60} \times 100 = 1.67$
Total	60	$\frac{60}{60} \times 100 = 100$

Table 3.13:

Know your progress

1. Prepare a percentage frequency table to the data contained in Table - 3.8.
2. A researcher decided to study the impact of drug abuses on the road accidents. He collected the number of accidents reported per day at the police station, due to the vehicles driven by drunkards and drug abusers. Data is given below. Construct a percentage frequency table.

5	2	2	0	1	4	6	2	2	1	5	4
2	5	0	0	1	3	0	3	3	1	1	1

The sum of relative frequencies is always 1 and the sum of percentage frequencies is 100.

Cumulative Frequency Tables

The number of observations less than or equal to a particular value is called **Less than cumulative frequency** of that value. Similarly the number of observations greater than or equal to a particular value is called **Greater than cumulative frequency** or **More than cumulative frequency** of that value. Consider the Table-3.8 The less than cumulative frequency of the observation 3 is 8, since 8 families have less than or equal to 3 family members. The less than cumulative frequency of the observation 5 is 31. The greater than cumulative frequency of the observation 9 is 5, since only 5 families have 9 or more than 9 family members. The tables representing cumulative frequencies are called Cumulative Frequency Tables. The preparation of a Less than Cumulative Frequency Table is demonstrated below.

Number of family members	Frequency	Less than Cumulative Frequency
2	2	2 + 0=2
3	6	6 + 2=8
4	9	9 + 8=17
5	14	14 + 17=31
6	13	13 + 31=44
7	7	7 + 44=51
8	4	4 + 51=55
9	2	2 + 55=57
10	2	2 + 57=59
11	1	1 + 59=60
Total	60	

Table 3.14:

Illustration 3.3

Construct a less than cumulative frequency table to the Table - 3.9

Weight	Upper Bounds	Frequency	Less than Cumulative Frequency
30 - 35	35	3	3
35 - 40	40	5	8
40 - 45	45	5	13
45 - 50	50	12	25
50 - 55	55	13	38
55 - 60	60	8	46
60 - 65	65	7	53
65 - 70	70	5	58
70 - 75	75	1	59
75 - 80	80	1	60
Total	-	60	-

Table 3.15:

Illustration 3.4

Construct a greater than cumulative frequency table for Table - 3.8.

Number of family members	Frequency	Greater than Cumulative Frequency
2	2	60
3	6	60-2=58
4	9	58-6=52
5	14	52-9=43
6	13	43-14=29
7	7	29-13=16
8	4	16-7=9
9	2	9-4=5
10	2	5-2=3
11	1	3-2=1
Total	60	

Table 3.16:

Illustration 3.5

Construct a Greater than Cumulative Frequency table to the Table - 3.9

Weight	Lower Bounds	Frequency	Greater than Cumulative Frequency
30 - 35	30	3	60
35 - 40	35	5	57
40 - 45	40	5	52
45 - 50	45	12	47
50 - 55	50	13	35
55 - 60	55	8	22
60 - 65	60	7	14
65 - 70	65	5	7
70 - 75	70	1	2
75 - 80	75	1	1
Total	-	60	-

Table 3.17:

Know your progress

1. The departure time of the train Mangala-Lakshadweep Express at the Kozhikode Railway Station is observed for three months. The following frequency table shows the number of minutes late by the train on these days. Construct a less than cumulative frequency table to the data. Determine the number of days in which the train was not late for more than five minutes.

Late (in Minutes)	0	1	2	3	4	5	6	7	8	9	10
No. of days	44	13	9	5	5	3	4	2	3	1	1

2. Prepare a greater than cumulative frequency table to the frequency table given in the above table and obtain the number of days in which train was late by at least two minutes.
3. The Compact Fluorescent Lamps (CFL) produced by a company is inspected to study about the life span of the lamps and the following frequency table is prepared.

Life in Hours ('00)	0-5	5-10	10-20	20-25	25-30	30-35	35-40
No. of CFLs	3	16	160	323	80	17	1

Compute the less than cumulative frequencies and obtain the number of CFLs damaged within 2500 hours.

4. Calculate the greater than cumulative frequencies to the frequency table given in the above table. Estimate the number of CFLs illuminated for at least 2000 hours.

3.8 Bivariate Frequency Distribution

If only one characteristic of the sampling units is measured for the study, it is called **Univariate Data.** If two characteristics are measured simultaneously from each unit, it is known as **Bivariate Data.** Similarly data containing measurements of more than two characteristics of each unit is called **Multivariate Data.** For example if only the height of the students is measured for the study, it is Univariate Data. Usually we represent it by x, y ,z, etc.

Height (in inches) : 52, 51, 57, 62, 68

If we measure the height and weight of each student for a study, it is a Bivariate Data. We represent it by (x, y) or (x_1, y_1) where first variable is the height and the second variable is the weight.

Height in inches and weight in Kg:

(52, 45), (51, 62), (57, 58), (62, 70), (68, 73)

The same data can also be represented as,

Height (x)	:	52	51	57	62	68
Weight (y)	:	45	62	58	70	73

The frequency distribution table of a Bivariate Data is called **Bivariate Frequency Table.**

Illustration 3.6

The heights (in inches) and weights (in Kg) of 40 students in a class are given. Construct a frequency table to the data.

(59, 60)	(62, 70)	(67, 65)	(72, 80)	(71, 58)	(68, 73)	(54, 42)	(59, 55)
(55, 53)	(60, 55)	(54, 42)	(65, 72)	(62, 71)	(69, 82)	(65, 47)	(68, 64)
(65, 74)	(64, 84)	(67, 69)	(72, 75)	(64, 65)	(71, 78)	(70, 74)	(67, 62)
(60, 57)	(59, 48)	(67, 71)	(60, 65)	(56, 49)	(63, 62)	(71, 69)	(58, 53)
(67, 62)	(57, 62)	(57, 55)	(62, 64)	(66, 73)	(66, 53)	(69, 72)	(56, 44)

Solution:

Height / Weight	54 - 58	59 - 63	64 - 68	69 - 73	Total
36-45	3				3
46-55	3	4	2		9
56-65	1	5	5	1	12
66-75		2	6	4	12
76-85			1	3	4
Total	7	11	14	8	40

The frequencies given in the last row and last column are called marginal frequencies. Frequencies in the last row are the marginal frequency of heights and those in the last column are the marginal frequency of weights. The marginal frequency tables of heights and weights are shown below.

Marginal frequency table of Height

Height	54-58	59-63	64-68	69-73	Total
Total	7	11	14	8	40

Marginal frequency table of weight

Weight	36-45	46-55	56-65	66-75	76-85	Total
Frequency	3	9	12	12	4	40

Know your progress

The amount spent for advertisement (In Lakh Rupees) and the profits (in Crore Rupees) of a company is observed for different months and the following data is obtained. Prepare a bivariate frequency table to the data. Also obtain the marginal frequency

tables.

(1, 13)	(3, 15)	(2, 16)	(6, 17)	(13, 23)	(12, 24)	(17, 25)	(11, 19)
(15, 23)	(18, 27)	(19, 28)	(18, 24)	(16, 16)	(13, 27)	(7, 15)	(8, 19)
(3, 14)	(4, 17)	(7, 21)	(19, 33)	(16, 35)	(17, 27)	(10, 21)	(11, 17)
(10, 16)	(13, 19)	(11, 18)	(12, 19)	(14, 21)	(16, 25)	(19, 24)	(20, 23)

3.9 Advantages of Tabulation

Following are some of the advantages of tabulation of data

1. Tables consolidate the data
2. A table presents the data in such a simple a3333nd systematic form that one can understand the values associated with the variable and / or attributes.
3. They reveal the association of a variable or attribute with the other.
4. Comparison of one factor with the another becomes easy and reliable.
5. Diagramatic and graphical presentation of data becomes simple and accurate with the help of tables.
6. Tables portray more information in less space.
7. Tables are time savers. Information can easily be gathered from a table.
8. Tables are the basis of statistical calculations and analysis of data.
9. Tables make interpretation of data easier and better as compared to raw data.

Let us sum up

In this Chapter we have discussed the need, objectives and advantages of Classification and Tabulation of data in statistics. It is the process of making the data ready for the analysis. We can classify the data under different ways called Qualitative, Quantitative,

Chronological and Geographical classifications. The process of arranging the classified data in columns and rows is called Tabulation. The tables representing frequencies of observations are called Frequency Tables. Relative frequency table, percentage frequency table and cumulative frequency table are some of the different types of statistical tables.

Learning outcomes

After transaction of this unit, the learner:-

- identifies the need of Classification and Tabulation.
- recognises the different methods of Classification and Tabulation.
- classifies raw data to useful information
- constructs frequency tables.
- interprets the data.

Evaluation Items

Choose the correct answer

1. The process of grouping the data according to their characteristics is called .....
   a) Organisation of data b) Classification of data
   c) Tabulation of data d) Presentation of data

2. Classification based on time is called .....
   a) Qualitative Classification b) Quantitative Classification
   c) Chronological Classification d) Geographical Classification

3. Tabulation of data saves .....
   a) Time b) Energy c) Space d) All the above

4. Classification of data to two disjoint classes is called .....
   a) One - Fold Classification b) Two - Fold Classification
   c) Many - Fold Classification d) None of these

5. The number of observations greater than a particular value is the .....
   a) Cumulative frequency b) Less than Cumulative frequency
   c) Relative frequency d) More than Cumulative frequency

6. An exclusive class excludes .....
   a) Lower limit b) Upper limit
   c) Both the limits d) Either of the limits

7. The less than cumulative frequencies of 7^{th} and 8^{th} observations are 32 and 84 respectively. then the frequency of 8^{th} observation is .....
   a) 52 b) 116 c) 42 d) 64

8. Relative frequency = ...
   a) $f \times N$ b) $\frac{f}{N}$ c) $f - N$ d) $f + N$

9. Percentage frequency = ...
   a) $f \times 100$ b) $\frac{f}{100}$
   c) $RelativeFrequency \times 100$ d) $\frac{RelativeFrequency}{100}$

10. A relative frequency table is prepared for 120 observations. The relative frequency of 5^{th} item is 0.1. Then the frequency of 5^{th} item is .....
    a) 100 b) 50 c) 12 d)10

11. Explain the objectives of classification and tabulation of data?

12. Make a comparison between Classification and Tabulation?

13. What are the different types of Classifications?

14. State the advantages of Tabulation?

15. Explain the term dichotomy?

16. There are three streams in a Higher Secondary School, Humanities, Commerce and Science. Boys and Girls are studying with three different second languages, Malayalam, Hindi and Urdu.

a) Draw a Classification Sketch.

b) Draw the skeleton of the Table

17. 1200 employees are working in a company as Managers, Executives and Clerks. Draw the Classification Sketch of the data of the employees according to the attributes Designation, Nationality (Indians or Non Indians), Gender, Age Group, etc Also tabulate the data by giving arbitrary frequencies.

18. 110000 students are registered for different courses in a university during 2000 - 2002. 63000 students registered for Arts stream and 30000 students registered for Science stream. 17000 students registered for Arts, 10000 students for Science and 4000 for Law in the year 2000. 25000 students among the 43000 students registered at the university in the year 2002 were in the Arts Stream. The students registered for Law in 2002 was one thousand more than that in 2001. Tabulate the data.

19. The number of flowers obtained from different plants in a garden in a day is given below.

3	9	7	6	0	10	6	8	5	4
9	8	4	3	6	7	1	2	5	7
6	7	8	4	6	5	6	8	3	6
7	2	9	7	6	3	5	0	3	8
2	8	5	9	7	4	5	8	6	7
10	9	7	6	7	7	8	6	5	6
6	7	7	7	8	9	6	5	7	7

a) Construct a frequency table to the number of flowers per plant.

b) Prepare a less than cumulative frequency table.

c) Prepare a greater than cumulative frequency table.

d) Obtain a relative frequency table

e) Obtain the percentage frequency table.

20. The weights of six months old rabbits in a farm are measured in grams and the following data is obtained.

862 816 971 932 877 958 854 928 802 950 946 928
837 952 855 812 836 958 933 946 902 925 941 882
900 861 907 832 917 858 888 868 860 827 946 886
976 889 937 806 944 916 951 951 855 940 890 828
802 822 926 808 916 914 943 828 954 892 844 938
809 882 918 928 979 830 935 840 809 919 873 915
865 901 894 863 870 862 814 913 861 875 971 922
906 829 938 969 828 910 972 876 961 930 949 864
864 955 935 907 870 980 839 940 843 885 938 920
801 873 877 847 856 842 921 958 906 914 878 829
898 898 852 925 896 867 939 975 849 917 922 904
852 848 927 820 864 952 911 975 963 930 802 823
976 890 816 856 841 906 867 929 921 929 896 965
809 967 928 943 816 895 813 804 880 970 847 972

a) Construct a frequency table by taking a suitable class interval.
b) Prepare a less than cumulative frequency table.
c) Prepare a greater than cumulative frequency table.
d) Obtain a relative frequency table.
e) Obtain the percentage frequency table.

21. Following table shows the cumulative frequency table of profit and number of companies.

Profit less than (in 000s)	15	30	45	60	75	90	105	120	135
No. of Companies	3	10	28	53	73	85	91	96	98

Construct the frequency table.

22. The students in a Statistics class were trying to study the heights of participants in a sports meet. They collected the height of 20 participants, as displayed in the table.

Height (x)	49	53	54	55	66	70	80
No of participants (f)	1	2	4	5	3	2	1

Construct

a) Less than cumulative frequency table
b) Greater than cumulative frequency table
d) Relative frequency table
e) Percentage frequency table

23. The table below show the age of 55 patients selected to study the effectiveness of a particular medicine.

Age	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No of Patients	5	7	17	12	5	2	7

Construct

a) Less than cumulative frequency table
b) Greater than cumulative frequency table
d) Relative frequency table
e) Percentage frequency table

24. The frequency distribution below represents the weights in kg of parcels carried by a small logistic company.

Weight	No of Parcels
10.0 - 10.9	2
11.0 - 11.9	3
12.0 - 12.9	5
13.0 - 13.9	8
14.0 - 14.9	12
15.0 - 15.9	15
16.0 - 16.9	13
17.0 - 17.9	11
18.0 - 18.9	6
19.0 - 19.9	2

Prepare a frequency table with exclusive classes.

25. Thirty automobiles were tested for fuel efficiency and the following table is obtained.

Mileage (in KMs)	Below 13	Below 18	Below 23	Below 28	Below 33
No: of Vehicles	3	8	23	28	30

Construct a frequency table.

26. Following table shows the relative frequencies of 500 workers in a company with respect to their daily wages (in hundred rupees).

Daily wages	250 - 300	300 - 350	350 - 400	400 - 450	450 - 500	500 - 550	550 - 600
Relative freq.	0.05	0.1	0.3	0.225	0.2	0.1	0.025

Construct a frequency table.

Answers:

1) b	2) c	3) d	4) b	5) d
6) b	7) a	8) b	9) c	10) c

Introduction

In the previous chapter we have discussed the classification and tabulation that help in summarising the collected data and presenting them in a systematic manner. Although tabulation is a good technique to present the data, diagrams are advanced technique to represent the data. Diagrams and Graphs are the methods for simplifying the complexity of quantitative data.

4.1 Significance of Diagrams and Graphs

- It gives a clear picture of the data
- We can make comparison between different samples easily
- The technique can be used universally at any place at any time. This technique is used almost in all subjects and other various fields
- It is more attractive and impressive

Graphs originated when ancient astronomers drew the position of the stars in the heavens. Roman surveyors also used coordinates to locate landmarks on their maps. William Henry Playfair, Scottish engineer and Political economist, is known as the founder of Graphical methods of Statistics. He invented four types of diagrams the line graph, bar chart, pie chart and circle graph. He used graphs to present economic data pictorially.

4.2 Diagrams

Commonly used diagrams are

1. Bar diagrams
2. Pie diagram

Bar diagram

A Bar Diagram consists of a set of separated rectangles. Each rectangle is known as bar. There are four types of bar diagrams.

(a) Simple Bar Diagram

(b) Multiple Bar Diagram

(c) Sub Divided Bar Diagram (Component Bar Diagram)

(d) Percentage Bar Diagram

Simple Bar Diagram

A simple bar diagram is used to represent only one variable. It is constructed by horizontal or vertical bars with same width. It is used in qualitative and quantitative cases.

Illustration 4.1

The following table gives the number of students in Science, Humanities and Commerce streams of a school. Represent the data by simple bar diagram

	Stream	No. of Students
	Science	240
	Humanities	120
	Commerce	100

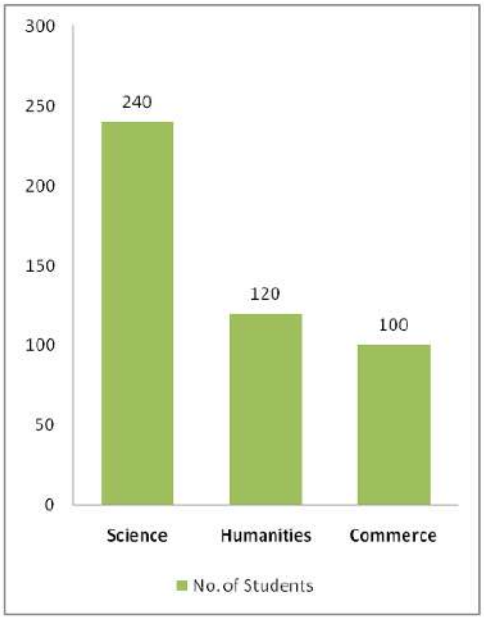
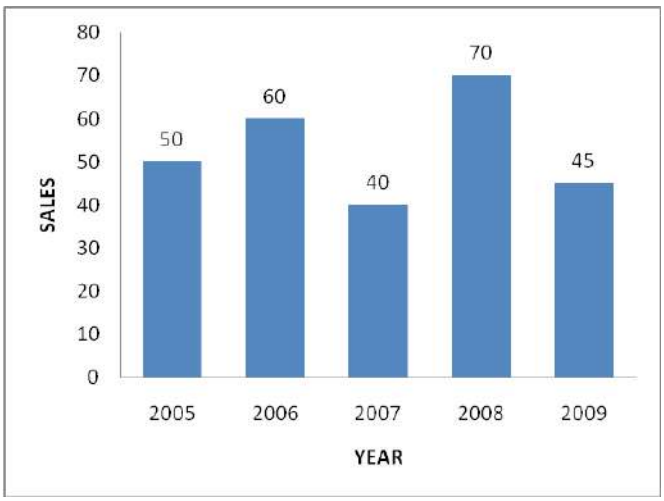


Illustration 4.2

Following table gives sales of a company for the last 5 years. Represent the data by simple bar diagram

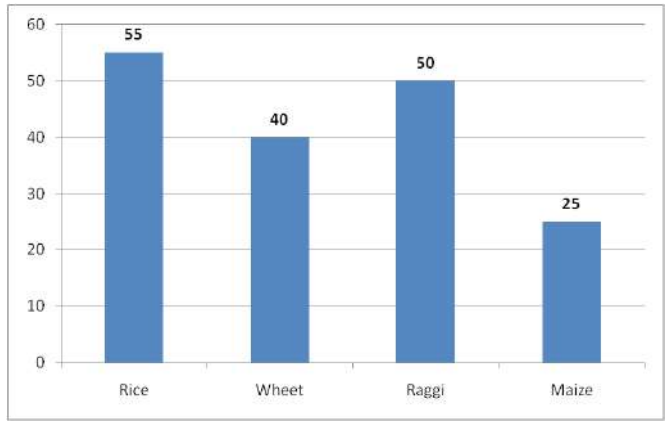
Year	*2005*	*2006*	*2007*	*2008*	*2009*
Sales in lakhs	*50*	*60*	*40*	*70*	*45*

Solution:



Know your progress

You are given a simple bar diagram representing data related to the production (in lakh tones) of different crops in India



Based on the above diagram, answer the following.

1. Which crop has recorded the highest production?
2. Represent the above diagram in tabular form
3. State the difference in production of the highest and the least production of crops

Activity

Collect the data of percentage results of HSE in last four years from your school and represent it by a simple bar diagram

Multiple Bar Diagram

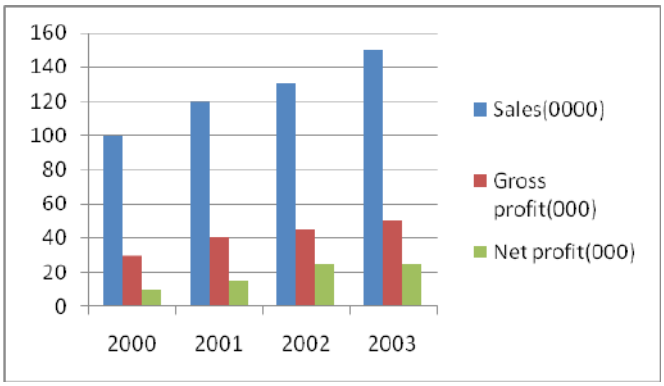
Multiple bar diagrams are those diagrams which show two or more sets of data simultaneously. Each set of variables is represented by a set of bars placed close to each other.

Illustration 4.3

Draw a multiple bar diagram for the following data.

Year	*Sales (000)*	*Gross Profit (000)*	*Net Profit (000)*
2000	*100*	*30*	*10*
2001	*120*	*40*	*15*
2002	*130*	*45*	*25*
2003	*150*	*50*	*25*

Solution:



Know your progress

Draw a suitable bar diagram from the following data

Items	Year	
	2000	2004
Industries	250	350
Agriculture	670	300
Internal trade	500	800

Activity

Collect the data of number of boys and girls in the last five years from your school and represent it in bar diagram

Sub Divided Bar Diagram

The subdivided or component bar diagrams are used to represent data in which the total magnitude (bar) is divided into different parts or components. In these

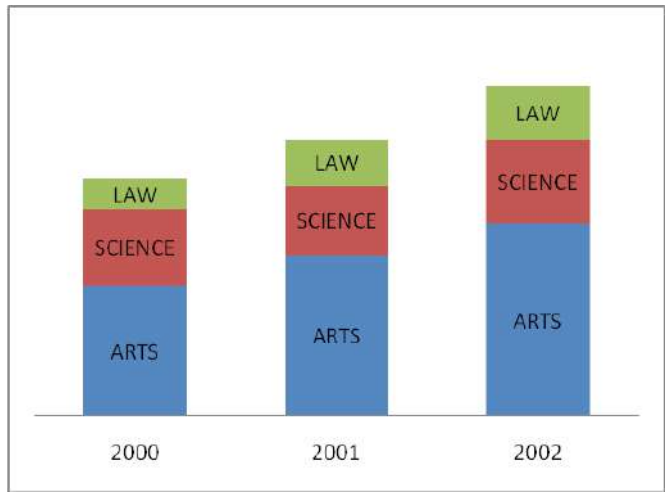
types of diagrams, first make simple bars for each class taking total magnitude in that class and then divided theses simple bars into parts in the ratio of various components.

Illustration 4.4

During 2000-02 the number of students in a university is as follows. Represent the data by using sub divided bar diagram

Year	*Arts*	*Science*	*Law*	*Total*
2000	*17000*	*10000*	*4000*	*31000*
2001	*21000*	*9000*	*6000*	*36000*
2002	*25000*	*11000*	*7000*	*43000*

Solution:



Percentage Bar Diagram

Percentage bar diagram is a modified form of the component bar diagram. It is used when the comparison of components is important. To construct a percentage bar diagram, the component bars correspond to the percentages of the total of each category.

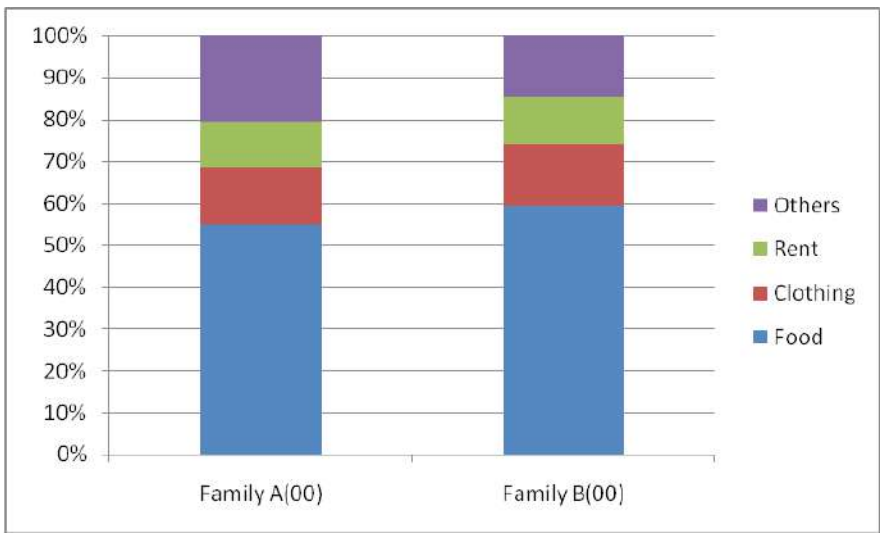
Illustration 4.5

The following are the expenditure of two families for various items. Represent it by using percentage bar diagram.

Types of items	Family A(00)	Family B(00)
Food	40	80
Clothing	10	20
Rent	8	15
Others	15	20
Total	73	135

Solution:

Types of items	Family A(00)	Percentage of family A	Family B(00)	Percentage of family B
Food	40	54.8	80	59.3
Clothing	10	13.7	20	14.8
Rent	8	11	15	11.1
Others	15	20.5	20	14.8
Total	73	100	135	100



Know your progress

During 2005-08 the number of students in a school are as follows

Year	humanities	science	commerce
2005-06	110	150	60
2006-07	118	180	55
2007-08	120	240	120

Represent the above data by percentage bar diagram

Activity

Collect the data of number of students studying in science, humanities and commerce from neighborhood of three schools and represent it in a bar diagram.

Pie diagram

Pie diagram is a circular diagram, with sectors representing the values of the data given. The areas of each sector will be proportional to the values of items and the area of the whole circle will be proportional to the totality.

In constructing a pie chart, the first step is to prepare the data so that the various component values can be transposed into corresponding degrees on the circle by using the formulae

$$Angle = \frac{Item frequency}{Total frequency} \times 360 \tag{4.1}$$

Illustration 4.6

The sales of an appliance in 5 cities in October,2013 is given in lakhs(Rs) is given below. Draw a pie diagram to represent the population in a town

City	*Sales in lakhs*
Kannur	*79*
Kozhikode	*99*
Cochin	*134*
Kollam	*70*
Trivandrum	*83*

Solution:

Cities	*Sales in lakhs*	*Angle*
Kannur	*79*	$\frac{79}{465} \times 360 = 61.2$
Kozhikode	*99*	$\frac{99}{465} \times 360 = 76.6$
Cochin	*134*	$\frac{134}{465} \times 360 = 103.7$
Kollam	*70*	$\frac{70}{465} \times 360 = 54.2$
Trivandrum	*83*	$\frac{83}{465} \times 360 = 64.3$

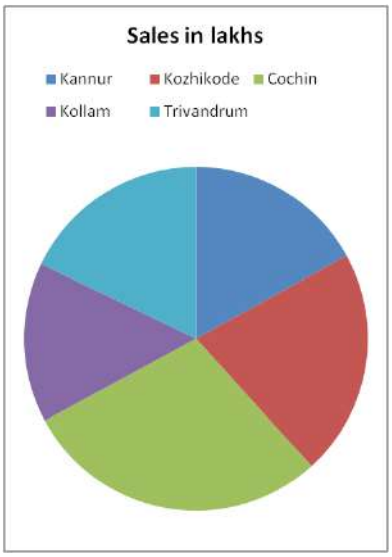


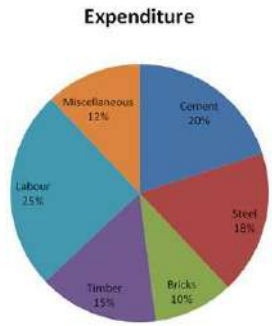
Illustration 4.7

The following figures relate to the cost of constructions of house in Delhi Draw a pie diagram to represent it.

Items	*Expenditure*
Cement	*20%*
Steel	*18%*
Bricks	*10%*
Timber	*15%*
Labour	*25%*
Miscellaneous	*12%*

Solution:

Items	*Corresponding angles*
Cement	$\frac{20}{100} \times 360 = 72$
Steel	$\frac{18}{100} \times 360 = 64.8$
Bricks	$\frac{10}{100} \times 360 = 36$
Timber	$\frac{15}{100} \times 360 = 54$
Labour	$\frac{25}{100} \times 360 = 90$
Miscellaneous	$\frac{12}{100} \times 360 = 43.2$



Know your progress

Draw a pie diagram to represent the following data of investment patterns in the 5 year plan

Agriculture	:	20%
Irrigation	:	18%
Small industries	:	22%
Transport	:	14%
Social service	:	15%
Others	:	11%

Activity

Collect the expenditure of various food items (rice, wheat, oil, et(c) from your home in a month and represent by means of pie diagram.

Pie Chart from William Henry Playfairs Statisical Breviory (1801), showing the proportions of the Turkish Empire located in Asia, Europe and Africa before 1789.

4.3 Graphs

The graphs are designed to make known clearly the characteristic features of a data .In the previous chapter we have seen that the frequency distribution can be represented in a table. But the graphs are more attractive to the eye than the tabulated data. The most commonly used graphs for representing a frequency distribution are

(a) Histogram

(b) Frequency Polygon

(c) Frequency Curve

(d) Ogives(cumulative frequency curve)

(e) Scatter Plot

Histogram

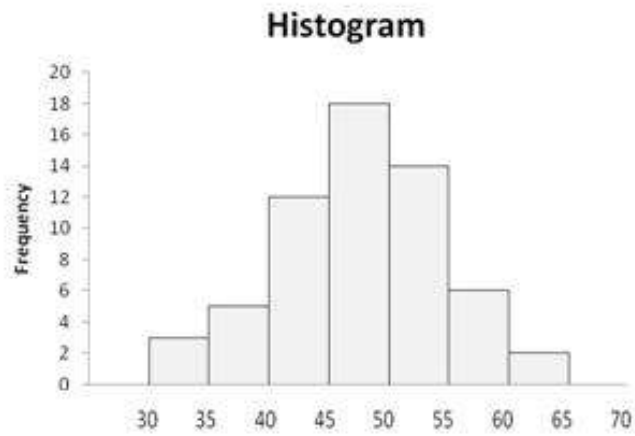
Histogram is an important method for displaying the frequency distribution .It is a set of vertical bars whose heights are proportional to the frequencies represented. In constructing histogram ,the variable is always taken on the X-axis and frequencies on the Y - axis. The width of the bars in the histogram will be proportional to the class interval. A histogram generally represents a continuous curve.

Illustration 4.8

The frequency distribution of weights of 60 students of a class in a school is given below. Draw histogram

Weight in kg	*30-35*	*35-40*	*40-45*	*45-50*	*50-55*	*55-60*	*60-65*
No. students	*3*	*5*	*12*	*18*	*14*	*6*	*2*

Solution:



Frequency Polygon

Mid points of the top of bars in a histogram are joined together by straight lines and then join to the X-axis at both extreme points. It gives a frequency polygon Another method of constructing frequency polygon is to take the midpoints of the various class intervals and then plot the frequency corresponding to each point and to join all these points with straight lines

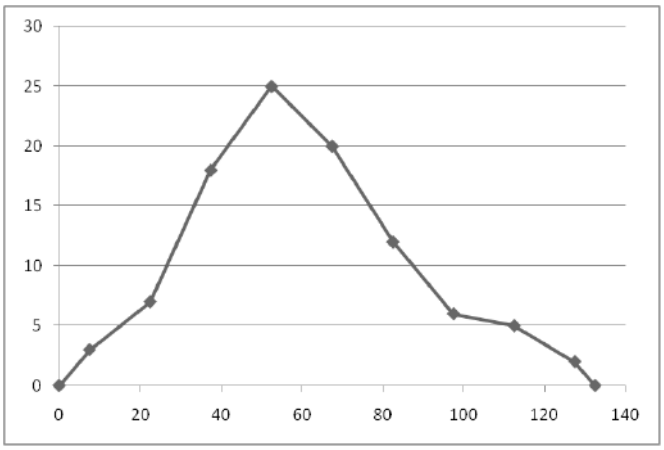
Illustration 4.9

Draw a frequency polygon for the following data

Profit (000)	*0-15*	*15-30*	*30-45*	*45-60*	*60-75*	*75-90*	*90-105*	*105-120*	*120-135*
No. Companies	*3*	*7*	*18*	*25*	*20*	*12*	*6*	*5*	*2*

Solution:

Mid Points	*7.5*	*22.5*	*37.5*	*52.5*	*67.5*	*82.5*	*97.5*	*112.5*	*127.5*
Frequency	*3*	*7*	*18*	*25*	*20*	*12*	*6*	*5*	*2*



Frequency Curve

A frequency curve is obtained by joining the points of a frequency polygon with a freehand smooth curve.

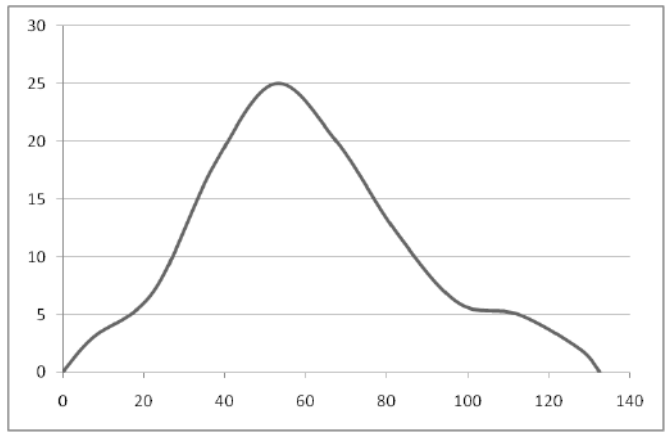
Illustration 4.10

Draw frequency curve for the following data

Profit (000)	*0-15*	*15-30*	*30-45*	*45-60*	*60-75*	*75-90*	*90-105*	*105-120*	*120-135*
No. Companies	*3*	*7*	*18*	*25*	*20*	*12*	*6*	*5*	*2*

Solution:

Mid Points	*7.5*	*22.5*	*37.5*	*52.5*	*67.5*	*82.5*	*97.5*	*112.5*	*127.5*
Frequency	*3*	*7*	*18*	*25*	*20*	*12*	*6*	*5*	*2*



Know your progress

1. Draw a histogram on the basis of following data

Mid values	18	25	32	39	46	53	60
Frequency	10	15	32	42	26	12	9

(hint:actual class=25-18=7, 18 ± 7/2, 14.5-21.5)

2. The following table gives the marks of 60 students in a class. Draw a histogram

marks	20-24	25-29	30-34	35-39	40-44	45-49
No.of students	10	16	7	11	9	7

(hint: change inclusive classes into exclusive classes)

3. Prepare histogram, frequency polygon and frequency curve from the following data

Marks:	0-5	5-10	10-15	15-20	20-25	25-30
No. of students:	8	6	26	40	30	8

Activity

Collect the marks of students from your class in a particular subject and draw histogram, frequency polygon and frequency curve

Ogives (Cumulative Frequency Curve)

Suppose a teacher is interested to knowing how many students have scored less than 30 marks in a class test or how many students have scored more than 50 marks in a test. To answer these questions, it is necessary to add the frequencies. When frequencies are added, they are called cumulative frequencies. These frequencies are then listed in a table called a cumulative frequency table. The curve obtained by plotting cumulative frequencies is called a cumulative frequency curve or ogive. There are two types of ogives

1. Less than Ogive
2. Greater than Ogive (More than Ogive)

Less than Ogive

It is drawn by plotting points with the upper bound of the classes along X-axis and the corresponding less than cumulative frequencies along Y-axis and joining these points with a smooth curve is called a less than ogive.

Illustration 4.11

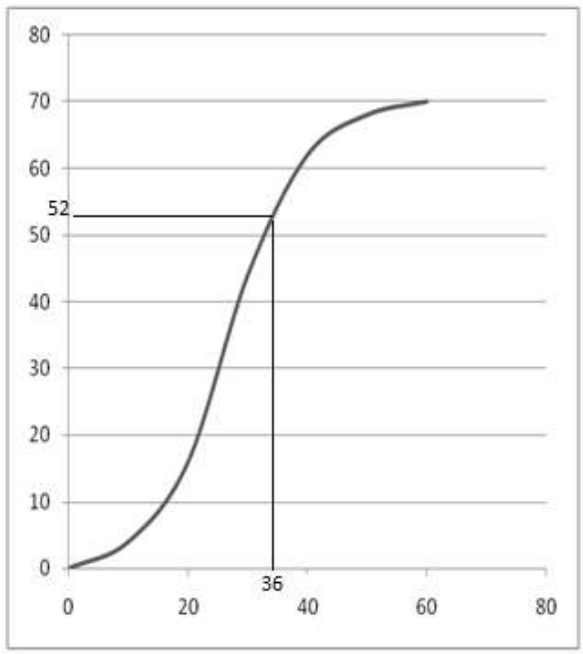
Below is given the frequency distribution of ages of 100 persons in a colony. Draw a less than ogive

Age	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
No. persons	*4*	*12*	*28*	*18*	*6*	*2*

Using this ogive, find how many persons have age less than 36

Solution:

Age	Frequency	Upper bound	Less than cumulative frequency
0-10	4	10	4
10-20	12	20	16
20-30	28	30	44
30-40	18	40	62
40-50	6	50	68
50-60	2	60	70



Number of persons having age less than 36 is equal to 52

Greater than Ogive (more than Ogive)

In order to get greater than ogive, we plot the lower limits along X - axis and corresponding greater than cumulative frequencies along Y - axis and joined these points with a smooth curve.

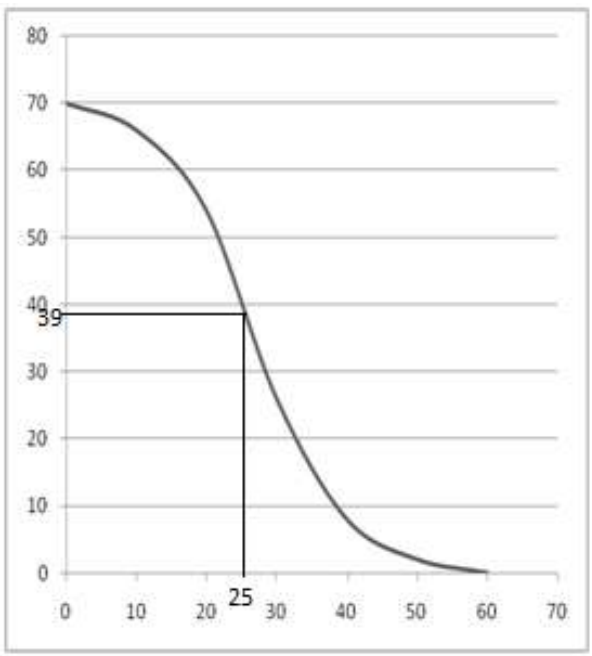
Illustration 4.12

Given below is the frequency distribution of ages of 100 persons in a colony.Draw a greater than ogive and also find how many persons have age greater than 25

Age	0-10	10-20	20-30	30-40	40-50	50-60
No.persons	4	12	28	18	6	2

Solution:

Age	Frequency	Lower bound	Greater than cumulative frequency
0-10	4	0	70
10-20	12	10	66
20-30	28	20	54
30-40	18	30	26
40-50	6	40	8
50-60	2	50	2



The number of persons greater than age 25 is equal to 39

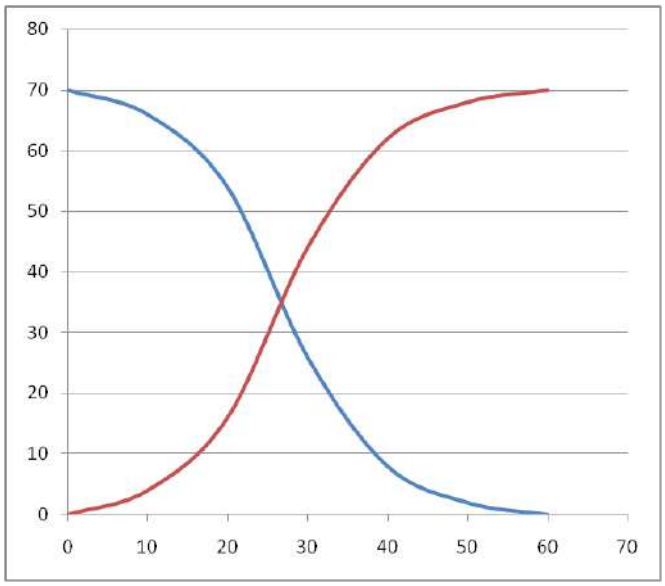
Illustration 4.13

Draw a less than ogive and greater ogive (Ogives) in the same graph.

Age	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
Number of persons	*4*	*12*	*28*	*18*	*6*	*2*

Solution:

Age	*Frequency*	*Upper bound*	*Less than cumulative frequency*	*Lower bound*	*Greater than cumulative frequency*
0-10	*4*	*10*	*4*	*0*	*70*
10-20	*12*	*20*	*16*	*10*	*66*
20-30	*28*	*30*	*44*	*20*	*54*
30-40	*18*	*40*	*62*	*30*	*26*
40-50	*6*	*50*	*68*	*40*	*8*
50 - 60	*2*	*60*	*70*	*50*	*2*



Know your progress

The following are the frequency distribution of daily wages of 60 persons in a company. Draw a greater than ogive and less than ogive (ogives) in a same graph

Wage (in tens)	30-40	40-50	50-60	60-70	70-80	80-90
Number of persons	5	10	20	15	7	3

Activity

Collect the data of height of students from your class and draw ogives

Scatter plot

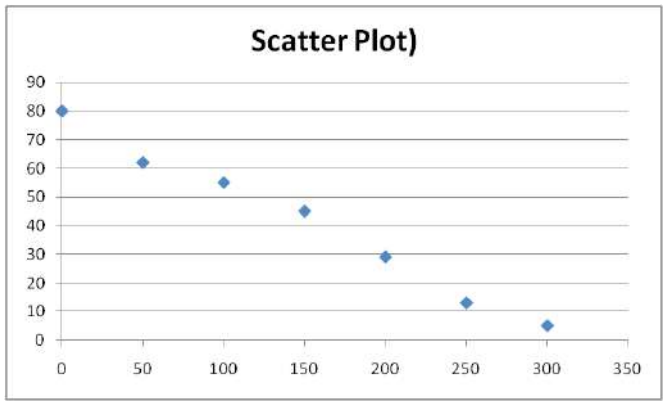
Scatter plot is a diagram to represent a bivariate data. It is used to analyse the relationship between two variables. In a scatter plot, one variable is represented along the X- axis and the other variable along the Y- axis. Each pair is represented by a single point.

Illustration 4.14

A driver keeps a record of the distance travelled and amount of fuel in his tank on a long journey. Draw a scatter plot.

Distance Travelled (in KM)	*0*	*50*	*100*	*150*	*200*	*250*	*300*
Fuel in Tank (in Ltr)	*80*	*73*	*67*	*61*	*52*	*46*	*37*

Solution:



Know your progress

An insurance company interested in studying the basic pay and LIC premium remittance of employees in a locality. They took a sample of 7 employees

Basic pay (00)	:	42	45	50	29	60	65
LIC remittance (in tens)	:	18	30	40	35	70	60

Represent the data on a graph

Activity

Collect the heights and weights of students from your class and represent it on a graph together.

Difference between Diagrams and Graphs

DIAGRAMS	GRAPHS
It is more attractive to eye	Less attractive
It is not based on the co-ordinate system	It is based on the co-ordinate system
Does not represent any mathematical relationship between variables	It represents mathematical relationship between variables
It is not so helpful in statistical analysis	Very much used in statistical analysis
It is used for a specific purpose	It is capable for further mathematical treatments

Let us sum up

In this chapter we have discussed different types of diagrams and graphs. It helps us to make comparison among various type of data very easily.Diagrams and graphs give a birds eye view of the entire data. The types of diagrams and graphs discussed in this chapter are very common and useful in day to day life

Learning outcomes

After transaction of this unit, the learner:-

- identifies importance of diagrammatic representation of data.
- explains different types of diagrams and graphs.
- creates different types of diagrams and graphs.
- interprets data using diagrams and graphs.

Evaluation Items

1. Which of the following is a one dimensional diagram?
   (a) bar diagram (b) frequency polygon
   (c)frequency curve (d) ogives

2. In an ogive ,the points are plotted for
   (a) the values and frequencies (b)the values and cumulative frequencies
   (c)frequencies and cumulative frequencies (d)none of the above

3. Pie chart represents the components of a factor by
   (a) percentage (b) angles
   (c) sectors (d)circles

4. Histogram is suitable for the data presented as
   (a) continous grouped frequency distribution (b) individual series
   (c)discrete grouped frequency distribution (d) all the above

5. With the help of histogram we can prepare
   (a) frequency polygon (b) frequency curve
   (c) both (d) none

6. Discuss the importance of diagrams and graphs

7. Prepare a table consisting of the merits and demerits of diagrams and graphs?

8. Illustrate four different types of bar diagrams

9. Write the importance of ogives?

10. Following table gives the birth rate per thousand of different countries over a period, represent by a suitable bar diagram

Country	Birth rate
India	32
Germany	18
UK	20
China	40

11. The data below show the yearly profits (in thousands of rupees) of the two companies A and B

Year	Company A	Company B
2000	120	90
2001	135	95
2002	140	108
2003	160	120

Represent the data by means of a suitable diagram (hint: multiple bar diagram)

12. The following table gives the height of trees

Height	**No. of trees**
Below 7 feet	26
Below 14 feet	57
Below 21 feet	92
Below 28feet	134
Below 35 feet	216
Below 42 feet	287
Below 49 feet	341
Below 56 feet	360

Represent the data in the form of histogram

13. Height distribution of a group of students are given below

Height (cm)	150-153	153-156	156-159	159-162	162-165	165-168	168-171	171-173
No. students	4	8	15	20	28	16	10	4

(a) Draw ogives

(b) How many students have heights less than 160 cm

(c) How many students have height greater than 166 cm

Answers

1. a , 2. b , 3. c , 4. a , 5. c

5 Central Tendency

Introduction

In the previous units we discussed how data can be collected and organized in a meaningful manner so that we can use it more conveniently for statistical analysis. The frequency table and graphic presentation of data make it more meaningful. This chapter takes you beyond frequency distributions.

Suppose you come to class, the day after a series of examinations. Someone asks, ″how much score you expect in this exams?″ you may answer like, about 70%, nearly 70%, around 70% etc. Have you thought why you use the words about, nearly, around etc before 70%. It is because you select a representative of the different scores you may get. ie sometimes we use one single value to represent a data. In this way it is easier to compare data of the same type. These representative values are usually known as **averages** or **measures of central tendencies.**

Central Tendency

Consider the distribution of the number of family members of 60 students in a class, we discussed in Chapter 3. What we can see that the observations are more concentrated towards the centre of the distribution. See the table below.

Number of family members	Tally Mark	Frequency
2	II	2
3	卌 I	6
4	卌 IIII	9
5	卌 卌 IIII	14
6	卌 卌 III	13
7	卌 II	7
8	IIII	4
9	II	2
10	II	2
11	I	1
Total		60

Similarly the distribution of body weights of 60 students (Chapter 3) also shows the same property. On examining the frequency distributions we can see that the observations in most distributions show a tendency to cluster around a central value. This property of the observations in a data to cluster or concentrate around a value is known as *Central Tendency.*

Measures of Central Tendency (Averages)

The values which give us an idea about the concentration of observations in the central part of the distribution are known as *Measures of Central Tendency* or *Averages*. It is a single value which represents an entire set of data, around which most of the values of the data cluster. According to *Professor Arthur Lyon Bowley,* averages are, "statistical constants which enable us to comprehend in a single effort, the significance of the whole".

Arthur Lyon Bowley

Sir Arthur Lyon Bowley, born on 6 November 1869 was a British statistician pioneered the use of sampling techniques in social surveys. Bowley's "Elements of Statistics" is generally regarded as the first English-language statistics text-book. It described the techniques of descriptive statistics that would be useful for economists and social sciences.

Desirable properties of a good average

An average should posses the following properties.

(i) Simple and rigid definition.

(ii) Simple to understand and easy to calculate.

(iii) Based on all the observations.

(iv) Least affected by extreme values.

(v) Least affected by fluctuations of sampling.

(vi) Capable of further mathematical treatment.

In 1796, Adolphe Quetelet, a Belgian astronomer, Mathematician, Statistician and Sociologist, investigated the characteristics heights, weights etc of French conscripts to determine the average man. Florence Nightingale, a celebrated British social reformer and statistician, and the founder of modern nursing, was so influenced by Quetelets work that she began collecting and analysing medical records in the military hospitals during the Crimean war. Based on her work, hospitals began keeping accurate records on their patients. Florence Nightingale was the first female member of the Royal Statistical Society.

Various measures of Central Tendencies

The measure of central tendency indicates the location or position of a value to describe the entire data. The various measures of central tendencies are

1. Arithmetic Mean (AM)
2. Median
3. Mode
4. Geometric Mean (GM)
5. Harmonic Mean (HM)

5.1 Arithmetic Mean (AM)

The *'Arithmetic Mean'* (referred to as *'mean'*) is a most common measure of central tendency. The mean is a common measure in which all the values play an equal role. Most of the time when we refer to the 'average' of a data, we are talking about its arithmetic mean. For example, to find the average life of a CFL bulb, the average temperature in a city etc, the average we refer to is the arithmetic mean.

The table below shows the number sales of five retail outlets in a day.

Retailer	:	1	2	3	4	5
Sales (in 1000 Rs)	:	8	23	4	8	2

To find the average sales per a retailer in the day, we sum the values and divide by the number of observations. This average, called the arithmetic mean is computed as

$$AM = \frac{8+23+4+8+2}{5} = \frac{45}{5} = 9$$

ie, the average sales per retailer is Rs. 9000/-.

The arithmetic mean of a set of data is defined as

$$\text{Mean} = \frac{\text{sum of the observations}}{\text{number of observations}}$$

We usually denote mean by the symbol '$\bar{x}$' (read as 'x bar').

$$\bar{x} = \frac{\text{sum of the observations}}{\text{number of observations}}$$

Computation of Arithmetic Mean

(i) Arithmetic Mean from a raw data

Consider a raw data containing n observations $x_1, x_2, x_3, \ldots, x_n$. The mean can be calculated by

$$\bar{x} = \frac{\text{sum of the observations}}{\text{number of observations}}$$

$$= \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$= \frac{\sum x}{n}$$

$$\text{Mean} = \frac{\sum x}{n}$$

The Greek letter $\sum$ represents summation

Remark: The sum of the observations in a series, $\sum x = n\bar{x}$

Illustration 5.1

An umbrella manufacturing company wants to launch a new product in a state. The rainfall (in cms) in the state for the last five years is, 120, 135, 110, 142 and 150. Find the average rainfall of the state for the last five years.

Solution. Average rainfall,

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{120 + 135 + 110 + 142 + 150}{5}$$

$$= 131.4\,cm$$

Illustration 5.2

In the first four class tests, a student got the scores 52, 48, 33 and 27 respectively. (a) Find the mean of the scores.
(b) If in the fifth test, he got a score of 45, find his new mean score.

Solution. (a) The mean score of the first four class tests is

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{52 + 48 + 33 + 27}{4}$$

$$= 40$$

(b) The new mean after 5 class tests is

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{52 + 48 + 33 + 27 + 45}{5}$$

$$= 41$$

Illustration 5.3

The mean of a group of 100 observations is known to be 50. Later it was discovered that two observations were misread as 92 and 8 instead of 192 and 88. Find the correct mean.

Solution. Given that $\bar{x} = 50$ and $n = 100$.

We have the sum of the observations,

$$\sum x = n\bar{x}$$

$$= 100 \times 50$$

$$= 5000$$

But it was wrong, because two observations 192 and 88 were misread as 92 and 8. So the corrected sum of observations is

$$\begin{aligned}\text{corrected sum} &= 5000 - 92 - 8 + 192 + 88\\ &= 5180\end{aligned}$$

$$\text{So corrected mean, } \bar{x} = \frac{\text{corrected sum}}{n}$$

$$= \frac{5180}{100}$$

$$= 51.80$$

(ii) Arithmetic Mean from a Discrete Frequency Distribution

A discrete frequency distribution consists of data in which the observations are expressed with their frequencies.

In discrete type distribution, for calculating the mean every item is multiplied by its corresponding frequencies and the total sum of this product is divided by the sum of the frequencies.

The Arithmetic Mean of a Discrete frequency distribution is

$$\bar{x} = \frac{\sum fx}{\sum f}$$

$$\text{ie. } \bar{x} = \frac{\sum fx}{N} \text{ where } N = \sum f \text{ is the total frequency}$$

The following steps can be used to find the mean of a given series $x_1, x_2, x_3, \ldots, x_n$ with corresponding frequencies $f_1, f_2, f_3, \ldots, f_n$.

Step 1: Find $f_1x_1, f_2x_2, f_3x_3, \ldots, f_nx_n$

Step 2: Find $\sum fx = f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n$

Step 3: Find $\sum f = f_1 + f_2 + f_3 + \cdots + f_n$

Step 4: The mean is obtained by the formula:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Illustration 5.4

The students in a Statistics class were trying to study the heights of participants in a sports meet. They collected the height of 20 participants, as displayed in the table.

Height (in inches)	:	*49*	*53*	*54*	*55*	*66*	*68*	*70*	*80*
No of participants	:	*1*	*2*	*4*	*5*	*3*	*2*	*2*	*1*

Calculate the mean height of the participants.

Solution.

Height (x)	No of participants (f)	fx
49	1	49
53	2	106
54	4	216
55	5	275
66	3	198
70	2	140
80	1	80
Total	$N = 20$	1200

$$\bar{x} = \frac{\sum fx}{N}$$
$$= \frac{1200}{20} \qquad = 60$$

Mean height = 60 inches

Illustration 5.5

A survey is taken by an insurance company to determine how many car accidents the average New Delhi City resident has gotten into in the past 10 years. The company surveyed 200 people who are getting off a train at a subway station. The following table gives the results of the survey.

Number of Accidents	*Number of People*
0	*60*
1	*10*
2	*40*
3	*10*
4	*80*

Calculate the mean number of accidents of this data set.

Solution.

No of accidents (x)	No of people (f)	fx
0	70	0
1	10	10
2	40	80
3	10	30
4	70	280
Total	$N = 200$	400

$$\bar{x} = \frac{\sum fx}{\sum f}$$
$$= \frac{400}{200} \qquad = 2$$

Mean of accidents = 2

(iii) Arithmetic Mean from a Continuous Frequency Distribution

A continuous frequency distribution consists of data that are grouped by classes.

The computation of AM is similar to the computation procedure for the discrete frequency distribution. But, since the data is grouped by classes, we do not know the individual values of every observation. So it is necessary to make an assumption about these values. The assumption is that every observation in a

class has a value equal to the midpoint of the class. The formula for computing AM is,

$$\bar{x} = \frac{\sum fx}{N} \text{ where } N = \sum f$$

This arithmetic mean is only an approximation. (Why?)

Illustration 5.6

The table below show the age of 55 patients selected to study the effectiveness of a particular medicine.

Age	*No of Patients*
0-10	*5*
10-20	*7*
20-30	*17*
30-40	*12*
40-50	*5*
50-60	*2*
60-70	*7*

Calculate the mean age of the patients.

Solution. To find mean, prepare the following table

Age	Mid point (x)	No of patients (f)	fx
0-10	5	5	25
10-20	15	7	105
20-30	25	17	425
30-40	35	12	420
40-50	45	5	225
50-60	55	2	110
60-70	65	7	455
Total		$N = 55$	1765

$$\bar{x} = \frac{\sum fx}{\sum f}$$
$$= \frac{1765}{55}$$
$$= 32.09$$

Illustration 5.7

The frequency distribution below represents the weights in kg of parcels carried by a small logistic company. Find the mean weight of parcels.

Weight	*No. of Parcels*
10.0-10.9	*2*
11.0-11.9	*3*
12.0-12.9	*5*
13.0-13.9	*8*
14.0-14.9	*12*
15.0-15.9	*15*
16.0-16.9	*13*
17.0-17.9	*11*
18.0-18.9	*6*
19.0-19.9	*2*

Solution.

Weight	Mid point (x)	No. of Parcels (f)	fx
10.0-10.9	10.45	2	20.90
11.0-11.9	11.45	3	34.35
12.0-12.9	12.45	5	62.25
13.0-13.9	13.45	8	107.60
14.0-14.9	14.45	12	173.40
15.0-15.9	15.45	15	231.75
16.0-16.9	16.45	13	213.85
17.0-17.9	17.45	11	191.95
18.0-18.9	18.45	6	110.70
19.0-19.9	19.45	2	38.90
Total		$N = 67$	1185.65

$$\bar{x} = \frac{\sum fx}{\sum f}$$
$$= \frac{1185.65}{67} \qquad = 17.70$$

Mean weight = 17.70 kg

Mathematical Properties of Arithmetic Mean

The AM of a distribution has the following mathematical properties.

1. The sum of deviations of items in a data from the AM is always Zero.

$$ie \sum (x - \bar{x}) = 0$$

2. The sum of squares of the deviations of the items in a data is the least when the deviation is taken about the Mean.

$$ie \sum (x - a)^2 \text{ is least when } a = \bar{x}$$

3. If the mean of n observations, $x_1, x_2, \ldots, x_n$ is $\bar{x}$ then the mean of the observations, $(x_1 \pm a), (x_2 \pm a), \ldots, (x_n \pm a)$ is $(\bar{x} \pm a)$.

   ie, If each observation is increased by 'a', then the mean also increased by a and if each observation is decreased by 'a', then the mean is also decreased by a.

4. The mean of n observations, $x_1, x_2, \ldots, x_n$ is $\bar{x}$. If each observation is multiplied by $p, p \neq 0$, then the mean of the new observations is $p\bar{x}$.

Activity

Observe the data 10, 25, 17, 22, 20, 35, 28, 42, 68 and 53

(a) Examine whether $\sum (x - \bar{x}) = 0$ for this data.

(b) Examine the admissibility of property 2 by giving some values to a

(c) What happens to the mean by

(i) adding 3 to all the observations in the data.

(ii) subtracting 3 from all the observations.

(d) What happens to the mean

(i) if all the observations are multiplied by 2.

(ii) if all the observations are divided by 2.

Merits and Demerits of AM

Arithmetic Mean is the measure which has most of the desirable properties of a good measure of central tendency. The following are some of the merits and demerits of it

Merits

1. It has a rigid definition.
2. It is easy to calculate and understand.
3. AM is based upon all the observations.
4. It is least affected by fluctuations of sampling.
5. It is capable of further mathematical treatment.

Demerits

1. AM is highly affected by extreme values.
2. It can't be determined by inspection.
3. It can't be used for qualitative characteristics like, intelligence, honesty, beauty etc.
4. It can't be calculated for open end classes.

Weighted Arithmetic Mean

Usually in computing Arithmetic Mean, equal importance is given to all the observations of the data. However there are cases where all the items are not of equal importance. In other words some items of a series are more important as compared to the other items in the same series. In such cases it becomes important to assign different weights to different items. For example if want to have an idea of the change in the living standards of a certain group of people, the AM we discussed so far can't be used. Because, all the commodities the people used may not be of equal importance. Rice, wheat etc may be more important when compared with sugar, tea, salt etc.

The AM that assign a weight to each observation on its importance related to other is called the weighted arithmetic mean. Weighted AM are widely used in the preparation of consumer and producer price index numbers.

Computation of Weighted Arithmetic Mean

Let $x_1, x_2, \ldots, x_n$ be the observations of a data and $w_1, w_2, \ldots, w_n$ be their corresponding weights. Then the weighted arithmetic mean is given by,

$$\begin{aligned}\bar{x} &= \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} \\ &= \frac{\sum wx}{\sum w}\end{aligned}$$

Illustration 5.8

A student's final scores in Mathematics, Physics, Chemistry and English are respectively 82, 86, 90 and 70. If the respective credits received for these courses are 3, 5, 3, and 1, determine the average score.

Solution. Here the weights associated to the observations 82. 86. 90 and 70 are 3, 5, 3 and 1.

x	:	82	86	90	70
w	:	3	5	3	1

Average,

$$\bar{x} = \frac{\sum wx}{\sum w}$$

$$= \frac{82 \times 3 + 86 \times 5 + 90 \times 3 + 70 \times 1}{3 + 5 + 3 + 1}$$
$$= \frac{1016}{12}$$
$$= 84.67$$

Combined Arithmetic Mean

In a class test taken by 4 boys and 6 girls, the boys obtained a mean score of 35 and the girls obtained a mean score of 75. What would be the mean score of all the students taken together ?

Is it $\frac{35+75}{2} = 55$?

Obviously not, because the number of boys and number of girls are not equal. The mean score of girls has a greater importance than boys (why?).

For all the students,

$$\text{the mean score} = \frac{\text{Total score of all the students}}{\text{Total number of students}}$$
$$= \frac{\text{Total score of boys} + \text{Total score of girls}}{\text{No. of boys} + \text{No. of girls}}$$
$$= \frac{4 \times 35 + 6 \times 75}{4 + 6}$$
$$= \frac{140 + 450}{10}$$
$$= 59$$

Therefore, the mean score of the combined group of boys and girls, called the *combined mean is 59.*

If $\bar{x_1}$ and $\bar{x_2}$ are the means of two groups of n_1 and n_2 observations respectively, the mean of the combined group of $n_1 + n_2$ observations is given by

$$\bar{x} = \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2}$$

Illustration 5.9

The mean score obtained in an examination by a group of 100 students was found to be 50. The mean of the scores obtained in the same examination by another group of 200 students was 57. Find the mean of scores obtained by both the groups taken together.

Solution. We are given that

$$\bar{x_1} = 50 \text{ and } \bar{x_2} = 57$$
$$n_1 = 100 \text{ and } n_2 = 200$$

We know that the combined mean is given by

$$\begin{aligned}\bar{x} &= \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2} \\ &= \frac{100 \times 50 + 200 \times 57}{100 + 200} \\ &= 54.67\end{aligned}$$

Illustration 5.10

The mean weight of 150 students in a certain class is 60 kgs. The mean weight of boys in the class is 70 kgs and that of girls is 55 kgs. Find the number of boys and number of girls in the class.

Solution. We are given that

The combined mean, $\bar{x} = 60$ kgs
Mean weight of boys, $\bar{x_1} = 70$ kgs
Mean weight of girls, $\bar{x_2} = 55$ kgs
The total no of students = 150

Let there be 'x' boys in the class. Therefore the number of girls in the class is

$150 - x$. We know that,

$$\bar{x} = \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2}$$

$$\text{ie,} 60 = \frac{70x + 55(150 - x)}{150}$$

$$\Rightarrow 9000 = 70x + 8250x - 55x$$

$$\Rightarrow 15x = 750 \Rightarrow x = 50$$

So number of boys in the class is 50 and number of girls in the class is $150 - 50 = 100$.

If there are k groups of sizes $n_1, n_2, \ldots\ldots, n_k$ respectively and $\bar{x_1}, \bar{x_2}, \ldots\ldots, \bar{x_k}$ are their respective means, then the combined mean of $n_1 + n_2 + \ldots + n_k$ observations is given by

$$\bar{x} = \frac{n_1\bar{x_1} + n_2\bar{x_2} + \ldots + n_k\bar{x_k}}{n_1 + n_2 + \ldots + n_k}$$

Know your progress

1. 10 novels were randomly selected and the number of pages were recorded as follows, 415, 398, 497, 399, 402, 405, 395, 412, 407 and 400. Find the mean of the number of pages.
2. The average of 20 values is calculated to be 25. Later it was discovered that two values 52 and 15 were misread as 42 and 51 respectively. Calculate the corrected value of the mean.
3. The weights of 70 workers in a factory are given below. Find the mean weight of a worker

Weight (in Kgs)	No of workers
60	5
62	10
63	12
65	18
67	15
68	10

4. 30 automobiles were tested for fuel efficiency (in kms/litter). The following frequency distribution was obtained. Calculate the average mileage of the automobiles.

Mileage	No of vehicles
7.5-12.5	3
12.5-17.5	5
17.5-22.5	15
22.5-27.5	5
27.5-32.5	2

5. For 50 antique car owners, the following distribution of the car's ages was obtained. Determine the mean age.

Car age	No of cars
16-18	20
19-21	18
22-24	8
25-27	4

6. The mean mark of 100 students in a class is 39. The mean mark of the boys is 35 while that of girls is 45. Find the number of boys and that of the girls in the class.

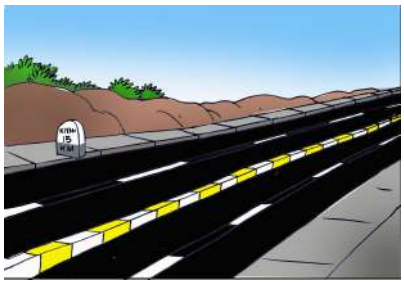
7. A candidate obtains the following marks in an examination. English- 46, Economics-58, Accountancy-72 and Statistics-67. It is agreed to give double weights to marks in English and Statistics. What is the mean mark.

5.2 Median

A family has 5 children aged 10, 8, 5, 4 and 12 years. Can you find the age of the middle child? To find the age of the middle child, we arrange the children's ages in ascending order.

$$4, 5, 8, 10, 12$$

The age of the middle child is the middlemost number in the data, which is 8. Here '8' is called the *median* of the five numbers.



Median devides a road into two equal parts

> *Median* is the value of the middlemost observations in the data when the data is arranged in ascending or descending order.

ie, *Median* of a distribution is the value of the variable which divide the distribution into two equal parts. The half of the observations are smaller than or equal to median and half are larger than or equal to median. So median is known as a *positional average.* Median of a data is the middle most observation in the data when the observations are arranged in ascending or descending order of their values.

Computing the Median

(i) Calculation of Median from a raw data

Consider a raw data having 'n' observations. To find the median, first arrange the data in ascending or descending order. Then median is the $\left(\frac{n+1}{2}\right)^{\text{th}}$ item in the data.

Median of a raw data is the $\left(\frac{n+1}{2}\right)^{\text{th}}$ item, when the data is arranged in ascending or descending order of magnitude.

Illustration 5.11

Rahna's maths quiz scores in 9 competitions were 88, 97, 87, 92, 90, 88, 93, 98 and 95. What was her median quiz score ?

Solution. Arranging the data in ascending order,

$$87, 88, 88, 90, 92, 93, 95, 96, 98$$

Here the number of observations, $n = 9$.

$$\begin{aligned}\text{Median} &= \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item} \\ &= \left(\frac{9+1}{2}\right)^{\text{th}} \text{ item} \\ &= 5^{\text{th}} \text{ item}\end{aligned}$$

The 5^{th} item in the series is 92. Median quiz score is 92

Illustration 5.12

Anand's family plans a trip from Thiruvananthapuram to Wayanad on their summer vacation. They drove through 8 districts. The following are the petrol prices in the 8 districts on those days. Rs.71.9, Rs.72.3, Rs.72.4, Rs. 72.32, Rs. 73, Rs.73.1, Rs.72.2 and Rs.72.48 What is the median petrol price ?

Solution. Arranging the observations in ascending order,

$$71.9,\ 72.2,\ 72.3,\ 72.32,\ 72.4,\ 72.48,\ 73,\ 73.1$$

Number of observations is 8.

$$\begin{aligned}\text{Median} &= \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item} \\ &= \left(\frac{8+1}{2}\right)^{\text{th}} \text{ item} \\ &= 4.5^{\text{th}} \text{ item}\end{aligned}$$

But there is no item in the series having a position 4.5. So we take the mean of the 4^{th} and 5^{th} items in the series as the median.

$$\begin{aligned}\text{Median} &= \text{Mean of } 4^{th} \text{ and } 5^{th} \text{ item} \\ &= \frac{72.32 + 72.4}{2} \\ &= 72.36\end{aligned}$$

Median petrol price = Rs. 72.36.

(ii) Median from a Discrete Frequency Distribution

For a discrete frequency distribution median is the observation having cumulative frequency $\frac{N+1}{2}$, when the observations are arranged in ascending order

The following steps can be used to find the median in a discrete distribution.

Step 1: Arrange the data in ascending or descending order of magnitude.

Step 2: Obtain the cumulative frequencies.

Step 3: Determine $\frac{N+1}{2}$, where N is the total frequency.

Step 4: Median is the value for the $\left(\frac{N+1}{2}\right)^{th}$ item of the data.

Illustration 5.13

The following data gives the daily wages of workers in a manufacturing company. Find the median wage.

Daily wage (in 100 Rupees)	:	6	8	10	12	15	18
Number of workers	:	20	14	7	16	12	2

Solution. The given data is in ascending order.

The cumulative frequencies are given by,

Wages (in 100 rupees)	Number of workers (frequency)	Cumulative frequency
6	20	20
8	14	34
10	7	41
12	16	57
15	12	69
18	2	71
	$N = 71$	

Total frequency $N = 71$. So $\frac{N+1}{2} = \frac{72}{2} = 36$

$\therefore$ Median is the value in the data which comes in the 36^{th} position. Which is the value of the item having cumulative frequency 36, which is 10.

$\therefore$ Median = Rs. 1000/-

Illustration 5.14

The table below shows the marks obtained by 42 students in an examination.

Marks	*9*	*20*	*25*	*40*	*50*	*80*
Number of students	*4*	*6*	*11*	*13*	*7*	*2*

Calculate the median mark.

Solution. The data given is in ascending order. The cumulative frequencies are

Marks	Frequency	Cumulative frequency
9	4	4
20	6	10
25	11	21
40	12	33
50	7	40
80	2	42
	$N = 42$	

Total frequency, $N = 42$.

$$\frac{N+1}{2} = \frac{43}{2} = 21.5$$

$\therefore$ Median is the value in the data which comes in the 21.5^{th} position. But there is no observation in the data which has the cumulative frequency, 21.5. Hence we consider the median as the mean of the 21^{th} and 22^{nd} observations. The 21^{th} observation is 25 and 22^{nd} observation is 40

$$\text{ie, Median} = \frac{25+40}{2} = \frac{65}{2} = 32.5$$

Hence median mark is 32.5.

(iii) Median from a Continuous Frequency Distribution

To find median in continuous frequency series, first we have to locate median class. Median class is the class where $(\frac{N}{2})^{th}$ observation lies. Median of a continuous frequency series is given by

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right)c}{f}$$

Where l - lower bound (actual class limit) of the median class.
c - class interval of the median class.
f - frequency of the median class and
m - cumulative frequency of the class preceding the median class.

The following steps can be used to determine the median for a continuous frequency series.

Step 1: Convert the inclusive type classes to the exclusive type classes (if any).

Step 2: Obtain the cumulative frequencies.

Step 3: Determine $\frac{N}{2}$, where N is the total frequency.

Step 4: Locate the class having cumulative frequency $\frac{N}{2}$.

Step 5: Find median using the above formula.

Illustration 5.15

The distribution of income of 63 families is,

Income (100 Rupees)	:	*30-40*	*40-50*	*50-60*	*60-70*	*70 -80*	*80-90*	*90 -100*
No of workers	:	*6*	*12*	*18*	*13*	*9*	*4*	*1*

Compute the median income.

Solution. The cumulative frequency table is

Class	Frequency	Cumulative frequency
30-40	6	6
40-50	12	18
50-60	18	36
60-70	13	49
70-80	9	58
80-90	4	62
90-100	1	63
	$N = 63$	

$$\frac{N}{2} = 31.5$$

.

The class having cumulative frequency 31.5 is 50-60.

∴ Median class is 50-60.

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right)c}{f}$$

Where, $l = 50, c = 10, f = 18$ and $m = 18$.

$$\begin{aligned} Median &= 50 + \frac{(31.5 - 18)10}{18} \\ &= 50 + \frac{135}{18} \\ &= 57.5 \end{aligned}$$

Illustration 5.16

The table below shows the distribution of marks obtained by 50 students in Economics. Find the median mark.

Marks	:	*10-14*	*15-19*	*20-24*	*25-29*	*30-34*	*35-39*
Number of students	:	*4*	*6*	*10*	*5*	*7*	*3*
	:	*40-44*	*45-49*				
	:	*9*	*6*				

Solution. Here the classes are of inclusive type. Before computing median, we have to convert it into exclusive form to get the actual class limits (class bounds).

Let us prepare the cumulative frequency table with the actual class limits.

Marks	Actual class	Frequency	Cumulative frequency
10-14	9.5-14.5	4	4
15-19	14.5-19.5	6	10
20-24	19.5-24.5	10	20
25-29	24.5-29.5	5	25
30-34	29.5-34.5	7	32
35-39	34.5-39.5	3	35
40-44	39.5-44.5	9	44
45-49	44.5-49.5	6	50
		$N = 50$	

Here $N = 50$. $\therefore \frac{N}{2} = 25$

Hence the median class is $24.5 - 29.5$.

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right)c}{f}$$

Where $l = 24.5, c = 5, f = 5$ and $m = 20$

$$Median = 24.5 + \frac{(25 - 20)5}{5}$$

$$= 29.5$$

5.2.1 Graphical location of Median

Median from Ogives

One of the advantages of median than mean is that it can be located graphically. Median is a positional average. There are two methods of locating median graphically.

(i) Presenting the data graphically by one ogive ('less than' or 'more than' ogive).

(ii) Presenting the data graphically by two ogives ('less than' and 'more than' ogives).

(i) Median by one ogive ('less than' or 'more than' ogive).

The steps involved in determining median using 'less than' (or 'more than') ogive are,

Step 1: Draw 'less than' (or 'more than') ogive.

Step 2: Find $\frac{N}{2}$ and mark it on the y-axis. Where N is the total frequency.

Step 3: Draw a perpendicular from $\frac{N}{2}$ to the right to cut the ogive at a point A (say).

Step 4: From the point A, draw a perpendicular on the x-axis. The point at which it touches the x-axis will be the median value of the series.

Illustration 5.17

The table below shows the marks obtained by 100 students in an examination. Locate median graphically.

Marks	:	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*	*60-70*
No of students	:	*7*	*10*	*21*	*27*	*22*	*9*	*4*

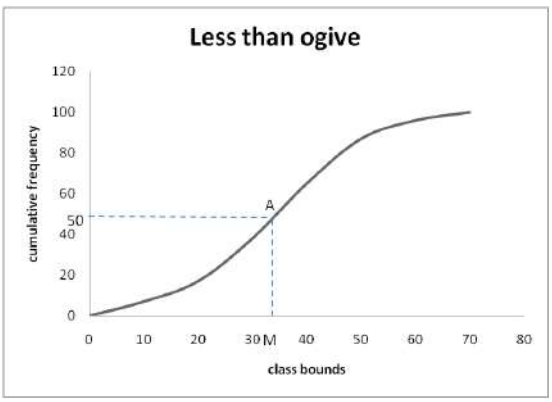
Solution.

(i) Median using 'less than' ogive.

In order to find median using 'less than' ogive, we have to construct the less than frequency table.

Upper bound	Less than cumulative frequency
10	7
20	17
30	38
40	65
50	87
60	96
70	100

$N = 100$ so that $\frac{N}{2} = 50$. Draw a less than ogive.



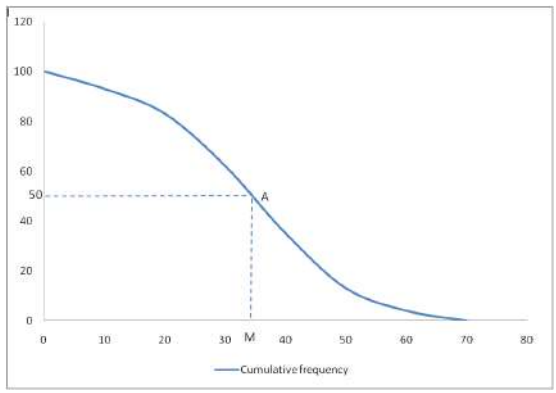
∴ Median is 34.

(ii) Median using more than ogive.

We have to first prepare a greater than cumulative frequency table.

Lower bound	Greater than cumulative frequency
0	100
10	93
20	83
30	62
40	35
50	13
60	4

Draw a more than ogive.



$\therefore$ Median is 34.

(ii) Median from two ogives ('less than' and 'more than' ogives).

The following are the steps involved in the determination of median from 'less than' and 'more than' ogives by simultaneously drawing them.

Step 1: Draw the two ogives on a paper with the same axis.

Step 2: Mark the point A, where the two ogives intersect.

Step 3: Draw perpendicular from A to the x-axis. The corresponding value on the x-axis would be the median of the data.

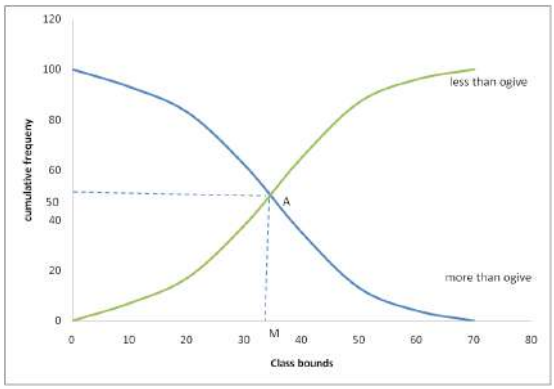
Illustration 5.18

Determine the Median using the data given in ***Illustration 5.17***

Solution. Prepare the two cumulative frequency tables.

Upper bound	Less than cumulative frequency	Lower bound	Greater than cumulative frequency
10	7	0	100
20	17	10	93
30	38	20	83
40	65	30	62
50	87	40	35
60	96	50	13
70	100	60	4

Draw the two ogives simultaneously on the same paper as shown below.



Median = 34.

5.2.2 Merits and demerits of Median

Merits

1. It has a rigid definition.
2. Median is easy to compute. In some cases it be located merely by inspection.
3. It is not affected by the extreme values.

4. It can be calculated for distribution with open end classes.

5. It is the only measure to be used while dealing with qualitative data which can measure quantitatively but can be arranged ascending or descending order of magnitudes.

6. The median may be a better indicator if a set of numbers has an *outlier*. An *outlier* is an extreme value that differs greatly from the other values.

Demerits

1. In some cases median can't be calculated exactly. For example, in the case of even number of observations, we take median as the mean of the two middle terms, as an approximation.

2. It is not based on all the observations.

3. It is affected by the fluctuations of sampling.

4. It is not capable of further mathematical treatments.

Know your progress

1. The manager of a sports shop recorded the number of cricket balls that he sold during seven months. The details are shown below. 132, 121,119, 116, 130, 121, and 131. Calculate median.
2. The weights (in kg) of 10 students are as follows. 31, 35, 27, 29, 32, 43, 37, 41, 34 and 28. Find the median. If the weight of 43 kg is replaced by 48 kg, find the new median.
3. The marks of 43 students are given in the following table. Find the median mark.

Marks	:	20	9	25	50	40	80
No of students	:	6	4	16	7	8	2

4. 80 randomly selected light bulbs were tested to determine their life time (in hours). The following frequency distribution was obtained. Calculate the median life of the bulb.

Life time (in hours)	:	52.5-63.5	63.5-74.5	74.5-85.5
Number of bulbs	:	12	23	18
	:	85.5-96.5	96.5-107.5	107.5-118.5
	:	6	14	5

5. The following is the distribution of marks of 100 students. Calculate the median mark.

Marks	:	30-39	40-49	50-59	60-69
Number of students	:	10	14	26	20
	:	70-79	80-89		
	:	18	12		

5.3 Mode

In everyday life we often apply the concept of majority. For example, in the election to the school parliament, the one who got majority of votes in your class become the class leader. The leader of your class represents your class in the school parliament. ie, Sometimes majority represents the data. *Mode* is the measure which expressing the concept of majority.

> Mode of a data is defined as the value that is repeated most often in the data. It is the observation having the maximum frequency in a data.

Mode is sometimes called the *Fashionable Average* or *Business Average*

Computing the Mode.

(i) Calculating the mode from a raw data.

For a raw data mode is the value which appears most often in the data. ie Mode of a raw data is the observation which appears a maximum number of times in the data.

Illustration 5.19

The ages of six persons who participated in an interview were 20, 21, 21, 24, 25, 24, 21 *and 27 years. Find the mode of the data.*

Solution. Here the observation 21 appears three times, 24 appears two and all others are appear in a single time. So the value which appears a maximum number of times is 21.

$\therefore$ Mode = 21 years.

In a distribution, there may be one, two or more than two modes. If in a distribution there are two observations which are appearing a maximum number of times, then both the observations are taken as the modes of the distribution. A distribution which has a single mode is called a unimodal distribution, which has two modes is called a bimodal and has more than two modes is called a multimodal distribution.

Illustration 5.20

Mr. Vijayakumar, the physical education teacher of a school is trying to determine the average height of students in the cricket team of the school. The height of the players in inches are 70, 72, 72, 74, 74, 74, 75, 76, 76, 76 *and 77. Calculate the mode of the heights.*

Solution. Here the data has two values, 74 and 76, which appears 3 times. All the other values appear less than 3 times.So the data set has two distinct modes 74 and 76.

Illustration 5.21

The generous CEO of a company wants to give all his employees a pay rise.

He is not sure whether to give everyone a straight Rs. 2000 rise or whether to increase the salary by 10%. The mean salary is Rs. 50000, the median is Rs. 20000 and the mode is Rs. 10000. What happens to mean, median and mode if

(a) everyone at the company is given Rs. 2000 pay rise.

(b) the salary is increased by 10%.

Solution. *Given that Mean, Median and Mode are respectively, Rs. 50000, Rs. 20000 and Rs. 10000.*

(a) If every employee are given a straight increase of Rs. 2000,

Let 'x' represents the original value, then the new salary becomes $x + 2000$

$$\begin{aligned} \text{Mean} &= \frac{\sum(x+2000)}{n} \\ &= \frac{\sum x + \sum 2000}{n} \\ &= \frac{\sum x}{n} + \frac{\sum 2000}{n} \\ &= 50000 + 2000 \\ &= 52000. \end{aligned}$$

(b) ie, by adding 2000 to every observation results an increase of 2000 in mean. Similarly the median and mode are also increased by 2000. If 10% increase in salary is given to all the employees,

Let 'x' represents the original values, then the new values become 110% of the original values. ie, This time the new observation is obtained by multiplying 1.1 to the actual values.

$$\begin{aligned} \text{Mean} &= \frac{\sum 1.1x}{n} \\ &= 1.1\frac{\sum x}{n} \\ &= 1.1 \times 50000 \\ &= 55000 \end{aligned}$$

Similarly,

$$\text{Median} = 1.1 \times 20000 = 22000$$
$$\text{Mode} = 1.1 \times 10000 = 11000$$

ie, the mean, median and mode are increased by 10%.

(ii) Mode from a discrete frequency distribution

Mode of a discrete frequency distribution is the observation having the maximum frequency.

The mode of a discrete frequency is the observation which appears a maximum number of times. ie, the observation having the highest frequency.

Illustration 5.22

The following distribution shows the sizes of shirts sold on a textile shop in Thiruvananthapuram on a month. Calculate the mode.

Size (in inches)	:	*36*	*38*	*40*	*42*	*44*
No of shirts sold	:	*15*	*22*	*31*	*30*	*20*

Solution. In the given frequency distribution, the observation having the maximum frequency is 40. Mode is 40.

(iii) Mode from a continuous frequency distribution.

As in the case of median, here also we have to locate a class called modal class to find mode in continous frequency series. Modal class is the class having highest frequency. Mode can be determined by the formula.

$$\text{Mode} = l + \frac{(f_1 - f_0)c}{(f_1 - f_0) + (f_1 - f_2)}$$

$$\text{ie. Mode} = l + \frac{(f_1 - f_0)c}{2f_1 - f_0 - f_2}$$

Where l - lower bound of the modal class.

f_1 - frequency of the modal class.

f_0 - frequency of the preceding class to the modal class.

f_2 - frequency of the succeeding class to the modal class.

c - class interval of the modal class.

For a continuous frequency distribution the calculation of mode involves the following steps.

Step 1: Locate the class having the highest frequency. This class is called the *modal class*.

Step 2: The mode can be determined using the above formula.

Illustration 5.23

For his research on the living standards, a researcher conducted a survey on 100 persons. The distribution of the ages of the persons is attached below.

Age	:	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
Number of People	:	*12*	*18*	*27*	*20*	*17*	*6*

Determine the mode of this distribution.

Solution. The highest frequency = 27. ∴ The modal class is 20-30.

$$\text{Mode} = l + \frac{(f_1 - f_0)c}{2f_1 - f_0 - f_2}$$

Where $l = 20, c = 10, f_0 = 18, f_1 = 27$ and $f_2 = 20$.

$$\therefore \text{Mode} = 20 + \frac{(27-18)\times 10}{2\times 27 - 18 - 20}$$
$$= 20 + \frac{9\times 10}{16}$$
$$= 25.625$$

So Mode is 25.625.

Illustration 5.24

The production per day of a company (in Tons) on 60 days are given below. Calculate the mode.

Production per day	:	*21-22*	*23-24*	*25-26*	*27-28*	*29-30*
Number of days	:	*7*	*13*	*22*	*10*	*8*

Solution. Here the classes are of inclusive type. We have to convert into exclusive class before determining the mode.

Class bounds	Number of days
20.5-22.5	7
22.5-24.5	13
24.5-26.5	22
26.5-28.5	10
28.5-30.5	8

The maximum frequency is 20 so the modal class is 24.5-26.5

$$\text{Mode} = l + \frac{(f_1 - f_0)c}{2f_1 - f_0 - f_2}$$

Where $l = 24.5, c = 2, f_0 = 13, f_1 = 22, f_2 = 10$

$$\text{Mode} = 24.5 + \frac{(22-13)\times 2}{2\times 22 - 13 - 10}$$
$$= 25.36$$

So mode = 25.36 tons.

5.3.1 Graphical location of Mode

Mode from Histogram

Like median, mode can also be located using graph. For locating mode, we use the Histogram. The steps involved in obtaining mode from histogram are

Step 1: Draw a histogram to the given data.

Step 2: Locate modal class (highest bar of histogram).

Step 3: Join diagonally the upper end points of the highest bar to the end points of the adjacent bars.

Step 4: Mark the point of intersection of the diagonals.

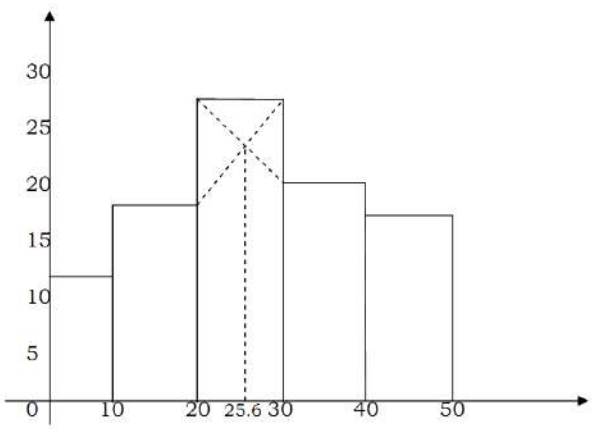
Step 5: Draw perpendicular from this point of intersection to the x-axis.

Step 6: The point where the perpendicular meets the x-axis gives the modal value.

Illustration 5.25

Determine the mode graphically using the data provided in Illustration 5.23

Solution. To locate mode, we have to first draw the histogram.



Mode = 25.6

5.3.2 Merits and demerits of Mode

Like the other measures of central tendency, mode has its own merits and demerits. Mode is widely used in industry.

Merits

1. Mode is easy to calculate and understand. It can sometimes located merely by inspection.
2. Mode is not affected by extreme values.
3. It can be determined for open end classes.
4. Mode is the only average that works with categorical data.

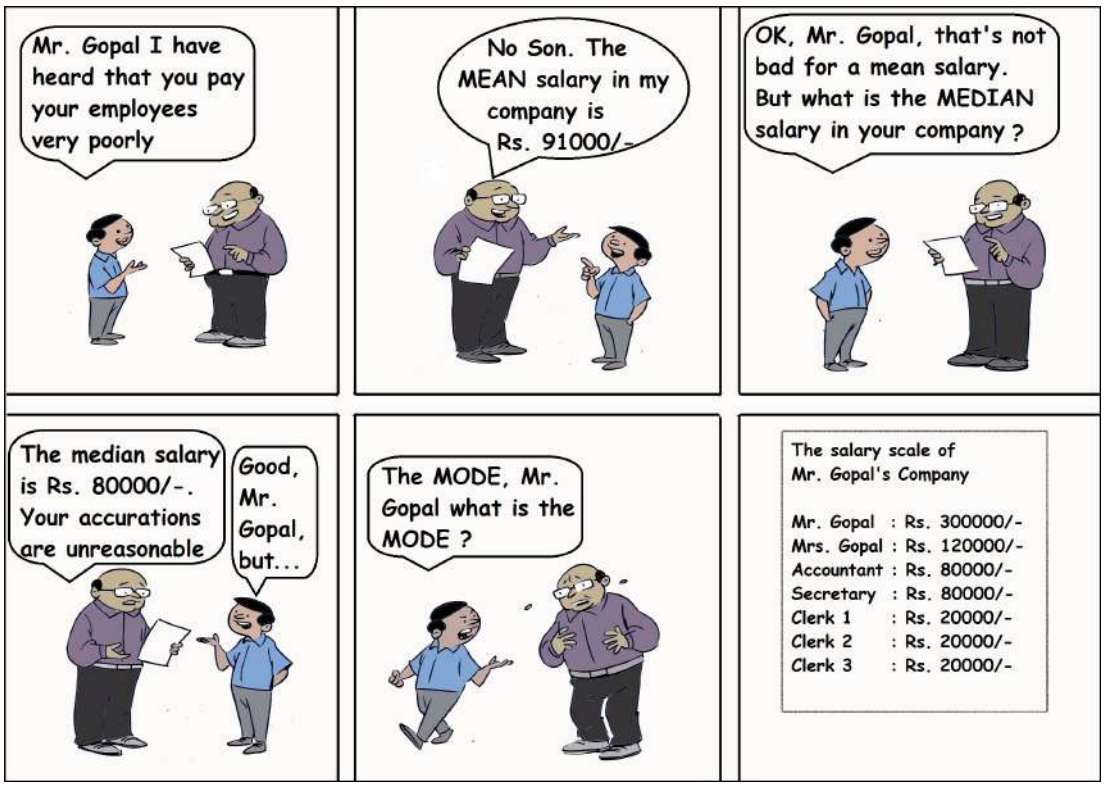
Demerits

1. Mode is sometimes ill-defined. It is not defined rigidly. There are distributions with no mode, one mode, two modes etc.
2. It is not based on all the observations.
3. It is not capable of further mathematical treatments.
4. It is affected by the fluctuations of sampling.

Comparitive table-Mean, Median and Mode

Sl No	Mean	Median	Mode
1.	Defined as the arithmetic average of all the observations.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the data set.

2.	Depend on all the observations.	Doesn't depend on all the observations.	Doesn't depend on all the observations.
3.	Uniquely and comprehensively defined	Can't determine in all the conditions.	Not uniquely defined for multimodal situations.
4.	Affected by extreme values.	Not affected by extreme values.	Not affected by extreme values.
5.	Can be treated mathematically. ie, means of several groups can be combined.	Can't be treated mathematically. ie, medians of several groups can't be combined.	Can't be treated mathematically. ie, modes of several groups can't be combined.
6.	More useful when data is cardinal.	More useful when the data is ordinal.	More useful when the data is nominal.



Empirical relationship among Mean, Median and Mode

For a distribution the empirical relation among mean, median and mode, given by Karl Pearson is,

$$\text{Mean - Mode} = 3(\text{Mean-Median})$$

OR

$$\text{Mode} = 3\,\text{Median}-2\,\text{Mean}$$

The importance of this relation is that we can estimate the value of any one of them by knowing the values of the other two. However, this relationship is true if the distribution is moderately asymmetric.

Illustration 5.26

From a partially destroyed data, it was obtained that the mode of the distribution is 63 and median is 77. Calculate the mean of the data.

Solution. We have the empirical relation,

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

$$\text{Mean} - 63 = 3(\text{Mean} - 77)$$

$$2\text{Mean} = 168$$

$$\text{Mean} = 84$$

Know your progress

1. The ages of workers in a firm are, 40, 50, 30, 20, 25, 35, 30, 30, 20 and 30. Find the modal age.
2. A shoe shop has sold 100 pairs of shoes of a particular brand on a certain day with the following distribution. Find the mode of the distribution.

Size of shoe	:	4	5	6	7	8	9	10
No of pairs	:	10	15	20	35	16	3	1

3. The daily wage of workers in a factory are given below. Determine the modal wage.

Wages (Rs.)	:	200-250	250-300	300-350	350-400
Number of workers	:	21	29	19	39
	:	400-450	450-500	500-550	550-600
	:	43	94	73	68

4. Calculate the mode the following distribution

Production per day (in tons)	:	21-22	23-24	25-26	27-28	29-30
Number of days	:	7	13	22	10	8

5.4 Geometric Mean(GM)

We have already studied how the mean, median and mode can be used as tools for obtaining averages. Sometimes, when we are dealing with quantities that changed over a period of time, we need to know an average rate of change, the mean, median or mode is inappropriate, because it may give the wrong answer. Here we need the use of another average called Geometric Mean.

The Geometric mean of 'n' observations of a series is the n^{th} root of the product of the n observations. If there are two observations, then geometric mean is the square root of the product of the two observations. If there are three observations, geometric mean is the cube root of the product of the three observations, etc.

> The Geometric Mean is the n^{th} root of the product of n observations in the data set.

Geometric Mean of a raw data

Consider a raw data of 'n' observations $x_1, x_2, x_3, \ldots, x_n$. The geometric mean of the data is given by,

$$\begin{aligned} GM &= \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} \\ &= (x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{\frac{1}{n}} \end{aligned}$$

The geometric mean of n items $x_1, x_2, x_3, \ldots, x_n$ is given by

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} = (x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{\frac{1}{n}}$$

Illustration 5.27

A textile has shows the following percentage increase in profit over the last 5 years.

year	:	*2008*	*2009*	*2010*	*2011*	*2012*
Percentage Increase	:	*5*	*10.5*	*9*	*6*	*7.5*

What is the average percentage of increase for the last five years.

Solution. Here the data gives the percentage increase of profits. The average income in 5 years is the GM of $105, 110.5, 109, 106$ and 107.5

$$\begin{aligned} GM &= \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} \\ &= \sqrt[5]{105 \times 110.5 \times 109 \times 106 \times 107.5} \\ &= 107.58 \end{aligned}$$

So average percentage of increase is 7.58 %.

Illustration 5.28

A town's population is increased at a constant annual rate from 40000 in 2006 to 42436 in 2008.

(a) Find the annual percentage increase.

(b) Find the population size in 2007.

Solution. *Given,*

The population in 2006 = 40000 The population in 2008 = 42436

(a) Le x be the annual rate of increase. Then,

$$40000 \times (1 + \frac{x}{100}) \times (1 + \frac{x}{100}) = 42436$$

$$\Rightarrow \quad 40000 \times (1 + \frac{x}{100})^2 = 42436$$

$$\Rightarrow \quad (1 + \frac{x}{100}) = 1.03 \Rightarrow x = 3$$

Therefore, annual percentage rate of increase = 3%

(b) Population in 2007 is the GM of 40000 and 42436.
So Population in 2007 $= \sqrt{40000 \times 42436} = 41200$

OR

Population in $2007 = 40000 \times 1.03 = 41200$

Uses and limitations of Geometric mean

The following are some of the special uses of geometric mean.

1. It is useful for calculating the average percentage increase or decrease, ratios, etc.
2. In the construction of index numbers, geometric mean is considered to be the best average.
3. The importance of GM is that it gives less weightage to the extreme values. Hence the influence of very small and very large values is minimized. In other words, small values get more weightage and large values get less weightage.

Some of the limitations of geometric mean are,

1. If some observations are negative, the Geometric mean will not be calculated.
2. If any one or more observations are zero, the calculation of geometric mean become meaningless because the product of the observations will always zero and hence the GM will be zero.

5.5 Harmonic Mean (HM)

Consider two places A and B.The distance between the places A and B is $30\,km$. A man travelled from A to B on his car at a speed of $60\,Km/hr$ and returns at a speed of $40\,Km/hr$. What will be his average speed?

We know that speed is the ratio between distance travelled and the time taken to travel the distance (ie speed = distance ÷ time). We are given that

Distance	Speed	Time taken to travel the distance
30 Km	60 Km/hr	$\frac{30}{60} = 0.5$ hr
30 Km	40 Km/hr	$\frac{30}{40} = 0.75$ hr
Total - 60 Km		1.25 hr

$$\begin{aligned}\text{Average speed} &= \frac{\text{Distance}}{\text{Time}} \\ &= \frac{60}{1.25} \\ &= 48\,\text{Km/hr}\end{aligned}$$

Now, calculate the AM and GM of the speeds, we get

$$AM = \frac{60+40}{2} = 50\,Km/hr$$
$$GM = \sqrt{60 \times 40} = 48.99\,Km/hr$$

Clearly, here the AM and GM can't used to determine the average speed, because the average speed is 48 Km/hr. The average we used here to find the average speed is Harmonic Mean. The harmonic mean of 60 and 40 is the reciprocal of the mean of the reciprocals of the observations.

$$\begin{aligned}ieHM &= \frac{1}{\frac{1}{2}\left(\frac{1}{60}+\frac{1}{40}\right)} \\ &= \frac{2 \times 60 \times 40}{60+40} \\ &= 48\,Km/hr\end{aligned}$$

The Harmonic Mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given observations.

Harmonic Mean of a raw data

Let $x_1, x_2, x_3, \ldots, x_n$ be the series of 'n' observations. The harmonic mean, HM of this series is given by

$$\begin{aligned}\frac{1}{HM} &= \frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)\\ &= \frac{1}{n}\sum\frac{1}{x}\end{aligned}$$

Therefore

$$HM = \frac{n}{\sum\frac{1}{x}}$$

> The harmonic Mean of a raw data is given by
>
> $$HM = \frac{n}{\sum\frac{1}{x}}$$

Illustration 5.29

In a cycle race competition, the speeds of 5 participants are 15 Km/hr, 18 Km/hr, 20 Km/hr, 22 Km/hr and 17 Km/hr. Find the average speed.

Solution. The average speed is best represented by the harmonic mean. The HM of the 5 observations is given by

$$\frac{1}{HM} = \frac{1}{n}\sum\frac{1}{x}$$

From the data given

$$\begin{aligned}\frac{1}{n}\sum\frac{1}{x} &= \frac{1}{5}\left(\frac{1}{15}+\frac{1}{18}+\frac{1}{20}+\frac{1}{22}+\frac{1}{17}\right)\\ &= 0.0553\end{aligned}$$

Therefore

$$HM = \frac{1}{0.0553} = 18.08$$

Average speed = 18.08 Km/hr

Uses and limitations of Harmonic Mean

Harmonic mean is specially used in the computation of average speed under various conditions. In the calculation of HM, smaller values get more weightage. Hence HM is suitable for a highly varying series.

The Harmonic Mean can't be determined if one of the values is zero.

Know your progress

1. Calculate the Geometric and Harmonic means of the following series of weekly expenditure of a batch of students. 125, 130, 75, 10 and 45.
2. Point out the mistake in the following statement, "A person goes x to y on cycle at 20 Km/hr and returns at 24 Km/hr. His average speed was 22 km/hr."
3. A train travels first 300 kilometers at an average speed of 40 Km/hr and further travels the same distance at an average speed of 30 Km/hr. What is the average speed?

Relation among Arithmetic Mean, Geometric Mean and Harmonic Mean

The AM, GM and HM are called the mathematical averages. The relations among them are,

1. $AM \geq GM \geq HM$, for a set of positive values

   When all the observations are the same, then $AM = GM = HM$.

2. $(GM)^2 = AM \times HM$,for two values

   ie, $GM = \sqrt{AM \times HM}$

   Activity

   For the observations $2, 4, 8, 12$ and 16, examine whether $AM \geq GM \geq HM$.

Illustration 5.30

The AM of two numbers is 10, the GM of those numbers is 8. Find the HM.

Solution. We have,

$$\begin{aligned}(GM)^2 &= AM \times HM \\ 8^2 &= 10 \times HM \\ HM &= \frac{64}{10} \\ &= 6.4\end{aligned}$$

5.6 Partition values- Quartiles, Deciles and Percentiles

Partition values are measures that divide the data into several equal parts. There are three types of partition values-Quartiles, Deciles and Percentiles. Quartiles divide the data into four equal parts, deciles divide the data into ten equal parts while percentiles divide the data into hundred equal parts. The method of computing partition values is similar to the method of computing median.

Quartiles

As mentioned earlier, quartiles are those values which divide a data into four equal parts. There are three quartiles, denoted by Q_1, Q_2 and Q_3. The first quartile Q_1 is the value for which 25% of the observations are below Q_1 and 75% are above it. The third quartile Q_3 is the value for which 75% of the observations are below Q_3 and 25% are above it. For Q_2, 50% are above Q_2 and 50% are below it. ie, Q_2 divides the data into two equal parts. Hence Q_2 is nothing but the Median.

Calculation of Quartiles

(i) Quartiles from a raw data

Let there be 'n' observations. Arrange them in ascending order of magnitude. Then,

$$Q_1 = \text{value of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item in the series}$$

$$Q_3 = \text{value of } \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{ item in the series}$$

Illustration 5.31

The daily wages (in rupees) of 7 workers are 300, 350, 400, 425, 450, 500 and 600. Compute the first and third quartiles.

Solution. The wages in ascending order are, 300, 350, 400, 425, 450, 500, 600. Here $n = 7$

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

Therefore

$$Q_1 = 2^{\text{nd}} \text{ observation} = 350$$

Also

$$\frac{3(n+1)}{4} = 3 \times 2 = 6$$

Hence

$$Q_3 = 6^{\text{th}} \text{ observation} = 500$$

Illustration 5.32

Compute Q_1 and Q_3 for the given data 9, 13, 14, 7, 12, 17, 8, 10, 6, 15, 18, 21, 20

Solution. Arranging the given data in ascending order, we get 6, 7, 8, 9, 10, 12,

13, 14, 15, 17, 18, 20, 21. Here, $n = 13$.

$$
\begin{aligned}
Q_1 &= \text{Value of } \frac{n+1}{4}^{\text{th}} \text{ item} \\
&= \text{Value of } \frac{13+1}{4}^{\text{th}} \text{ item, ie, } 3.5^{\text{th}} \text{ item} \\
&= \text{Value of } 3^{\text{rd}} \text{ item } + 0.5(\text{Value of } 4^{\text{th}} \text{ item } - \text{Value of } 3^{\text{rd}} \text{ item}) \\
&= 8 + 0.5(9 - 8) \\
&= 8.5 \\
Q_3 &= \text{Value of } \frac{3(n+1)}{4}^{\text{th}} \text{ item} \\
&= \text{Value of } 3 \times 3.5^{\text{th}} \text{ item, ie, } 10.5^{\text{th}} \text{ item} \\
&= \text{Value of } 10^{\text{th}} \text{ item} + 0.5 \times (\text{Value of } 11^{\text{th}} \text{ item} - \text{Value of } 10^{\text{th}} \text{ item}) \\
&= 17 + 0.5(18 - 17) \\
&= 17.5
\end{aligned}
$$

Hence $Q_1 = 8.5$ and $Q_3 = 17.5$

(ii) Quartiles from a discrete frequency distribution

For computing quartiles, prepare the less than frequency table. Let N be the total frequency. Then,

$$
\begin{aligned}
Q_1 &= \text{observation having cumulative frequency } \frac{(N+1)}{4} \\
Q_3 &= \text{observation having cumulative frequency } \frac{3(N+1)}{4}
\end{aligned}
$$

Illustration 5.33

The heights in inches of 49 persons are given below

Height	:	*58*	*59*	*60*	*61*	*62*	*63*	*64*	*65*	*66*
No of persons	:	*2*	*3*	*6*	*15*	*10*	*5*	*4*	*3*	*1*

Calculate the first and third quartiles.

Solution. To find the quartiles, we have to prepare the cumulative frequency table

Height	Frequency	Cumulative frequency
58	2	2
59	3	5
60	6	11
61	15	26
62	10	36
63	5	41
64	4	45
65	3	48
66	1	49
	$N = 49$	

Here $N = 49$

$$Q_1 = \text{Observation having cumulative frequency } \frac{N+1}{4}$$
$$= \text{observation having cumulative frequency } 12.5$$
$$= 61$$
$$Q_3 = \text{Observation having cumulative frequency } \frac{3(N+1)}{4}$$
$$= \text{Observation having cumulative frequency } 37.5$$
$$= 63$$

Hence $Q_1 = 61$ and $Q_3 = 63$.

(iii) Quartiles from a continuous frequency distribution

Prepare the cumulative frequency table. Let N be the total frequency. Locate the classes having cumulative frequencies $\frac{N}{4}$ and $\frac{3N}{4}$. These classes are called the quartile classes. Then Q_1 and Q_3 are given by the formula,

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m_1\right)c_1}{f_1}$$
$$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m_3\right)c_3}{f_3}$$

Where l_1 and l_3 are the lower bounds of quartile classes

f_1 and f_3 are the frequencies of the quartile classes

c_1 and c_3 are the class intervals of the quartile classes

m_1 and m_3 are the cumulative frequencies preceding the quartile classes

Illustration 5.34

The marks of 80 students in an examination are given below. Calculate the lower and upper quartiles.

Marks	:	*0-10*	*10 -20*	*20-40*	*40-60*	*60-80*	*80-100*
No of students	:	*8*	*10*	*22*	*25*	*10*	*5*

Solution. Let us prepare the following cumulative frequency table

Marks	No of students f	Cumulative frequency
0-10	8	8
10-20	10	18
20-40	22	40
40-60	25	65
60-80	10	75
80-100	5	80
	$N = 80$	

Here $N = 80$, $\frac{N}{4} = 20$ and $\frac{3N}{4} = 60$. Therefore, the quartile classes are 20-40 and 40-60. Then

$$Q_1 = l_1 + \frac{\left(\frac{N}{4} - m_1\right)c_1}{f_1}$$

where $l_1 = 20, c_1 = 20, f_1 = 22$ and $m_1 = 18$. So

$$Q_1 = 20 + \frac{(20-18)\times 20}{22}$$
$$= 20 + \frac{2\times 20}{22}$$
$$= 21.8$$

Also

$$Q_3 = l_3 + \frac{\left(\frac{3N}{4} - m_3\right)c_3}{f_3}$$

where $l_3 = 40, c_3 = 20, f_3 = 25$, and $m_3 = 40$. So

$$\begin{aligned} Q_3 &= 40 + \frac{(60-40)\times 20}{25} \\ &= 20 + \frac{20\times 20}{25} \\ &= 56 \end{aligned}$$

Hence $Q_1 = 21.8$ and $Q_3 = 56$.

Deciles and Percentiles

Deciles are those values which divide a distribution into ten equal parts. There are 9 deciles.

Percentiles are those values which divide a distribution into hundred equal parts. There are 99 percentiles.

Median is the 5^{th} decile and 50^{th} percentile.

Know your progress

1. Compute the quartiles for following data 13, 14, 7, 12, 17, 8, 10, 6, 15, 18, 21, 20.
2. The following data relate to the sizes of shoes sold at a store during a given week. Find the upper and lower quartiles.

Size of shoes	:	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
Frequency	:	10	18	22	25	40	15	10	8	7

3. Compute Q_1 and Q_3 for the distribution

Marks	:	0-10	10-20	20-30	30-40	40-50
Number of students	:	3	10	17	7	6
	:	50-60	60-70	70-80		
	:	4	2	1		

5.7 Box plot

The **box plot** of a data is a graphical representation based on its quartiles, as well as its smallest and largest values. It attempts to provide a visual image of the data distribution. The box plot is also referred to as *Box and Whisker plot or Box and Whisker diagram*.

A box plot is a graph of a data set that consists of a line extending from the minimum value to the maximum value and a box with lines drawn at the first quartile Q_1, the median, and the third quartile Q_3. The lines extending from the box are called whiskers. The perpendicular line in the box is the median.

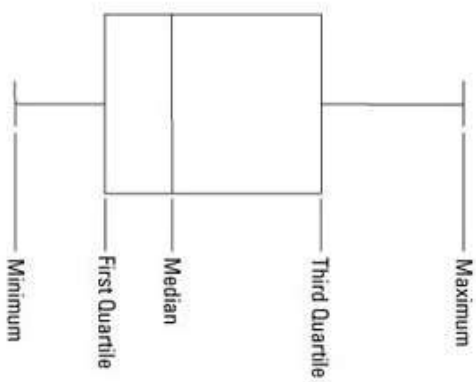


Illustration 5.35

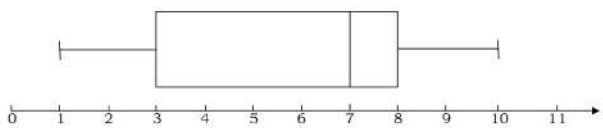
Eleven secretaries were given a test and the scores obtained by them are as follows, 8, 7, 6, 9, 1, 3, 10, 3, 8, 4 and 7. Represent the data using a box plot.

Solution. Arranging the data in ascending order,

$$1, 3, 3, 4, 6, 7, 7, 8, 7, 9, 10$$

Number of observations, $n = 11$. To draw the box plot, we have to determine the following.

$$\begin{aligned}
\text{Minimum value} &= 1 \\
Q_1 &= \tfrac{n+1}{4}^{\text{th}} \text{ item} \\
&= \tfrac{11+1}{4}^{\text{th}} \text{ item} \\
&= 3^{\text{rd}} \text{ item} \\
&= 3 \\
\text{Median} &= \tfrac{n+1}{2}^{\text{th}} \text{ item} \\
&= 6^{\text{th}} \text{ item} \\
&= 7 \\
Q_3 &= \tfrac{3(n+1)}{4}^{\text{th}} \text{ item} \\
&= 9^{\text{th}} \text{ item} \\
&= 8 \\
\text{Maximum value} &= 10
\end{aligned}$$



Let us sum up

Central tendency is the tendency of the observations in a data to concentrate around a central value. A measure of central tendency is a single value which is used to represent an entire set of data. In Statistics there are various measures of central tendencies. They are Arithmetic Mean, Median, mode, Geometric Mean and Harmonic Mean. The AM, GM and HM are also known as the mathematical averages and Median and Mode are known as the positional averages.

The AM of a set of observations is their sum divided by the number of observations. The weighted mean enables us to calculate an average value that takes into account the importance of each value with respect to the overall total. Median of a distribution is the value of the variable which divides it into two equal parts. Mode is the value that is represented most often in the data. Geometric mean is the nth root of the product of n items of a series. It is useful in calculating the average percentage increase or decrease. Harmonic mean of series is the reciprocal of the arithmetic mean of the

reciprocals of the observations. It has a specific application in the computation of average speed under various conditions.

AM, GM and HM are mathematical in nature and measure the quantitative characteristics of the data. To measure the qualitative characteristics of the data the other measures, namely median and mode are used.

Partition values are measures which divide the data into several equal parts. Quartiles divide the data into four equal parts, deciles into ten equal parts and percentiles into hundred equal parts.

Learning outcomes

Check your understanding in the following

- Meaning of central tendency and various measures of central tendency.
- Arithmetic mean and calculation of arithmetic mean. Properties, merits and demerits of arithmetic mean.
- Median and calculation of median. Merits and demerits of median.
- Mode and calculation of mode. Merits and demerits of mode.
- Geometric mean-calculation, uses and limitations.
- Harmonic mean-calculation, uses and limitations.
- Partition values-Quartiles, deciles and percentiles.
- Box plot.

Evaluation items

1. The variable value in a series which divides the series into halves is called _____
2. The two ogives intersect at _____
3. Out of all measures of central tendency _____ is the only measure which is not unique.
4. Second quartile is same as _____

5. The Am and HM of a distribution are 10 and 8.1, then GM is
(81, 0.81, 9, 2)

6. If each value of a series is multiplied by 10; the median is _____

(a) not affected, (b) 10 times of the original median value,
(c) one-tenth of the original median value, (d) increased by 10

7. The GM of two values can be calculated if

(a) both the values are positive, (b) one of the two values is zero,
(c) one of them is negative, (d) both of them are zero

8. Find the mean, median and mode

(a) 12, 17, 17, 39, 41, 44, 54, 67, 82
(b) 123, 115, 98, 107, 115, 109, 113, 98
(c) 1.2, 1.4, 1.9, 2.0, 2.4, 3.5, 3.9, 4.3, 5.2

9. The weights of eight babies at birth in kgs are 2.4, 2.8, 3.2, 1.9, 2.7, 4.2, 3.8, 2.2. Find the average weight at birth

10. The frequency of reaction of an individual to a certain stimulant were measured by a psychologist are 0.53, 0.46, 0.5, 0.49, 0.52, 0.53, 0.44, 0.55. Determine the mean reaction time of the individual to the stimulant.

11. For a set of 8 scores, the mean is 5. If seven of these scores are 9, 3, 4, 5, 6, 4, 7, what must be the remaining score?

12. The average income of a person for the first five days of a week is Rs. 350 per day and if he works for the first six days in a week, his average income becomes Rs. 400 per day. Find his income for the sixth day.

13. The average age of 30 students in a class is 17 years. 4 students with an average age of 14.5 years left the class and 5 new students joined the class whose ages were 23, 25, 17.5, 19.5, and 21 years respectively. Calculate the average age of the class.

14. The average of 60 students in a class is 18 years. 7 students left the class, whose average age was 16.25 years and 5 new students joined the class whose ages were respectively 19, 18, 21.5, 23.75 and 14 years respectively. Find out the present average age of the students in the class

15. The mean of 100 items is 49. It was discovered that 3 items which should have been 60, 70 and 80 were wrongly taken as 16, 17 and 18. What is the corrected mean?

16. Ten coins were tossed together and the number of the resulting from them was observed. The operation was performed 1050 times and the frequencies obtained for different number of trials (x) are shown in the following table. Calculate the arithmetic mean

x :	0	1	2	3	4	5	6	7	8	9	10
f :	2	8	43	133	207	260	213	120	54	9	1

17. The following table gives the number of children of 150 families in a village.

No of children :	0	1	2	3	4	5
No of families :	10	21	55	42	15	7

Find the average number of children per family.

18. The arithmetic mean of the following distribution is 115.61, find out the missing frequency.

Values :	110	112	113	117	120	125	128	130	Total
Frequency :	25	?	13	?	14	8	6	2	100

19. The mean of the following series is 68.25, find out the missing value

Wages :	50	58	60	65	70	–	80	100
Number of workers :	2	20	5	35	8	10	16	4

20. The following is the data of weekly earnings of 200 persons. Calculate the mean.

Weekly wages	:	1000-1200	1200-1400	1400-1600	1600-1800
Number of persons	:	3	21	35	57
	:	1800-2000	2000-2200	2200-2400	2400-2600
	:	40	24	14	6

21. An agency interviewed 200 persons as part of an opinion poll. The following distribution represents the age of the persons interviewed. Calculate the average age.

Age	:	80 -89	70 -79	60 -69	50 -59	40 -49	30 -39
Frequency	:	2	2	6	20	56	40
	:	20 -29	10 -19				
	:	42	32				

22. Calculate the average temperature from the following data

Temp ($°C$)	:	$-40 - -30$	$-30 - -20$	$-20 - -10$	-10-0
Number of days	:	10	28	30	42
	:	0-10	10-20	20-30	
	:	65	180	10	

23. The following distribution is the ages of 360 patients getting medical treatment in a hospital on a day. Calculate the mean of the ages.

Age	:	10-20	20-30	30-40	40-50	50-60	60-70
No of patients	:	90	50	60	80	50	30

24. The following table gives the distribution of total household expenditure (in rupees) of manual workers in a city. Find the average expenditure.

Expenditure	No of workers	Expenditure	No of workers
100 - 150	24	300 - 350	30
150 - 200	40	350 - 400	22
200 - 250	33	400 - 450	16
250 - 300	28	450 - 500	7

25. A group of 134 girls and 166 boys appeared for an English examination. The boys obtained a mean score of 68.5 and the mean score for all the students was 64.35. Calculate the mean score for the girls.

26. The mean height of 25 male workers in a factory is 161 cms and the mean height of 35 female workers in the same factory is 158 cms. Find the mean height all the workers in the factory.

27. The grades of a student in lab, lecture and recitation parts of a Physics course were 71, 78 and 89 respectively. a) If the weight accorded the grades are 2, 4 and 5 respectively, what is the appropriate average grade? b) What is the average grade if equal weights are assigned?

28. The mark of a student in Economics, Statistics and Commerce are 82, 68 and 89 respectively. If the weights accorded to these marks are 2, 3 and 5 respectively, then a) Which is the appropriate average? b) Calculate the value of the above average

29. Find the median of the observations 46, 64, 87, 41, 58, 77, 35, 90, 55, 92, 33. If 92 is replaced with 99 and 41 with 43 in the above data, find the new median.

30. The numbers 4, 7, 8, x+1, 2x - 3 , 15, 16, 20 are arranged in ascending order. Find the value of x if the median is 12.5

31. The median of the following observations arranged in ascending order 11, 12, 14, 18, x+2, x+4, 30, 32, 35, 41 is 24. Find x.

32. Find out the median.

Income:	1000	1500	3000	2000	2500	1800
No of persons:	24	26	16	20	6	30

33. Calculate the median.

Values:	1	2	3	4	5	6	7
Frequency:	6	8	10	14	13	9	4

34. The marks out of 60 obtained by 58 students in a certain examination are given below. Calculate the median

Marks	No of students	Marks	No of students
15 - 20	4	40 - 45	8
20 - 25	5	45 - 50	9
25 - 30	11	50 - 65	6
30 - 35	6	65 - 70	4
35 - 40	5		

35. Determine the median wage for the following distribution

Wages	200 - 300	300 - 400	400 - 500	500 - 600	600 - 700
No of labourers	3	5	20	10	5

36. Determine the median from the following distribution

Marks	No of students	Marks	No of students
45 - 50	10	20 - 25	31
40 - 45	15	15 - 20	24
35 - 40	26	10 - 15	15
30 - 35	30	5 - 10	10
25 - 30	42	0 - 5	5

37. The following frequency distribution represents the commission earned (in rupees) by 100 salesmen employed at different branches of a multilevel marketing company.

Commission	No of people	Commission	No of people
150 - 158	5	186 - 194	20
159 - 167	16	195 - 203	15
168 - 176	20	204 - 212	3
177 - 185	21		

Find out the median.

38. Car batteries have a nominal voltage of 12 volts. A manufacturer tested a sample of 160 batteries and obtained the following results.

Voltage	11.8-12.0	12.1-12.3	12.4-12.6	12.7-13.1	13.2-13.4	13.5-13.7
No of batteries	18	26	12	78	18	8

Calculate the median voltage.

39. The following data represent the expenditure (in million of rupees) of 45 municipal corporations. Find the median.

Class	10-20	21-31	32-42	43-53	54-64	65-75
Frequency	2	8	15	7	10	3

40. Obtain the missing frequency of the following distribution where the value of the median is 86.

Class	40-50	50-60	60-70	70-80	80-90	90-100	100-110
Frequency	2	1	6	6	f	12	5

41. Which measure is the most helpful to a shoe maker?

42. The following data gives the sizes of 15 shoes sold at a shop on a particular day 5, 7, 9, 9, 8, 5, 6, 8, 7, 7, 7, 9, 2, 7. Estimate the modal size.

43. The table shows the scores obtained when a die is thrown 60 times. Calculate the mode.

Score :	1	2	3	4	5	6
Frequency:	12	9	8	13	9	9

44. Find the mode of the following distribution

Shoe size :	2	3	4	5	6
Frequency :	8	15	23	20	14

45. The table below shows the number of men in various age groups with some form of paid employment in a village. Find the modal age.

Age	14-20	20-30	30-40	40-50	50-60	60-70	70-90
Frequency	12	14	26	35	23	5	1

46. Calculate the mode of the following distribution.

Age	20-24	25-29	30-34	35-39	40-44	45-49
Frequency	20	24	32	28	20	26

47. The live span of 80 electric bulbs were recorded in hours to the nearest hour and grouped as shown below. Find the mode.

Life (in hours)	660-669	670-679	680-689	690-699	700-709	710-719	720-729	730-739
No ob Bulbs	4	5	12	24	15	10	7	3

48. The frequency table below illustrates the heights of the 87 trees in a public park.

Height (M)	2-5	6-9	10-13	14-17	18-21	22-25	26-29
No of Trees	4	6	14	26	25	7	3

Calculate the mean, median and mode of the heights.

49. A variable has nine values, 4, 11, 25, 37, 11, 26, 35, 11 and p.

a) (i) Which measure of central tendency can be found without knowing the value of p?

(ii) Find out the value of that measure.

b) It is known that p is more than 30, then

(i) Which other measure of central tendency can now be found?

(ii) Find out that measure.

c) If the remaining measure of central tendency has the value 22, find the value of p.

50. Each of the sentences below uses the word average. For each sentence indicate which measure of central tendency the word average refers to.
a) The average car being driven in India is white.
b) The average number of children in 8 classes of a primary school is 32.25.
c) Half number of students secured less than the average marks in an exam.

51. The set of integers p, 13, 18, 29 and 29 has mode 29, mean 20 and median 18. Without calculating the value of p, find out the mean, median and mode of the following distributions
a) p+2, 15, 20, 31, 31.
b) p - 5 , 8, 13, 24, 24
c) 2p, 26, 36, 58, 58
d) p/2 , 6.5, 9, 14.5, 14.5

52. For a set of observations, mode is twice the mean. If median is 23, find out the mean.

53. The value of mean and mode of a frequency distribution are 50 and 45 respectively. Find the median.

54. The average monthly salary of 100 employees of a factory is Rs.4500. If the median salary is found to be Rupees 4900/- per month, find out the appropriate modal value.

55. Find the GM of 4.2 and 16.8.

56. Determine the GM of 4 and 36.

57. Find the GM of 8, 16 and 62.5 .

58. Find the GM of 18, 16, 22, 12

59. 'Kapil & Sons' have shown the following percentage of increase in their business over the last 5 years:

year :	2008	2009	2010	2011	2012
Increase :	7%	8%	10%	12%	18%

Find the average increase in percentage.

60. Find the HM of
a) 2, 3, 6
b) 3.2, 5.2, 4.8, 6.1, 4.2

61. A motor car covered a distance of 50 km in 4 phases, the first phase at 50 k.p.h, the second phase at 20 k.p.h, the third and fourth at 40 k.p.h and 25 k.p.h.respectively. Calculate the average speed.

62. A cyclist travels from his house to school at a speed of 10 km/hr and returns to house at 14 km/hr. Which is the most suitable average to find the average speed? Calculate it.

63. Find the average speed of an aeroplane which flies among the four sides of a square at a speed 100, 200, 300 and 400 km/hr respectively.

64. For the values 8, 6, 4 and 3, prove that $AM > GM > HM$.

65. The following are the scores obtained by 9 students in a class test
38, 7, 43, 25, 20, 15, 12, 18, 11
Calculate the quartiles.

66. Calculate the quartiles.

Marks :	20	30	40	50	60
No of students:	4	16	20	18	11

67. Find the first and third quartiles for the distribution,

Height (In Inches) :	58	59	60	61	62	63	64	65	66
No of person :	2	3	6	15	10	5	4	3	1

68. Calculate the quartiles from the following distribution.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No of Students	5	7	8	12	28	22	10	8

69. Find the quartiles of the following distribution.

Scores	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No of Students	1	3	11	21	43	32	9

70. Draw a box plot for the following data.
13, 14, 7, 12, 17, 8, 10, 6, 15, 18, 21, 20.
$(Q_1 = 8.5, Q_3 = 17.5, Q_2 = 13)$

Answers:

1) Median	2) Median	3) Mode
4) Median	5)9	6) b
7) a	8) (a)41.44, 41, 17	8)(b) 109.75, 111, 98 and 115
8)(c)2.87, 2.4, 1.46	9)2.9 Kg.	10) 0.5025
11)2	12) 650	13)18
14) 18.32	15)50.69	16)5.01
17) 2.35	18) 22,10	19)75
20) 1768	21)35.8	22)4.29
23)39.84	24) 266.25	25)59.21

26) 159.25
27) (a). 81.73
27) (b). 79.33
28) (a). Weighted AM,
28 (b). 81.30
29) 58,58
30)9
31) 21
32)1800
33) 4
34)38
35) 467.5
36)27.74
37) 180.36
38)12.80
39) 40.67
40)10
41) Mode
42) Mode=7
43)4
44) 4
45)44.29
46) 32.83
47)695.21
48) Mean=15.64,
48) Median=16
48) Mode=17.19
49) (a) i) Mode,
49) (a) ii) 11
49) (b) i) Median,
49) (b) ii) 25
49) (c) p = 38
50) (a) Mode
50) (b) Mean
50)(c) Median
51) (a) Mean = 22
Median = 20
Mode = 31
51) (b) Mean = 15,
Median = 13,
Mode = 24
51) (c) Mean = 40
Median = 36
Mode = 58
51) (d) Mean = 10,
Median = 9,
Mode = 14.5
52)17.25
53)48.33
54)5700
55)8.40
56)12
57)20
58)16.61
59)10.39%
60) (a) 3
60)(b) 4.484
61)29.63
62)11.67
63)192
64) AM = 5.25
GM = 4.9
HM = 4.57
63) $AM > GM > HM$
65) Q1 = 11.5,
Q2 = 18,
Q3 = 31.5
66) Q1 = 30
Q2 = 40,
Q3 = 50
67) Q1 = 61,
Q3 = 63
68) Q1 = 34.17
Q2 = 46.43,
Q3 = 56.82
69) Q1 = 66.64
Q2 = 75.08
Q3 = 82.94

Introduction

The various measures of central tendency that we studied in the last chapter describe the characteristic of the entire data by means of a single value that is the central value of the distribution. But with these measures alone we cannot form a clear idea about the distribution. Consider the following series :

A	28	29	30	30	32
B	30	30	30	30	30
C	1	2	30	30	87

Compute the mean, median and mode of the two data sets. As you see from your results, the two datasets have the same mean, median and mode, all are equal to 30. The three data sets also contain same number of elements. But the three sets are different. What is the main difference among them?

Three data sets have same average, but they differ in deviation of observations. ie, Data set C is more scattered than in data sets A and B. The values in C are more spread out. ie, They lie farther away from their mean than in the case of A and B. Thus, The degree to which numerical data tend to spread about an average value is called the dispersion, or variation, of the data.

Dispersion

> Dispersion is the degree of scatter or variation of the variable about a central value.

There are sevaral measures of variability or dispersion, they are

1. Range
2. Quartile Deviation
3. Mean Deviation

4. Standard Deviation

Properties of Measure of Dispersion

A good measure of dispersion should possess the following properties

- It should be rigidly defined.
- It should be based on all observations.
- It should be simple to understand and easy to calculate.
- It should be amicable to further algebraic treatment.
- It should not be unduly affected by extreme values.

Importance of Measuring Dispersion

- It tells about the variability of a data.
- It enables us to compare two or more distributions.
- It is of great importance in advanced statistical analysis.

6.1 Range

Range(R) is the difference between the highest value (H) and lowest value (L) in a set of data. In symbols $R = H - L$

- Higher value of range implies higher dispersion and vice versa.

Merits

- It is the simplest measure of dispersion.
- It is simple to understand and easy to calculate.

Demerits

- It gives importance to the two extreme values only,and therefore it may be unduly influenced by extreme values.
- It is not a reliable measure of dispersion on many occasions.

Illustration 6.1

Given below the data of price (Rupees) of 1gm gold in the first week of August 2013.

Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*
2530	*2500*	*2430*	*2560*	*2680*	*2600*

Calculate range of price in the week

Solution.

$$\begin{aligned}\text{Range} &= \text{highest value} - \text{lowest value}\\ &= 2680 - 2430\\ &= 250 \text{ rupees}\end{aligned}$$

Activity

Look at the series 10,30,50,100

- Calculate range if 30 is replaced by 20.
- Calculate range if 100 is not present in it.

6.2 Quartile Deviation (QD)

One drawback of range is that it depends only on extreme values and we cannot find range in open end classes. As a modification the difference between 3^{rd} quartile and 1^{rd} quartile is called inter quartile range and half of the difference between 3^{rd} quartile and 1^{st} quartile is called Quartile Deviation (QD) . Quartile deviation is also called semi-inter quartile range.

$$\text{Inter quartile range} = Q_3 - Q_1$$

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

Where $Q_3 = 3^{rd}$ quartile, $Q_1 = 1^{st}$ quartile

Quartile Deviation for a Raw Data

To find QD for a raw data consisting of n values ,first arrange the data in ascending order of magnitude, then find Q_1 and Q_3 as explained in Chapter 5 ,Section 6

$$\text{QD} = \frac{Q_3 - Q_1}{2}$$

Illustration 6.2

A statistical data was collected from 11 school children on the number of hours they spend watching television in one week. The data is given below 3, 8.5, 12, 9, 16.5, 9, 14, 20, 18, 19, 20. Find quartile deviation

Solution.

Arrange the data in ascending order of magnitude as follows

3, 8.5, 9, 9, 12, 14, 16.5, 18, 19, 20, 20

$$Q_1 = \text{value of } \left(\frac{n+1}{4}\right)^{th} \text{ item}$$

$$= \text{value of} \left(\frac{11+1}{4}\right)^{th} \text{ item}$$

$$= 3^{rd}\text{item}$$

$$= 9$$

$$Q_3 = \text{value of } \left[3\left(\frac{n+1}{4}\right)\right]^{th} \text{ item}$$

$$= \text{value of } 9^{th}\text{item.}$$

$$= 19$$

$$QD = \frac{19 - 9}{2}$$
$$= 5$$

Quartile Deviation for a Discrete Frequency Distribution

Consider a discrete frequency distribution consisting of N values. To find QD first arrange the values in the ascending order of magnitude, then find Q_1 and Q_3 for a discrete frequency distribution as explained in chapter 5, section 9. Then $QD = \frac{Q_3 - Q_1}{2}$.

Illustration 6.3

Calculate Q D from the following data

Score	*5*	*10*	*15*	*20*	*25*	*30*
Number of students	*4*	*7*	*15*	*8*	*7*	*2*

Solution.

Less than cumulative frequency table is given below

Scores (x)	Number ofstudents(f)	Less than cum. frequency
5	4	4
10	7	11
15	15	26
20	8	34
25	7	41
30	2	43
	$N = 43$	

$$Q_1 = \text{Value of} \left(\frac{N+1}{4}\right)^{\text{th}} \text{item}$$
$$= \text{Value of} \left(\frac{43+1}{4}\right)^{\text{th}} \text{item}$$
$$= \text{Value of } 11^{\text{th}} \text{ item}$$
$$= 10$$
$$Q_3 = \text{Value of} \left(\frac{3(N+1)}{4}\right)^{\text{th}} \text{item}$$
$$= \text{Value of} \left(\frac{3(43+1)}{4}\right)^{\text{th}} \text{item}$$
$$= \text{Value of } 33^{\text{rd}} \text{ item}$$
$$= 20$$

$$QD = \frac{Q_3 - Q_1}{2}$$
$$= \frac{20 - 10}{2}$$
$$= 5$$

Quartile Deviation for a Continuous Frequency Distribution

You have already learnt to find Q_3 and Q_1 in a continuous frequency distribution in Chapter 5, Section 6

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

Illustration 6.4

A survey of domestic consumption of electricity in a colony gave the following distribution of units consumed

units	*Below 100*	*100-200*	*200-300*	*300-400*	*400-500*	*500-600*	*600-700*	*above 700*
No.of consumers	*20*	*21*	*30*	*46*	*20*	*25*	*16*	*10*

Find quartile deviation.

Prepare less than cumulative frequency table as explained in unit 3

No.of units	No.of consumers(f)	Less than frequency
Below 100	20	20
100-200	21	41
200-300	30	71
300-400	46	117
400-500	20	137
500-600	25	162
600-700	16	178
700& above	10	188
	$N = 188$	

$$Q_1 = l_1 + \left(\frac{\frac{N}{4} - m_1}{f_1} \right) \times C_1$$

$$\frac{N}{4} = \frac{188}{4}$$

$$= 47$$

The value 47 lies in 200-300 class

Therefore Q_1 lies in 200-300 class

$$Q_1 = 200 + \frac{47 - 41}{30} \times 100$$

$$= 200 + 20$$

$$= 220$$

$$Similarly, Q_3 = l_3 + \left(\frac{\frac{3N}{4} - m_3}{f3} \right) \times C_3$$

$$3 \times \frac{N}{4} = 141$$

The value 141 lies in 500-600 class

$$Q_3 = 500 + \left[\frac{141 - 137}{25}\right] \times 100$$
$$= 500 + 16$$
$$= 516$$

$$QD = \frac{516 - 220}{2}$$
$$= 148$$

Merits

- It can be calculated for open end classes.
- It is not unduly affected by extreme values.

Demerits

- It is not based on all observations.
- It is not capable of further algebraic treatments.

Know your progress

1. The following data represents the profit(in lakhs) of 50 business men in a large city. Find QD of the data

Profit	20-29	30-39	40-49	50-59	60-69
No.of Business man	8	12	20	7	3

2. The heights of students (in inches) is given below. Calculate QD of the data
   Heights: 55, 54, 57, 67, 60, 61, 58, 63

3. Prices of shares of a company were as under from Monday to Saturday for 40 weeks. Find QD of shares

Day	Mon	Tue	Wed	Thu	Fri	Sat
Price	150	200	190	210	230	180
No.weeks	5	5	8	10	5	7

Activity

Collect a data on height of students in your class and find quartile deviation and range. Interpret your results.

6.3 Mean Deviation (MD)

Even though quartile deviation is an improvement over range, it depends only on two values Q_3 and Q_1. As a solution we try to introduce another measure of dispersion which depends on all values in a series. Mean deviation is the arithmetic mean of absolute deviation of the observations from an assumed average.

Mean Deviation for a Raw Data

Let us consider a raw data of 'n observations. mean deviation is calculated as follows

Mean deviation= $\frac{\sum |x-A|}{n}$, where A is any average

Compute an average (A) which is required to calculate mean deviation (mean or median or mode)

Find the absolute deviation of the observations from the average to each value. It is denoted by $|x-A|$

The arithmetic mean of these absolute deviations is mean deviation

Illustration 6.5

Eleven students were selected and asked how many hours each of them studied the day before the final examination in statistics. Their answers were recorded here 8, 11, 5, 4, 5, 0, 2, 6, 9, 3, 2. Calculate Mean Deviation about mean

Solution.

$$\text{Mean deviation about mean} = \frac{\sum |x - \text{Mean}|}{n}$$

Duration of study(x)	$\lvert x - \text{Mean}\rvert$
8	$\lvert 8-5\rvert = 3$
11	$\lvert 11-5\rvert = 6$
5	$\lvert 5-5\rvert = 0$
4	$\lvert 4-5\rvert = 1$
5	$\lvert 5-5\rvert = 0$
0	$\lvert 0-5\rvert = 5$
2	$\lvert 2-5\rvert = 3$
6	$\lvert 6-5\rvert = 1$
9	$\lvert 9-5\rvert = 4$
3	$\lvert 3-5\rvert = 2$
2	$\lvert 2-5\rvert = 3$
$\sum x = 55$	$\sum \lvert x - \text{mean}\rvert = 28$

$$\text{Mean} = \frac{\sum x}{n} = \frac{55}{11} = 5$$

$$\text{Mean deviation} = \frac{\sum |x - \text{Mean}|}{n} = \frac{28}{11} = 2.54.$$

Mean Deviation for a Discrete Frequency Distribution

For a discrete frequency distribution consisting of N observations, the procedure and formulae is as follows

Mean deviation= $\frac{\sum f|x-A|}{N}$ where A is any average

Step 1

Find the average (A) which is required to calculate mean deviation.

Step 2

Find the absolute deviations of the observations from the average to each value.ie $|x-A|$.

Step 3

Multiply $|x-A|$ by their frequency f. Thus we get $f|x-A|$ to each observations.

Step 4

The arithmetic mean of these values is the mean deviation.

Illustration 6.6

The following table shows the number of books read by students in a literature class consisting of 28 students,in a month.

No. of Books	*0*	*1*	*2*	*3*	*4*
No.of students	*2*	*6*	*12*	*5*	*3*

Calculate mean deviation about mode of number of books read

Solution. Mean deviation about mode = $\frac{\sum f|x-\text{Mode}|}{N}$

mode=2 (the value which has maximum frequency).

No. of books	$\|x - \text{mode}\|$	f	$f\|x - \text{mode}\|$
0	$\|0-2\| = 2$	2	4
1	$\|1-2\| = 1$	6	6
2	$\|2-2\| = 0$	12	0
3	$\|3-2\| = 1$	5	5
4	$\|4-2\| = 2$	3	6
		$N = 28$	$\sum f\|x - \text{mode}\| = 21$

$$\text{Mean deviation about mode} = \frac{21}{28}$$
$$= 0.75$$

Mean Deviation for a Continuous Frequency Distribution

To find mean deviation, convert the continuous frequency distribution into discrete frequency distribution by taking middle values of each classes. The procedure and formulae are same in a discrete frequency distribution

Illustration 6.7

Calculate mean deviation from median of the following data

Height in cm	*100-120*	*120-140*	*140-160*	*160-180*	*180-200*
No.of students	*4*	*6*	*10*	*8*	*5*

Solution. Mean deviation about median= $\frac{\sum f|x - \text{median}|}{N}$

Find middle values of each class taken then prepare a table as given below

Class	Mid. value(x)	Freq. (f)	Cum. Freq.	$\|x - \text{median}\|$	$f\|x - \text{median}\|$
100-120	110	4	4	$\|110-153\| = 43$	172
120-140	130	6	10	$\|130-153\| = 23$	138
140-160	150	10	20	$\|150-153\| = 3$	30
160-180	170	8	28	$\|170-153\| = 17$	136
180-200	190	5	33	$\|190-153\| = 37$	185
Total		33			661

$$N = 33\ ,\ \sum f|x - \text{median}| = 661$$

$\frac{N}{2} = \frac{33}{2} = 16.5$

The class having cum.frequency 16.5 is 140-160

Therefore 140-160 is median class

$$\begin{aligned}\text{median} &= l + \frac{\frac{N}{2} - m}{f} \times c \\ &= 140 + \frac{16.5 - 10}{10} \times 20 \\ &= 153\end{aligned}$$

$$\begin{aligned}\text{Mean deviation about median} &= \frac{\sum f|x - \text{median}|}{N} \\ &= \frac{661}{33} \\ &= 20.03\end{aligned}$$

Merits

- It is based on all observations.
- It can be calculated from any value.
- It is not much affected by extreme values.

Demerits

- Ignoring signs of deviations may create artificiality.
- Computation is not much easier.

Know your progress

1. The following table shows the frequency distribution of grade points

Grade point	8	6	4	2
No.of students	4	20	5	1

Find mean deviation of the grade points about mode.

2. Find mean deviation about mean of the data which relate to the sales of 100 companies.

Sales(Rs.thousand)	40-50	50-60	60-70	70-80	80-90	90-100
No.of days	10	15	25	30	12	8

Activity

Look at the data 5, 8, 3, 2, 12

- Calculate mean deviation about mean
- Calculate mean deviation about median
- Calculate mean deviation about mode
- Try yourself to find which of them is minimum

6.4 Standard Deviation(SD)

In order to have a more meaningful measure to compute the variability of a data we use SD. The concept of SD was introduced by Karl Pearson. A relatively small standard deviation indicates high degree of uniformity in the data with not much variation of individual values from their mean. In particular, if all the observations are equal the SD is equal to zero.

Standard deviation (SD) is defined as the positive square root of the mean of squares of deviations from the arithmetic mean. It is denoted by Greek letter σ (sigma).It cannot be negative. It is the best measure of dispersion.

Properties

1. The minimum value of SD is zero.
2. The sum of squared deviation is minimum when taken about mean.

Standard deviation for a raw data

Let us consider a raw data of n observations. SD can be calculated as follows

$$SD, \sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

For computation purpose we can use another formula derived from the above. The procedure is discussed below

Step 1

Calculate arithmetic mean $\overline{x}$ of the given data

Step 2

Calculate x^2 for each observation.

Step 3

Calculate $\frac{\sum x^2}{n}$, where 'n' is no. of values

Step 4

Calculate $\sigma = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2}$

$$SD, \sigma = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2}$$

Illustration 6.8

A survey was conducted for the number of road accidents in a major city during 11 successive weeks. The results are given below 8 ,6 ,3 ,0 ,5 ,9 ,2 ,1 ,3 ,5 ,2 . Clculate SD of road accidents

Solution.

$$SD = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2}$$

x	8	6	3	0	5	9	2	1	3	5	2	$\sum x = 44$
x^2	64	36	9	0	25	81	4	1	9	25	4	$\sum x^2 = 258$

$$\begin{aligned}\bar{x} &= \frac{44}{11} \\ &= 4 \\ \sigma &= \sqrt{\frac{258}{11} - 4^2} \\ &= \sqrt{23.45 - 16} \\ &= 2.73\end{aligned}$$

Standard Deviation for a Discrete Frequency Distribution

For a discrete frequency distribution consisting of N $(\sum f)$ observations, the procedure and formulae is discussed below

Step 1

Calculate the arithmetic mean of the data, $\bar{x} = \frac{\sum fx}{N}$

Step 2

Calculate x^2 and fx^2 for each value

Step 3

Calculate $\sum fx^2$

Step 4

Calculate SD $= \sqrt{\frac{\sum fx^2}{N} - (\bar{x})^2}$

$$SD, \sigma = \sqrt{\frac{\sum fx^2}{N} - (\bar{x})^2}$$

Illustration 6.9

25 students were given an arithmetic test. The time in minute to complete the

test is as follows

Time in minutes	*1*	*2*	*3*	*4*	*5*
No.of students	*4*	*3*	*10*	*5*	*3*

Calculate SD of their time to complete the test

Solution.

Prepare a table given below

x	f	x^2	fx	fx^2
1	4	1	4	4
2	3	4	6	12
3	10	9	30	90
4	5	16	20	80
5	3	25	15	75
	$N = 25$		$\sum fx = 75$	$\sum fx^2 = 261$

$$\begin{aligned}\overline{x} &= \frac{\sum fx}{N} \\ &= \frac{75}{25} \\ &= 3\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fx^2}{N} - \overline{x}^2} = \sqrt{\frac{261}{25} - 3^2} \\ &= \sqrt{10.44 - 9} \\ &= \sqrt{1.44} \\ &= 1.2\end{aligned}$$

Standard Deviation for a Continuous Frequency Distribution

To find SD, convert the continuous frequency distribution into discrete frequency distribution by taking mid values of each class. The procedure and formulae are same as in the case of a discrete frequency distribution

Illustration 6.10

A study of 100 engineering companies gives the following information

Profit(in crore)	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
No.of companies	*8*	*12*	*20*	*30*	*20*	*10*

Find SD of profit earned.

Solution.

Computation table

x	f	x^2	fx	fx^2
5	8	25	40	200
15	12	225	180	2700
25	20	625	500	12500
35	30	1225	1050	36750
45	20	2025	900	40500
55	10	3025	550	30250
	$N = 100$		$\sum fx = 3220$	$\sum fx^2 = 122900$

$$\bar{x} = \frac{\sum fx}{N}$$
$$= \frac{3220}{100}$$
$$= 32.2$$
$$\sigma = \sqrt{\frac{\sum fx^2}{N} - \bar{x}^2}$$
$$= \sqrt{\frac{122900}{100} - (32.2)^2}$$
$$= \sqrt{192.16}$$
$$= 13.86$$

Merits

- It is rigidly defined and based on all observations.
- It is capable of further algebraic treatments.
- It is less affected by the fluctuations of sampling than other measures of dispersion.

Demerits

- It is difficult to calculate.
- It is impossible to find it in open end classes.

 *Know your progress*

1. Calculate SD of daily income of 10 persons in rupees given below 227, 235, 255, 269, 292, 299, 312, 321, 333, 348
2. The following table gives the distribution of expenditure of 100 families in a village. Calculate SD

Income	0-1000	1000-2000	2000-3000	3000-4000	4000-5000	5000-6000
No.of families	18	26	30	12	10	4

Activity

If a student scored equal scores in 6 subjects in an examination, calculate SD and interpret your result.

Activity

Observe the data 40 ,42 ,38 ,44 ,46 ,48 ,50

- Compute SD of the data.
- Add 3 to each value then find new SD.

- Subtract 3 to each value then find new SD.
- Multiply 3 by each value then find new SD.
- Divide 3 by each value then find new SD.
- Interpret your findings

Variance

The term variance was introduced by R.A.Fisher. It is defined as the square of SD

$$\text{Variance} = \sigma^2$$

$$Or, \text{SD} = \sqrt{\text{variance}} = \sigma$$

It has many applications in advanced statistical analysis

6.5 Relative measures of dispersion

The measures of dispersion so far discussed are called as absolute measures of dispersion (Range, QD, MD, SD). Because they measures the variability or deviation of values in a data and expressed in the original unit of that data. If the units are different, absolute measures of dispersion is not suitable for a direct comparison of two or more series because it may result an incorrect conclusion regarding variation of the data. For these reasons, a measure of dispersion which is independent of unit of measurement is suggested. Such a measure is called relative measure of dispersion. It is the ratio of absolute measures of dispersion to an average with which the measure of dispersion is computed

Features of relative measure:

- It is a ratio.
- It is a pure number.
- It is free from the unit of measurement of observations.

- It is used for comparing two or more sets of data.

The two important relative measures of dispersion we are going to discuss are

1. Coefficient of variation
2. Coefficient of quartile deviation

Coefficient of Variation (CV)

It is the most commonly used measure to compare the consistency or stability between two or more sets of data. Coefficient of variation is defined as SD divided by its arithmetic mean expressed in percentage

$$Coefficient\ of\ variation,\ CV = \frac{\sigma}{\overline{x}} \times 100$$

Suppose, between two groups we have to find the consistent group. First find the CV for each group. The group with less CV is considered to be the consistent group. It is free from the unit of original measurements.

Illustration 6.11

In two factories A & B located in same industrial area, the weekly wage(in Rupees) and standard deviations are as follows

Factory	*Average*($\overline{x}$)	$SD(\sigma)$	*Number of workers*
A	*500*	*5*	*476*
B	*600*	*4*	*524*

1. *Which factory pays larger amount as weekly wages?*
2. *Which factory has greater consistency in individual wages?*

Solution. Here $n_1 = 476, \overline{x}_1 = 500\ \sigma_1 = 5, n_2 = 524, \overline{x}_2 = 600, \sigma_2 = 4$

1.

$$\text{Total wages paid by factory A} = 500 \times 476$$
$$= 238000$$
$$\text{Total wages paid by factory B} = 600 \times 524$$
$$= 314400$$

2.

$$\text{CV of factory A} = \frac{\sigma_1}{\bar{x}_1} \times 100$$
$$= \frac{5}{500} \times 100$$
$$= 1$$
$$\text{CV of factory B} = \frac{\sigma_2}{\bar{x}_2} \times 100$$
$$= \frac{4}{600} \times 100$$
$$= 0.67$$

Since total wages of factory B is greater than factory A, factory B pays larger amount as weekly wages. Since CV is greater for A compared to B, factory A has greater variability.

Know your progress

Prices of a particular commodity in 5 years at two different cities in Kerala and Tamilnadu are as follows

Tamilnadu(RS)	20	22	19	22	23
Kerala(RS)	18	12	10	20	15

Which state has stable price?justify your answer

Activity

Form two groups in your class, collect the grades scored by statistics examination and then find which of the two groups is more stable

.

Coefficient of Quartile Deviation

Quartile deviation is an absolute measure of dispersion .The relative measure corresponding to quartile deviation is coefficient of quartile deviation.

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It can be used to compare the degree of variation in different situations

6.6 Covariance

It is a measure of strength of linear relationship between two variables. Covariance indicates whether the variables are positively related or negatively related in a bivariate distribution.

$COV(x,y) = \frac{\sum(x-\overline{x})(y-\overline{y})}{n}$, where '$n$ is the number of observations in a data.

For computation purpose we can use another formulae

$$COV(x,y) = \frac{\sum xy}{n} - \overline{x} \times \overline{y}$$

If $COV(x,y)$ is positive the variables move in same direction. If it is negative they move in opposite direction .If two variables are unrelated, then $COV(x,y)$ should be zero.

Illustration 6.12

The following data pertains to the length of service(in years) and the annual income in thousands for a sample of 8 employees in an industry. Find covariance.

Length of service(x)	6	8	9	10	11	13	15	16
Annual income (y)	14	17	15	18	16	22	25	25

Solution. $COV(x,y) = \frac{\sum xy}{n} - \bar{x} \times \bar{y}$

x	y	xy
6	14	84
8	17	136
9	15	135
10	18	180
11	16	176
13	22	286
15	25	375
16	25	400
$\bar{x} = 11$	$\bar{y} = 19$	$\sum xy = 1772$

$$\text{Cov(x,y)} = \frac{1772}{8} - 11 \times 19$$
$$= 12.5$$

Let us sum up

In this chapter we have discussed the concept of dispersion and some absolute and relative measures of dispersion . The absolute measures of dispersion helps us to calculate the deviation of values among themselves or deviation of values from an average. Among this, Range and quartile deviation attempt to measure deviation of values among themselves. But mean deviation and standard deviation attempt to measure variation of values from their average. The relative measures of dispersion are free from the unit of measurement of observations and are used to compare two sets of data. A measure of relationship between two variables is also introduced using covariance.

Learning outcomes

After transaction of this unit, the learner:-

- recognises the importance of measuring dispersion.
- explains and evaluates the measures of dispersion-Range, QD,MD and SD.
- distinguishes absolute and relative measures of dispersion.

Evaluation Items

1. If a constant is subtracted from each observation then variance _____
2. If each observation is divided by 10 then SD of the new observations is _____
3. If first 25% of values in a data is 20 or less and last 25% is 50 or more then QD is _____
4. If the lowest value of a set is 9 and its range is 57 then the highest value is _____
5. If the CV of a distribution is 50 and its SD 20,the arithmetic mean shall be _____
6. The SD of five observations 5 ,5 ,5 ,5, 5 is _____
7. The mean of squared deviation from mean is called _____
8. Indicate whether the following statements are true or false
   (a) Range is the best measure of dispersion
   (b) QD is more suitable in the case of open-end distributions
   (c) Absolute measure of dispersion can be used for the purpose of comparison
   (d) Mean deviation is least when deviations are taken from median
   (e) SD can never be negative

9. Distinguish between absolute and relative measures of dispersion

10. Relative measures of dispersion are used for comparison , why?

11. What do you understand by dispersion of a set of values? How do you measure dispersion?

12. What are the different absolute measures of dispersion?

13. Why SD is called the best measures of dispersion?

14. Suppose each measurement in a distribution is multiplied by 2 what happens to the following?

    (a) Mean of the distribution.

    (b) Variance of the distribution.

    (c) SD of the distribution.

15. For a data consisting of 9 observations, the sum is 9 values is 360. And the sum of squares of deviation taken from mean is 288. Find

    (a) standard deviation

    (b) $C.V.$

16. The number accidents occurred due to careless driving on a busy road during 7 days in week is reported below

No.of accidents	7	10	4	7	9	3	2

    Determine mean deviation about mean.

17. Following are the responses from 52 students to the question about how much money they spend everyday

Money spent	No.students
5	2
10	5
15	10
20	7
25	8
30	5
35	6
40	4
45	5

Find quartile deviation.

18. For a newly created post the manager interviewed the candidates in 5 days. The number of candidates in each day were 16,19,15,15,&14 respectively . Find variance

19. A factory produces two types of electric lamps A & B. In an experiment relating to their life the following results were obtained.

Length of life (Hrs)	Number of lamps (A)	Number of lamps (B)
500-700	5	4
700-900	11	30
900-1100	26	12
1100-1300	10	8
1300-1500	8	6

(a) Which bulb A or B has more average life.

(b) Which is more consistent.

20. Annual tax paid by certain employees is given below

Tax Paid in thousand	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of Emplyees	18	30	46	28	20	12	6

Find semi inter quartile range.

21. For 108 randomly selected higher secondary students the following I Q distribution was obtained

I Q	90-98	99-107	108-116	117-125	126-134
No. students	6	10	25	15	4

Find standard deviation of IQ.

22. Prices of a particular commodity in 5 years in two cities are given below

Price in city A	20	22	19	23	16
Price in city B	10	20	18	12	15

Which city has stable prices using coefficient of quartile deviation.

23. The number of motor cycle accident cases reported in a hospital is as follows

Age	0-10	10-20	20-30	30-40	40-50	50-60
No. of accidents	15	49	37	20	6	1

Find mean deviation about median.

24. The following income distribution was obtained from 1000 persons

Income more than	1000	900	800	700	600	500	400	300	200
No.of persons	0	50	110	200	400	650	825	950	1000

Find quartile deviation.

25. Nine students of B.Com class of a college have obtained the following marks in statistics out of 100. Calculate standard deviation and variance

Sl no	1	2	3	4	5	6	7	8	9
Marks in Statistics	5	10	20	25	40	42	45	48	70

26. The age group and the number of diabetic patients treated in a hospital is listed below

Age	20	30	40	50	60	70	80
No.of .patients	3	10	30	40	35	10	7

Determine mean deviation about median.

27. The number of television sets sold in a week from a home appliance show room is given below

Number of sets	5	6	7	8	4	3	1

Find standard deviation and variance.

28. In a village ,25% of the persons earned more than Rs.10000/- whereas 75% earned only more than Rs.5000/-. Calculate relative and absolute values of dispersion.

29. The following table relates to advertisement expenditure(y) and sales (x) of a company for a period of 10 years. Calculate $COV(x,y)$

Sales(x)	50	5	50	40	30	20	20	15	10
Adv.Exp.(y)	700	200	650	500	450	400	300	250	210

Answers

1. does not change 2. $1/10$th of SD of original obs 3. 15 4. 66 5. 40 6. 0 7. SD 8. (a) false (b)true (c) false (d)true (e) true 14. (A) .2 times (B) 4 times (C) .2 times 15. SD=5.66,CV=6.28 16. MD about mean= 2.57 17. QD=10 18. var=2.96 19. (A) bulb A has more average life (B) bulb A is more consistent 20. semi inter quartile range=5.48 21. SD=9.36 22. coe:QD of cityA=0.125,coe: QD of cityB=0.267 therefore cityA has more stable price 23. MD about median=8.906 24. QD=137.5 25. SD=19.46, variance=378.69 26. MD about median=10 27. SD=2.231,var=4.979 28. QD= 2500,coefficient of QD=0.33 29. $COV(x,y) = 2694.44$

7 Skewness and Kurtosis

Introduction

In the last two units we discussed the most important characteristics (Central Tendency and Dispersion) of a statistical data. Knowledge of averages and measures of dispersion is of great importance in assessing the properties of a group of data. An average represents the central tendency of the distribution and the dispersion measures the degree of variation around the central value.

These measures of central tendency and dispersion do not reveal whether the dispersal of values on either side of an average is symmetrical or not. If the spread of the frequencies is the same on both sides of the centre point of the curve then that curve is called symmetric curve.

Mirror image is symmetric

In a study of a frequency distribution it would be of great help to know whether it would give a symmetric curve and if not, what extend it would deviate from symmetry. The word skewness indicates the absence of symmetry in a data set. We study skewness to have an idea about the shape of the frequency curve which we can draw with the help of given data.

Two or more distributions may differ in terms of flatness or peakedness of their frequency curve. The method of measuring this characteristic of a data set is termed as Kurtosis. Thus the word Kurtosis relates to the degree of flatness or peakedness of a data.

7.1 Skewness

Consider the following frequency distributions which give the scores obtained by the students who were studying in Commerce, Humanities and Science groups in a higher secondary school.

1.	Scores	:	0-10	10-20	20-30	30-40	40-50	50-60
	No. of students	:	5	8	15	15	8	5
2.	Scores	:	0-10	10-20	20-30	30-40	40-50	50-60
	No. of students	:	4	7	16	11	7	5
3.	Scores	:	0-10	10-20	20-30	30-40	40-50	50-60
	No. of students	:	5	7	11	16	7	4

Mean and variance of the above distributions are same but they differ widely in their overall appearance as we can seen from the following diagrams

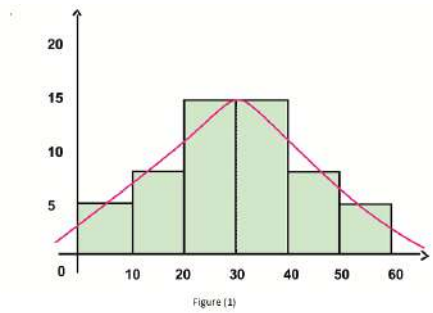


Figure (1)

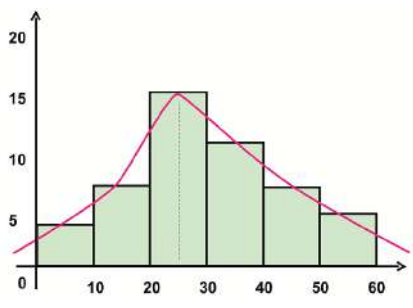


Figure (2)

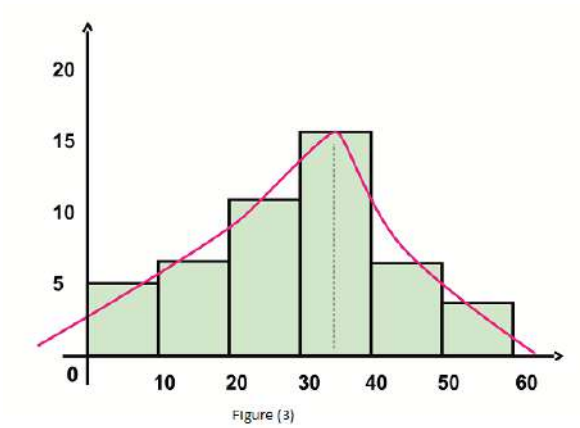
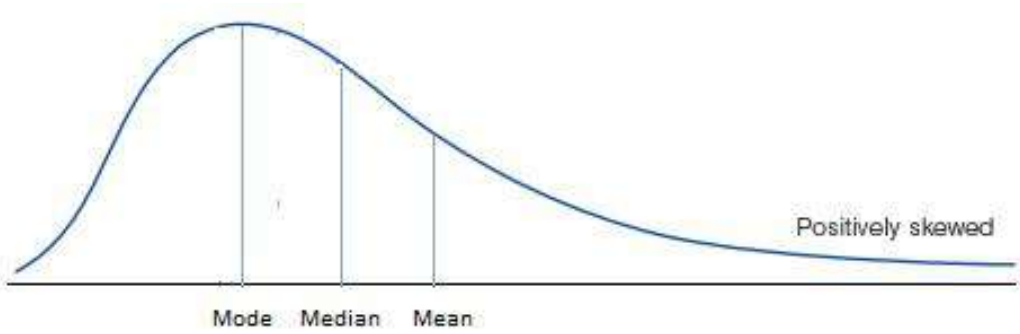


Figure (3)

In figure(1) the right and left of mode(highest ordinate) are perfect mirror images of one another. They are called as symmetric distributions.
In figure (2) you can see more items on the right side of mode and have a longer tail to the right side of mode. Similarly in figure (3), more items on the left of mode and have longer tail to the left of mode. When frequency curves are drawn for different frequency distributions, there is an apparent common characteristic, which is striking to the eye, called symmetry or lack of symmetry. The lack of symmetry of a distribution is known as skewness. Here it is clear that figures (2) and (3) are not symmetric or they are skewed. Thus, there are 2 types of skewness 1) Positive skewness 2) Negative skewness.

Positive Skewness

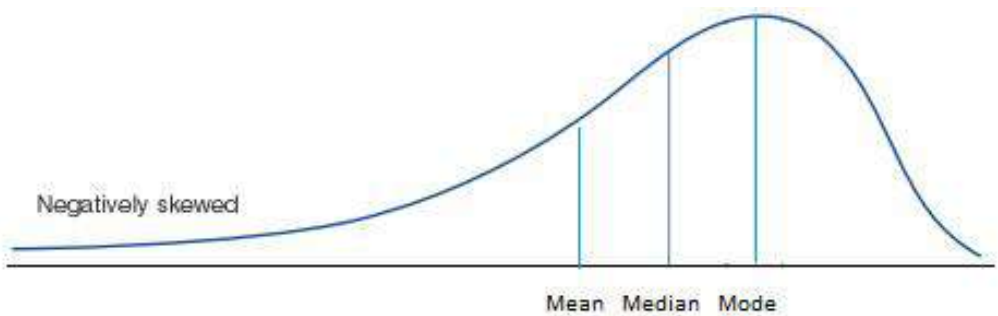
The frequency curve is said to be positively skewed if more items are found to the right side of the mode. In this case the frequency curve will have a longer tail to the right. Also Mode, Median and Mean are in the ascending order of their magnitude.



,]

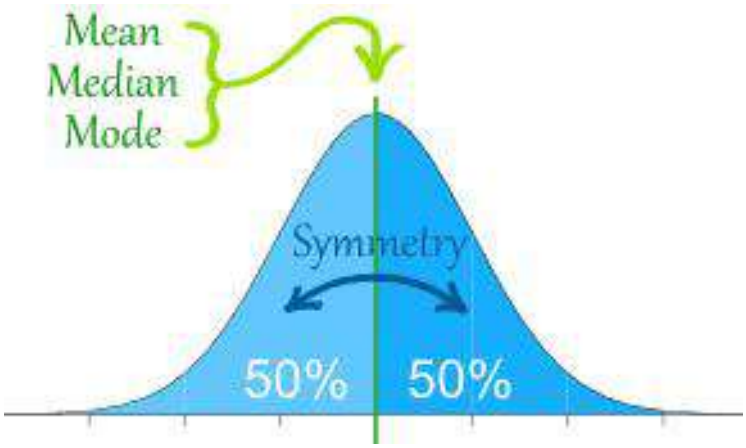
Negative Skewness

The frequency curve is said to be negatively skewed if more items are found to the left side of the mode. In this case the curve will have a longer tail to the left and Mode, Median and Mean are in the descending order of their magnitude.



For a symmetric frequency curve

- Mean = Median = Mode
- The highest ordinate (mode) divides the total area under the curve into two equal parts
- Quartiles are equidistant from median. Ie. Q_3 – Median = Median – Q_1.



For a positively skewed data, Q_3 – Median > Median – Q_1
For a negatively skewed data, Q_3 – Median < Median – Q_1

In a symmetric distribution the mean, median, mode coincides. The symmetry of such a curve can be changed by adding some items to one side or other side of the mode. If there are more items to the right of the highest ordinate of the frequency curve, the curve is said to be positively skewed. Addition of more items on right side pulls AM away from mode towards the right. As there are some more items on the right side of the mode, the median is also slightly pulled from mode towards right. Similar thing will happen if more items are added to the left. Thus skewness has got the effect of pulling the arithmetic mean and the median away from the mode, sometimes towards the right and sometimes towards the left.

Activity

Calculate Mean, Median, Mode and Quartiles of three distributions given below and Compare the results.

1.	Scores	:	0-10	10-20	20-30	30-40	40-50
	No. of students	:	5	12	16	12	5
2.	Scores	:	0-10	10-20	20-30	30-40	40-50
	No. of students	:	5	20	12	8	5
3.	Scores	:	0-10	10-20	20-30	30-40	40-50
	No. of students	:	5	8	12	20	5

Also draw frequency curves of above three distributions and compare them.

Activity

Collect data regarding

(a) Marks in various subjects in the class.
(b) Heights of students in the class.
(c) Data on income of parents
(d) Consumption of electricity

Compute Mean, Median and Mode. Comment about the skewness of the distributions.

7.2 Measures of Skewness

Measures of skewness indicate to what extend and in what direction the distribution of variable departs from symmetry of a frequency curve. It gives the information about the shape of the distribution and the degree of variation on either side of the central value.

There are 3 important measures of Skewness

1. Karl Pearsons coefficient of Skewness
2. Bowleys coefficient of Skewness
3. Coefficient of Skewness based on Moments.

Karl Pearsons Coefficient of Skewness

For a symmetrical data we have mean = mode. If there is skewness, the mean is separated from the mode. If mean - mode is positive then there is positive Skweness and if mean - mode is negative then there is negative skewness. This measure is a natural way of measuring skewness.

The above measure is inadequate for comparison since they may be same for two distributions with different dispersions. So we use coefficient of skewness which is numerical figures independent of units of measurement.

Karl Pearson derived the coefficient of skewness denoted by S_k and is defined as

Karl Pearsons Coefficient of Skewness,

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\overline{X} - \text{Mode}}{\sigma}$$

For positive skewness, $S_k > 0$ (since Mean> Mode).

For negative skewness, $S_k < 0$ (since Mean < Mode).

For symmetry $S_k = 0$ (since Mean= Mode).

Illustration 7.1

For a distribution Mean=30, Mode=26.8 and variance=64. Find the coefficient of skewness. Interpret the result

Solution. *Given* $\overline{X} = 30$, *Mode* $= 26.8, \sigma^2 = 64, \sigma = 8$.

Karl Pearsons Coefficient of Skewness,

$$S_K = \frac{\overline{x} - Mode}{\sigma} = \frac{30 - 26.8}{8} = 0.40$$

Since $S_K > 0$, *the distribution is positively Skewed.*

Illustration 7.2

For a group of 20 items, $\sum x = 1452$, $\sum x^2 = 144280$ and mode = 63.7. Obtain Karl Pearsons coefficient of skewness.

Solution. Given Mode=63.7,

$$\begin{aligned}\overline{x} &= \frac{\sum x}{n} \\ &= \frac{1452}{20} \\ &= 72.6\end{aligned}$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \\ &= \sqrt{\frac{144280}{20} - \left(\frac{1452}{20}\right)^2} \\ &= 44.08\end{aligned}$$

Karl Pearsons coefficient of Skewness,

$$\begin{aligned}S_K &= \frac{\overline{X} - \text{Mode}}{\sigma} \\ &= \frac{72.6 - 63.7}{44.08} \\ &= 0.2019\end{aligned}$$

Since $S_K > 0$, the distribution is positively skewed

Illustration 7.3

The number of accidents reported at city hospital in a week as follows 40,62,40,25, 40,34 and 60. Calculate Karl Pearsons coefficient of skewness.

Solution. Given

$$\begin{aligned}\overline{x} &= \frac{\sum x}{n} \\ &= \frac{301}{7} \\ &= 43 \\ \sigma &= \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \\ &= \sqrt{\frac{14025}{7} - \left(\frac{301}{7}\right)^2} \\ &= 12.43 \\ \text{Mode} &= 40 (\text{Most frequently occurred item})\end{aligned}$$

Karl Pearsons coefficient of skewness,

$$\begin{aligned}S_K &= \frac{\overline{X} - \text{Mode}}{\sigma} \\ &= \frac{43 - 40}{12.43} \\ &= 0.24\end{aligned}$$

Since $S_K > 0$, the distribution is positively skewed.

Illustration 7.4

Find the coefficient of Skewness by Karl Pearsons method for the following data and comment upon the nature of skewness.

Value	:	*6*	*12*	*18*	*24*	*30*	*36*	*42*
Frequency	:	*4*	*7*	*9*	*18*	*15*	*10*	*3*

Solution.

$$\bar{x} = \frac{\sum fx}{n}$$
$$= \frac{1638}{66}$$
$$= 24.82$$
$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$
$$= \sqrt{\frac{46188}{66} - \left(\frac{1638}{66}\right)^2}$$
$$= 9.16$$
$$\text{Mode} = \text{observation having highest frequency}$$
$$= 24$$

Karl Pearsons coefficient of Skewness,

$$S_K = \frac{\overline{X} - \text{Mode}}{\sigma}$$
$$= \frac{24.82 - 24}{9.16}$$
$$= 0.089$$

Since $S_K > 0$, the given distribution is positively skewed.

Illustration 7.5

The monthly income distribution of 100 persons living in a village is attached

Income(In '000s)	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*	*50-60*
No. of Persons	*12*	*18*	*27*	*20*	*17*	*6*

Determine (a) Mode (b) Standard Deviation (c) Coefficient of Skewness of this distribution.

Solution. The highest frequency = 27, hence the model class is 20-30

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_2 - f_0}\right)c$$

Where $l = 20$, $c = 10$, $f_0 = 18$, $f_1 = 27$ and $f_2 = 20$

$$\begin{aligned} Mode &= 20 + \frac{90}{60} \\ &= 25.625 \\ \text{Mean} &= \frac{\sum fx}{N} \\ &= \frac{2800}{100} \\ &= 28 \end{aligned}$$

$$\begin{aligned} \text{Standard deviation}, \sigma &= \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} \\ &= \sqrt{\frac{98300}{100} - (28)^2} \\ &= \sqrt{199} = 14.11 \end{aligned}$$

Karl Pearsons coefficient of Skewness

$$\begin{aligned} S_K &= \frac{\overline{X} - \text{Mode}}{\sigma} \\ &= \frac{28 - 25.625}{14.11} \\ &= 0.16832 \end{aligned}$$

Since $S_K > 0$, the given distribution is positively skewed.

Know your progress

1. *A sample distribution of 200 employees according to their gross monthly salary drawn by them shows an average salary Rs. 3590, mode Rs. 3660 and Variance Rs. 625. Obtain Karl Pearsons coefficient of skewness and interpret it.*

2. *The sum of 20 observations is 300, the sum of their squares is 5000 and Mode is 15. Find the coefficient of skewness and coefficient of variations.*

Bowleys Coefficient of Skewness

For symmetrical data we have Median $-Q_1 = Q_3 -$ Median. It will not be equal for skewed data , hence the difference $(Q_3 -$ Median$) - ($Median $- Q_1)$ can be used to measure skewness.

Sir Arthur Bowley derived another measure of skewness based on Quartiles which is known as Bowleys coefficient of skewness denoted as S_B and is defined as

Bowleys coefficient of skewness, $S_B = \dfrac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$

The value of S_B lies between -1 to +1.
If $S_B > 0$ the distribution is positively skewed.
If $S_B < 0$ the distribution is negatively skewed.
If $S_B = 0$ the distribution is symmetric.

Note:

1. Bowleys coefficient of skewness is very useful especially in the case of open-end classes
2. Results of S_B and S_k are not to be compared with each other.

Illustration 7.6

For a certain distribution the upper and lower quartiles are 56 and 44 respectively. If the median for the same data is 55 then identify the nature of skewness.

Solution. Given $Q_1 = 44$, $Q_3 = 56$, median $= 55$
Bowleys coefficient of skewness,

$$S_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$
$$= \frac{56 + 44 - 2 \times 55}{56 - 44}$$
$$= -0.83$$

Since $S_B < 0$, the distribution is negatively skewed.

Illustration 7.7

If 25% of the total observations lie above 70 and 50% of the total observations are less than 38 and 75% of the total observations are greater than 30, then find coefficient of skewness and interpret the result.

Solution. Since 25% observations are above 70, we have $Q_3 = 70.$,
Since 50% observations are below 38, we have $Q_2 = 38.$
Since 75% observations are greater than 30, we have $Q_1 = 30.$

Bowleys coefficient of skewness,

$$\begin{aligned} S_B &= \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} \\ &= \frac{70 + 30 - 2 \times 38}{70 - 30} \\ &= 0.6 \end{aligned}$$

Since $S_B > 0$, the distribution is positively skewed.

Illustration 7.8

The Bowleys coefficient of skewness is 0.8, the sum of two quartiles is 80 and median is 30. Find the values of upper and lower quartiles.

Solution. Given $S_B = 0.8$, $Q_1 + Q_3 = 80$, Median $= 30$
We have, Bowleys coefficient of skewness,

$$\begin{aligned} S_B &= \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} \\ 0.8 &= \frac{80 - 2 \times 30}{(80 - Q_1) - Q_1} \\ 0.8 &= \frac{20}{80 - 2Q_1} \end{aligned}$$

$$\text{i.e.,} Q_1 = 27.5,\ Q3 = 52.5$$

Illustration 7.9

Following marks were obtained in Statistics by 15 students. 15, 20, 20, 21, 22, 22, 24, 28, 28, 29, 30, 32, 25, 33 and 35. Calculate quartile coefficient of skewness.

Solution. Arranging the data in ascending order we have

$$15, 20, 20, 21, 22, 22, 24, 25, 28, 28, 29, 30, 32, 33, 35$$

$$Q_1 = \left[\frac{n+1}{4}\right]^{th} \text{ item} = 4^{th} \text{ item} = 21$$

$$\text{Median} = \left[\frac{n+1}{2}\right]^{th} \text{ item} = 8^{th} \text{ item} = 25$$

$$Q_3 = \left[\frac{3(n+1)}{4}\right]^{th} \text{ item} = 12^{th} \text{value} = 30$$

Bowleys coefficient of skewness,

$$S_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

$$= \frac{30 + 21 - 2 \times 25}{30 - 21}$$

$$= 0.11$$

Since $S_B > 0$, the distribution is positively skewed.

Illustration 7.10

The following data shows daily wages of 124 workers of a factory.

Wages(Rs.)	*200*	*250*	*300*	*350*	*400*	*450*	*500*	*550*
No. of workers	*10*	*15*	*18*	*30*	*26*	*15*	*8*	*2*

Find (1) Quartiles (2) Coefficient of Skewness

Solution.

X	C.f
200	10
250	25
300	43
350	73
400	99
450	114
500	122
550	124

$$Q_1 = \left[\frac{n+1}{4}\right]^{th} \text{ item} = 31.25^{th} \text{ item}$$

$$= 300$$

$$\text{Median} = \left[\frac{n+1}{2}\right]^{th} \text{ item} = 62.5^{th} \text{ item}$$

$$= 350$$

$$Q_3 = \left[\frac{3(n+1)}{4}\right]^{th} \text{ item} = 93.75^{th} \text{ item}$$

$$= 400$$

Bowleys coefficient of skewness,

$$S_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$
$$= \frac{400 + 300 - 2 \times 350}{400 - 300}$$
$$= 0$$

Since $S_B = 0$, the distribution is symmetric.

Illustration 7.11

Following data show the ages of family members in a colony consisting of 20 families

Ages(Years)	*0-20*	*20-40*	*40-60*	*60-80*	*80-100*
No. of members	*4*	*10*	*15*	*20*	*11*

Find out the nature of skewness for the above data using quartile coefficient.

Solution.

Age	No. of members	c.f
0 -20	4	4
20-40	10	14
40-60	15	29
60-80	20	49
80-100	11	60

Find Quartiles as in chapter 5.

$$Q_1 = 40 + \left(\frac{15 - 14}{15}\right)20 = 41.33$$
$$\text{Median} = 60 + \left(\frac{30 - 29}{20}\right)20 = 61$$
$$Q_3 = 60 + \left(\frac{45 - 29}{15}\right)20 = 76$$

Bowleys coefficient of skewness,

$$S_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$
$$= \frac{76 + 41.33 - 2 \times 61}{76 - 41.33}$$
$$= -0.1347$$

Since $S_B < 0$, the distribution is negatively skewed.

Illustration 7.12

Find Quartile coefficient of Skewness of the two groups given below and point out which distribution is more skewed.

Scores	Group A	Group B
55-58	12	20
58-61	17	22
61-64	23	25
64-67	18	13
67-70	11	7

Solution.

For Group A

$$Q_1 = 58 + \left(\frac{20.25 - 12}{17}\right)3 = 59.46$$
$$\text{Median} = 61 + \left(\frac{40.5 - 29}{23}\right)3 = 62.5$$
$$Q_3 = 64 + \left(\frac{60.75 - 52}{18}\right)3 = 65.46$$

Quartile coefficient of skewness for group A,

$$\begin{aligned} S_B &= \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} \\ &= \frac{65.46 + 59.46 - 2 \times 62.5}{65.46 - 59.46} \\ &= -0.013 \end{aligned}$$

Similarly for Group B, $Q_1 = 58.24$, Median $= 61.18$ and $Q_3 = 63.79$
Quartile coefficient of skewness for group B,

$$\begin{aligned} S_B &= \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} \\ &= \frac{63.79 + 58.24 - 2 \times 61.18}{63.79 - 58.24} \\ &= -0.0595 \end{aligned}$$

S_B for group B is greater than that of group A in magnitude. So B is more skewed than A.

7.3 Moments

Moments are statistical constants used to describe the various characteristics of a frequency distribution like central tendency, variation, skewness and kurtosis. It is a convenient and unifying method for summarising descriptive statistical measures.

Moments are calculated using the arithmetic mean. The arithmetic mean of the various powers of deviations of observations in any distribution is called the moments of the distribution. If the deviations are taken from arithmetic mean then the moments are termed as central moments and it is denoted by μ (pronounced as 'mu'), a Greek letter. The first four central moments are defined below. Let $x_1, x_2, x_3, \ldots, x_n$ be n observations then

First central moment

$$\mu_1 = \frac{\sum(x-\overline{x})}{n}$$

$$= 0 \text{(Since sum of deviations of items from the mean is always zero}$$

Second central moment

$$\mu_2 = \frac{\sum(x-\overline{x})^2}{n}$$

$$= \text{variance}$$

Third central moment

$$\mu_3 = \frac{\sum(x-\overline{x})^3}{n} \text{ and}$$

Fourth central moment

$$\mu_4 = \frac{\sum(x-\overline{x})^4}{n}$$

Coefficient of Skewness based on Moments.

Coefficient of skewness Based on moments, β_1 (pronounced as beeta one) is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

If $\beta_1 = 0$ the curve is symmetric. The greater the values of β_1 the more skewed the distribution. β_1 does not gives the direction of skewness, since μ_3^2 and μ_2^3 are always positive. Hence Karl Pearson defined γ_1(pronounced as Gamma one) as

$$\begin{aligned}\gamma_1 &= \frac{\mu_3}{\sqrt{\mu_2^3}} \\ &= \frac{\mu_3}{\sigma^3} \\ &= \sqrt{\beta_1}\end{aligned}$$

if $\mu_3 > 0$ then $\gamma_1 > 0$ the distribution is positively skewed. If $\mu_3 < 0$ then $\gamma_1 < 0$ the distribution is negatively skewed. If $\mu_3 = 0$ then $\gamma_1 = 0$ the distribution is symmetric. Thus μ_3 determines the nature of skewness.

Illustration 7.13

The first four central moments of a distribution are 0, 14.75, 39.75 and 142.31. Find coefficient of Skewness and state the nature of Skewness.

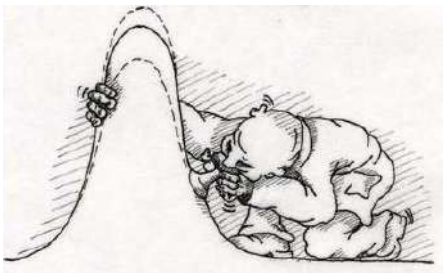
Solution. Given$\mu_2 = 14.75$, $\mu_3 = 39.75$

$$\begin{aligned}\text{Coefficient of Skewness } \beta_1 &= \frac{\mu_3^2}{\mu_2^3} \\ &= \frac{39.75^2}{14.75^3} \\ &= 0.4924\end{aligned}$$

Since $\mu_3 > 0$, the distribution is positively skewed.

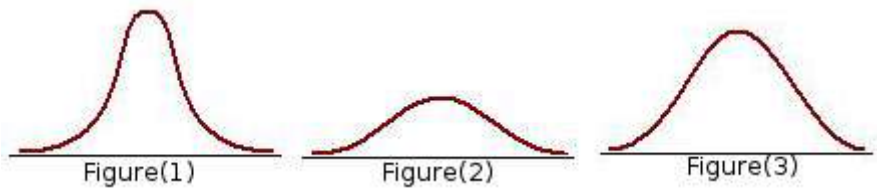
7.4 Kurtosis

Two or more distributions may have identical average, variation and skewness, but they show different degrees of concentration of values of observation around mode and hence they may show different degrees of peakedness of the distributions. Kurtosis is the measure of Peakedness or flatness of the frequency distribution.

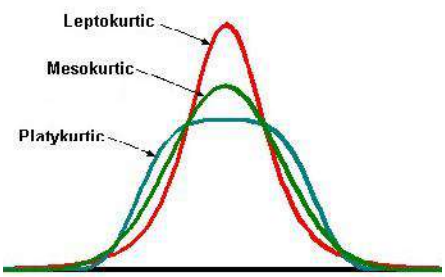


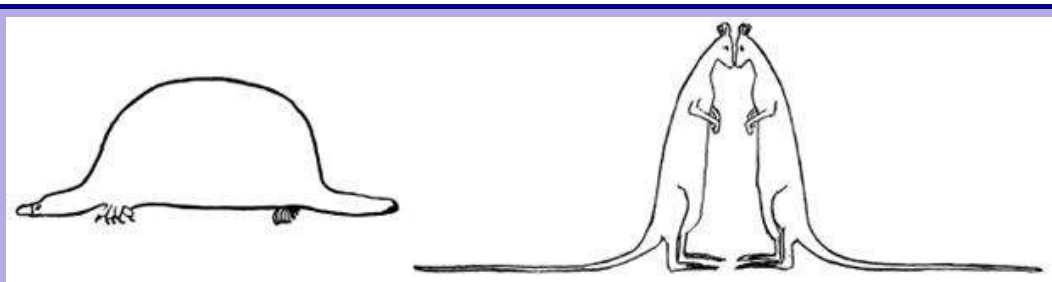
Types of Kurtosis

Look at the following frequency curves



All of them have same centre of location, dispersion and are symmetrical, but they differ in peakedness. In figure 1 the curve is more peaked than others and is called as Lepto Kurtic. In figure 2 the curve is less peaked than others or it is more flat topped, is called as Platty Kurtic. In figure 3 curve is neither more peaked nor more flat topped or it is moderately peaked, is called Meso Kurtic. Meso Kurtic is also known as Natural Curve or Normal Curve.





The word "Kurtosis" is derived from the Greek word meaning "Humped" or "Bulginess". Famous British Statistician William S Gosset (Student) has very humorously pointed out the nature of kurtosis in his research paper "Errors of Routine Analysis" that platy kurtic curves, like the platypus, are squat with short tails; lepto kurtic curves are high with long tails like the Kangaroos noted for lepping. Gossets little sketch is reproduced as shown in the figure.

7.5 Measures of Kurtosis

Kurtosis is measured by coefficient β_2(pronounced as beeta two) is defined by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

Karl Pearson defined one more coefficient of kurtosis as $\gamma_2 = \beta_2 - 3$
If $\beta_2 = 3$ (ie,$\gamma_2 = 0$) the curve is meso kurtic.
If $\beta_2 > 3$ (ie, $\gamma_2 > 0$) the curve is lepto kurtic.
If $\beta_2 < 3$ (ie, $\gamma_2 < 0$) the curve is platy kurtic.

Illustration 7.14

The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the kurtosis of the distribution

Solution. Given $\mu_1 = 0, \mu_2 = 2.5, \mu_3 = 0.7, \mu_4 = 18.75$

$$\text{Coefficient of Kurtosis}, \beta_2 = \frac{\mu_4}{\mu_2^2}$$
$$= \frac{18.75}{2.5^2}$$
$$= 3$$

Since $\beta_2 = 3$ the distribution is meso kurtic.

Illustration 7.15

For the following data calculate coefficient of skewness and coefficient of kurtosis and comment on it 2, 3, 7, 8, 10

Solution.

X	$(x-6)$	$(x-6)^2$	$(x-6)^3$	$(x-6)^4$
2	-4	16	-64	256
3	-3	9	-27	81
7	1	2	1	1
8	4	4	8	16
10	4	16	64	256
30	0	46	-18	610

$$\bar{x} = \frac{30}{5} = 6$$
$$\mu_1 = \frac{\sum(x-\bar{x})}{n} = 0 \text{(Always)}$$
$$\mu_2 = \frac{\sum(x-\bar{x})^2}{n} = \frac{46}{5} = 9.2$$
$$\mu_3 = \frac{\sum(x-\bar{x})^3}{n} = \frac{-18}{5} = -3.6$$
$$\mu_4 = \frac{\sum(x-\bar{x})^4}{n} = \frac{610}{5} = 122$$

$$\text{Coefficient of skewness}, \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-3.6)^2}{9.2^3} = 0.0166$$

Since $\mu_3 < 0$, the distribution is negatively skewed

$$\text{Coefficient of kurtosis}, \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{122}{9.2^2} = 1.44$$

Since $\beta_2 < 3$, the distribution is platy kurtic

Let us sum up

In this chapter the concept of skewness and kurtosis were introduced. Skewness means lack of symmetry where as kurtosis is the measure of peakedness. There are two types of skewness, positive skewness and negative skewness. If a frequency curve has longer tail towards right side of mode, then it is said to be positively skewed. If a frequency curve has longer tail towards left side of mode, it is said to be negatively skewed. If a curve is relatively narrow and peaked at the top, it is called leptokurtic. The curve which is more flat topped is called platy kurtic. The curve which is neither more peaked nor more flat topped is called meso kurtic. Various measures of skewness and kurtosis were also discussed here. Measures of skewness indicate to what extend and in what direction the distribution of variable differs from symmetry of a frequency curve. Measures of kurtosis denote the shape of top of a frequency curve. For a symmetric curve Karl Persons coefficient of skewness , Bowleys coefficient of skewness and moment coefficient of skewness are all equal to zero.

Learning outcomes

After transaction of this unit, the learner:-

- distinguishes symmetric and asymmetric distributions.
- recognises skewness of distributions.
- evaluates and interprets nature of skewness.
- explains kurtosis of distributions.
- evaluates and interprets types of kurtosis.

Evaluation Items

1. For a positively skewed frequency distribution, which of the following is true always ?

   (A) $Q_1 + Q_3 > 2Q_2$ (B) $Q_1 + Q_2 > 2Q_3$ (C) $Q_1 - Q_3 > Q_2$
   (D) $Q_3 - Q_1 > Q_2$

2. For negatively skewed data, which of the following is true?

   (A) Mean = Median = Mode (B) Median < Mean < Mode
   (C) Mean < Median < Mode (D) Mode < Mean < Median

3. For a negatively skewed distribution, more observations clustered

   (A) on the left tail (B) on the right tail (C) in the middle
   (D) Anywhere

4. The limits of Bowleys coefficient of skewness are

   (A) ± 1 (B) ± 2 (C) ± 3 (D) 0 to 1

5. Given $\mu_2 = 7$, $\mu_4 = 98$, then the curve is

   (A) meso kurtic (B) platy kurtic (C) positively skewed
   (D) lepto kurtic

6. For a symmetric distribution

   (A) $\beta_1 = 0$ (B) $\beta_1 < 0$ (C) $\beta_1 > 0$ (D) $\beta_1 \neq 0$

7. For a set of data, third central moment is -1.6, then the coefficient of skewness is

   (A) positively skewed (B) negatively skewed (C) symmetric
   (D) Can't detertmined

8. In an economic study, a sample of persons earning up to Rs. 30,000 per month were considered. It is found that 30% are earning less that 5000 per month, 95% are earning less than Rs. 15,000 per month and 98% less than 24,000 per month. Then the frequency curve for the data will be

   (A) Symmetric (B) Positively skewed (C) Negatively skewed
   (D) Nothing can be inferred

9. Which of the following is not correct statement

   (A) For a symmetric distribution Mean = Median = Mode
   (B) If Median $= 24$ and Mean $= 26$ then the skewness is positive
   (C) $\beta_1 = 0$ for a positively skewed data
   (D) If $\beta_2 = 3$, the distribution is meso kurtic

10. Old age distribution is an example for _____ Skewed distribution

11. If $\beta_2 > 3$ then the curve is _____

12. Skewness is _____ if $(Q_3 - Q_2) < (Q_2 - Q_1)$

13. If Karl Pearsons coefficient of Skewness is 0.40, standard deviation is 8 and mean is 30. Find the mode of the distribution.
    Ans. Mode $= 26.8$

14. For a group of 10 times $\sum x = 452$, $\sum x^2 = 24270$ and mode $= 43.7$. Find the Pearsonian coefficient of Skewness.
    Ans. Mean $= 45.2$,SD $= 19.6$ and $S_K = 0.08$)

15. The income distribution of two villages revealed the following information.

	Mean	Mode	Standard deviation
Village I	500	475	10
Village II	600	590	5

    What is the nature of Skewness of the two distributions ? Which distribution is more skewed?

    Ans. Positively Skewed, Village 1 is more skewed

16. Tests were conducted for 3 subjects namely Economics, English and Statistics for 37 students in a class. The marks obtained by them are tabulated as follows

Marks	No.of students		
	Economics	English	Statistics
12	2	2	1
14	5	12	3
16	7	8	5
18	9	6	6
20	7	5	8
22	5	3	12
24	2	1	2
Mean	18	16.6	18.96
Median	18	16	20
Mode	18	14	22

(A) Identify the Positively skewed, Negatively Skewed and Symmetric distributions

(B) Indicate the position of Mean, Median and Mode in three cases by drawing frequency curves of the distribution.

Ans. Eco-Symmetric, English - Positive, Statistics- Negative

17. For a frequency distribution the Mean=100, coefficient of skewness is 0.2 and Pearsons coefficient of variance is 35. Find mode of the distribution.

Ans. (SD =35, Mode=93)

18. In a distribution of wages of workers in a factory, the difference of upper and lower quartiles is 15, their sum is 35 and the Median is 20. Find the coefficient of Skewness.

Ans. $(S_B = 0.33)$

19. Suppose the distribution of scores of some students is symmetric. If the values of Q1and Q3 are 20 and 40 respectively, what is the Median mark? If the Median mark is 35, what would be the skewness of the distribution ?

Ans. Median $= 30$, $S_k = -0.5$

20. The following data represent the monthly salary (in thousands)of seven Assistant Professors in the department Statistics of Presidency college 26, 30, 32, 26, 29, 28, 60. Find (a) Mean (b) Mode (c) Coefficient of Skewness.

    Ans. Mean $= 32.57$,Mode $= 26$, $SD = 11.2$,$S_k = 0.59$

21. The following figures represent the weights (in Kg) of 10 new born babies in a hospital on a particular day 2, 3, 3, 3, 4, 2, 2.5, 3.5, 3.7, 3

    (A) Compute Mean, Mode and Standard deviation

    (B) Are these data Skewed? Give reason

    Ans. Mean $= 2.97$, Mode $= 3$, $SD = 0.63$, $S_k = -0.048$

22. The following table gives the height (in inches) of 100 students in a higher secondary school.

Class interval	No. of Students
60 and up to 62	5
62 and up to 64	18
64 and up to 66	42
66 and up to 68	20
68 and up to 70	8
70 and up to 72	7

    Calculate (a) Mean (b)Mode (c) Coefficient of Skewness

    Ans. Mean $= 65.58$, Mode $= 65.04$, $SD = 2.41$, $S_k = 0.23$

23. The president of company, which employs 50 persons, wants to study the pattern of absenteeism of all employees. The distribution of the number of days these employees were absent is given as follows:

No. of days absent	No. of Employees
0 to 2	15
3 to 5	20
6 to 8	8
9 to 11	5
12 to 14	2

Calculate (a) Mean (b) Standard deviation (c) Karl Pearsons Coefficient of Skewness. **Ans.** Mean = 4.54, Mode= 3.59, SD= 3,27, $S_K = 0.29$

24. Calculate Karl Pearsons coefficient of skewness for the following data

Life in Hrs.	:	80-160	160-240	240-320	320-400	400-480
No. of Tubes	:	24	90	45	12	30
Life in Hrs.	:	480-560	560-640	640-720		
No. of Tubes	:	120	39	30		

Ans. Mean=403.1, Mode=522.1,$SD = 174.2$, $S_K = -0.68$

25. Compute Karl Pearsons coefficient of skewness from the following data

Marks	:	Above 0	Above 10	Above 20	Above 30	Above 40
No. of Students	:	140	130	115	95	80
Marks	:	Above 50	Above 60	Above 70	Above 80	
No. of Students	:	70	30	14	0	

Ans. Mean=43.14,Mode=55.56,$SD = 20.96$, $S_k = -0.59$

26. Calculate coefficient of skewness using Quartiles

Mid value	:	15	20	25	30	35	40
Frequency	:	30	28	25	24	10	21

Ans. $Q_1 = 18.3$, $Q_2 = 24.7$, $Q_3 = 31.8$, $S_B = 0.052$

27. Calculate Bowleys coefficient of skewness for the following data

Life(in month) :	< 87.5	< 112.5	< 137.5	< 162.5
No. of bulbs :	35	75	123	223
Life(in month) :	< 187.5	< 212.5	< 237.5	< 262.5
No. of bulbs :	348	428	478	500

Ans. $Q_1 = 138$, $Q_2 = 167.9$, $Q_3 = 195.94$, $SB = -0.03$

28. Find Karl Pearsons coefficient of skewness for the two series and point out which one is more skewed

Age	No. of children	
(in Yrs)	School A	School B
6	3	1
8	9	10
9	15	9
10	8	7
11	5	3
Total	40	30

Ans. For A, Mean $= 9$, Mode $= 9$, $SD = 1.26$, $S_k = 0$; For B, Mean $= 9$, Mode $= 8$, $SD = 1.13$,$S_k = 0.88$. B is more Skewed

29. Calculate Bowleys coefficient of skewness for the marks of the students if

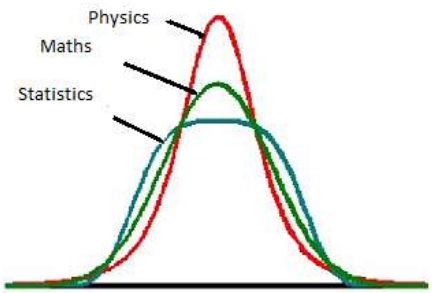
    3 students get 3 marks each
    5 students get 5 marks each
    8 students get 7 marks each
    6 students get 8 marks each
    2 students get 10 marks each.

Ans. $Q_1 = 5$, $Q_2 = 7$, $Q_3 = 8$, $SB = -0.33$

30. If 2^{nd} and 4^{th} central moments of a frequency distribution are 2 and 16 respectively. What would be the nature of kurtosis of the distribution ?

Ans. $\beta_2 = 4$, lepto kurtic

31. The frequency curves of marks obtained in three subjects are given



Comment on the type of kurtosis.

Ans. Physics is lepto, Maths is meso and Statistics is platy kurtic

32. The first four central moments of a distribution are 0,9.56,-3.29 and 215.72. Compute the measures of skewness and kurtosis. State the nature of the curve.

Ans. $\beta_1 = 0.012$, skewness is negative,$\beta_2 = 2.36$ platy kurtic

33. The 1st four moments about mean of a distribution are 0, 16,-64 and 162. Calculate moment coefficients of skewness and kurtosis.

Ans. $\beta_1 = 1$, skewness is negative ,$\beta_2 = 0.633$ platy kurtic

34. For two distributions the second central moments are 9 and 16 and their third central moments are -8.1 and -12 respectively. Which of the two distributions is more skewed to the left ?

Ans. For I ,$\beta_1 = 0.9$, $\gamma_1 = -0.3$; For II $\beta_1 = 0.04$, $\gamma_1 = -0.2$; First one is More Skewed.

35. The first Four central moments of a distribution are 0, 9.2, -3.6 and 122. Calculate coefficient of kurtosis of the distribution.

Ans. $\beta_2 = 1.44$, platy kurtic

36. The values of β_1 and γ_1 for three distributions are given below

Identify the meso kurtic distribution.

Ans. Distribution II is meso kurtic

37. For a negatively skewed distribution, the values of β_1 and μ_2 are 9 and 4 respectively. Find the values of γ_1 and μ_3.

    Ans. $\gamma_1 = -3$, $\mu_3 = -24$ since skewness is negative

38. First two central moments of a meso kurtic distribution are 0 and 2.5 respectively. Find fourth central moment.

    Ans. $\mu_4 = 18.75$

39. Compute first four central moments for first four even numbers. Also calculate coefficients of skewness and kurtosis

    Ans. $\mu_2 = 5$, $\mu_3 = 0$, $\mu_4 = 41$, $\beta_1 = 0$, $\beta_2 = 1.64$

40. Determine the coefficient of kurtosis and comment on the nature of data 3, 6, 8, 10,18

    Ans. $\mu_2 = 25.6$, $\mu_3 = 97.2$, $\mu_4 = 1588$, $\beta_2 = 2.42$, platy kurtic

Answers:

1) a 2) c 3) a 4) a 5) b 6) a 7) b 8) c 9) c
10) positively skwed 11) lepto kurtic 12) negative

Introduction

Uncertainty is a part of our everyday life. We are familiar with questions like, "Is there any chance for a rain tomorrow?", "Is there any chance for meeting someone?", "What are the chances of getting a job?", "What are the chances of getting admission to a particular course?" etc. People answer such questions with percentages, fractions or statements like 'fifty-fifty'. But for scientific purposes it is necessary to give the word chance a clear interpretation. This turnsout to be hard and mathematicians have struggled for centuries with this job.

J. Bernoulli (1654-1705)

De Moivre (1667-1754)

History: Probabiliaty theory originated in the 16th century when an Italian physician and mathematician J.Cardan wrote the first book on the subject, 'The Book on Games of Chance'. Since its inception, the study of probability has attracted the attention of great mathematicians. James Bernoulli (1654-1705), A. de Moivre (1667-1754), and Pierre Simon Laplace (1749-1827) are among those who made significant contributions to this field. Laplace's 'Theorie Analytique des Probabilites' (1812) is considered to be the greatest contribution by a single person to the theory of probability. Later, A.N.Kolmogorov (1903-1987), a Russian mathematician laid down some axioms to interpret probability, in his book 'Foundation to Probability' published in 1933. Kolmogorov treated probability as a function of outcomes of the experiment. In recent years probability has been used extensively in many areas such as Biology, Medicine, Economics, Genetics, Physics, Sociology, Insurance etc.

Probability

The word chance indicates that there is an element of uncertainty. Sometimes it is impossible to say what will happen the next moment. But certain events are more likely to occur than others, and that is where probability theory comes into play. Probability helps in the prediction of future by assessing how likely outcomes are and knowing of what could happen in the future will help us make informed decision. The uncertainty and risk are measured numerically in probability theory. Thus probability is the way of measuring the chances of something to happen. We can use it to indicate how likely or unlikely an occurrence is. Probability is the base of inferential statistics. There are different approaches to probability. The basic concepts of probability are discussed in this chapter. Probability is measured on a scale of 0 to 1. If an event is impossible, it has the probability of zero. If it is an absolute certainty, then the probability is 1. In most of the cases we will be dealing with the probability somewhere in between.

Look at a coin tossing experiment. When a coin is tossed, there are two possible outcomes:

- head (H) or
- tail (T)

We say that the probability of the coin landing H is $\frac{1}{2}$. And the probability of the coin landing T is $\frac{1}{2}$.

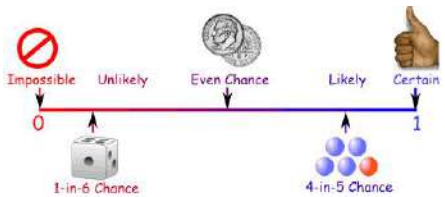
See another experiment:

When a single die is thrown, there are six possible outcomes: 1, 2, 3, 4, 5, 6. The probability of any one of them occuring is $\frac{1}{6}$.

See one more example: There are 5 balls in a bag, out of which 4 are blue, and 1 is red. What is the probability that a blue ball is selected?

Number of ways it can happen: 4 (There are 4 blue balls.)
Total number of outcomes: 5 (There are 5 balls in total.)
So, the probability is $\frac{4}{5} = 0.8$

So, probability of something happening is somewhere in between 0 and 1



Basic properties of probabilities

- The probability is always between 0 and 1.
- The probability of occurrence of an impossible event is 0.
- The probability of something certain to occur is 1.
- Probability cannot be negative.

To understand probability on the basis of proper reasoning and logic we have to be clear about the terms which are in common use but having specific meaning when we talk about random phenomena.

8.1 Random experiment

An experiment consists of a number of trials. Drawing ten cards from a pack of cards, flipping a coin two times, etc are examples for random experiment.

An experiment is called random experiment if it satisfies the following conditions:

1. It has more than one outcome.
2. It is not possible to predict the outcome in advance.
3. It can be repeated any number of times under identical conditions.

In this chapter, we shall refer the random experiment by experiment only unless stated otherwise.

Trial A trial is an action which results in one of several possible outcomes. Flipping a coin, rolling a die, drawing a card from a pack of cards etc. are trials.

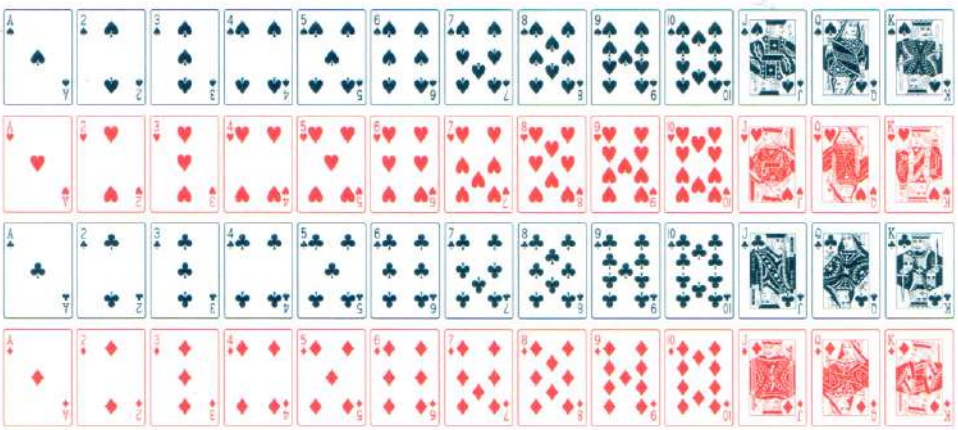
Outcome A possible result of the single trial of a random experiment is called its outcome.

Sample space The set of all possible outcomes of a random experiment is called the sample space. Sample space is usually listed in curly brackets; { }.

Sample point Each element in the sample space is called a sample point.

Let us consider some examples:

Experiment	Sample Space
Tossing a coin	 { Head, Tail} or {H,T}
Roll a die	 { 1,2,3,4,5,6}
Answer a true or false question	{ True, False}
Tossing two coins	{ HH,TT,HT,TH}



Sample space for choosing a card from a deck

Activity

Write the sample space for tossing two dice.

8.2 Events

When we say 'Event' we mean one or more outcomes. The following are examples of events:

- Getting a tail in a coin toss
- Rolling a '5' on a roll of die

An event can include several outcomes:

- Choosing a 'King' from a deck of cards (any of the 4 Kings)
- Rolling an even number (2, 4 or 6) on a die

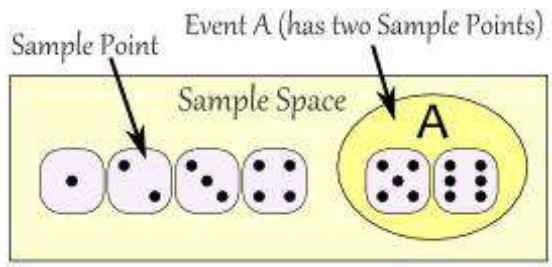
Therefore, **An event is a set of outcomes which have some characteristics in common.**

Getting a number less than 5 in a die tossing experiment, getting black or face card when 4 cards are drawn from a pack of cards etc. are also events.

An event with one outcome is called a **simple event**. An event which consists of more than one outcome is called a **compound event**. That is, a compound event consists of two or more simple events.

Basic properties of probabilities

- The **Sample Space** is the set of all possible outcomes.
- A **Sample Point** is just one possible outcome.
- An **Event** can be one or more of the possible outcomes.



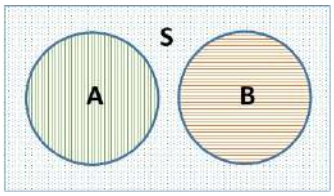
Probability is the likelihood for the occurrence of an event.
Event is an outcome or occurrence with a probability assigned to it.

Equally likely events

Two or more events that have the same probability of occurrence are called equally likely events. For example, In a coin tossing experiment the events Head , Tail have equal chance to occur. Hence they are equally likely events.

Mutually exclusive events

Two events are mutually exclusive if they cannot occur simultaneously. All simple events (elementary outcomes) are mutually exclusive. For example, if a coin is tossed, either the head or tail can occur, but not both. Similarly, a newborn can either be a boy or a girl. Therefore in a trial, the occurrence of head excludes the occurrence of tail. The occurrence of boy excludes the occurrence of girl and vice-versa.

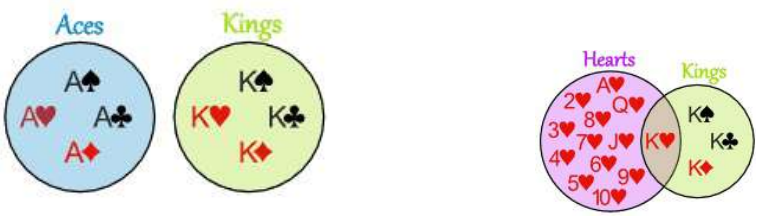


The following pairs of events are mutually exclusive:

- Turning left and turning right (You cannot do both at the same time.)
- Occurrence of head and occurrence of tail, in a coin toss
- A card drawn from a deck of cards is King, and that it is an Ace

What events are not mutually exclusive?

The event that a card drawn from a deck of cards is a King, and that is a Heart, are not mutually exclusive, since you can have a King of Hearts.



Aces and Kings are mutually exclusive but Hearts and Kings are not.

Mutually Exclusive means you can't get both events at the same time. It is either one or the other, but not both.

Exhaustive events

A set of events are said to be exhaustive if they altogether constitute the sample space. For example, consider the experiment of rolling a die, sample space S = { 1,2,3,4,5,6 }. Let the events A ={ 1,3,5 } and B ={2,4,6 }. The events A and B are exhaustive since they togther constitute the sample space.

8.3 Classical Definition of Probability

Classical definition uses sample spaces to determine the numerical probability that one event will happen. One does not have to perform the experiment to determine the probability. Classical probability assumes that all outcomes in the sample space are equally likely and mutually exclusive.

Pierre Simon Laplace (1749-1827)

Formula for classical probability

The probability of any event A is

$$\frac{\text{Number of outcomes in } A}{\text{Number of outcomes in } S}$$

$P(A) = \frac{N(A)}{N(S)}$. This probability is called classical probability.

Illustration 8.1

A spinner has 4 equal sectors coloured yellow, blue, green and red. What are the chances of landing on blue after spinning the spinner? What are the chances of landing on red?

Solution:

$$P(Yellow) = \frac{Number\ of\ Yellow\ Sectors}{Total\ Number\ of\ Secters} = \frac{1}{4}$$

$$P(Blue) = \frac{Number\ of\ Blue\ Sectors}{Total\ Number\ of\ Secters} = \frac{1}{4}$$

$$P(Green) = \frac{Number\ of\ Green\ Sectors}{Total\ Number\ of\ Secters} = \frac{1}{4}$$

$$P(Red) = \frac{Number\ of\ Red\ Sectors}{Total\ Number\ of\ Secters} = \frac{1}{4}$$

Illustration 8.2

A single 6-sided die is rolled. What is the probability of each outcome? What is the probability of rolling an even number? What is that of rolling an odd number?

The possible outcomes of this experiment are 1, 2, 3, 4, 5 and 6.

Solution:

$$P(1) = \frac{\textit{number of ways to roll a 1}}{\textit{total number of sides}} = \frac{1}{6}$$

$$P(2) = \frac{\textit{number of ways to roll a 2}}{\textit{total number of sides}} = \frac{1}{6}$$

$$P(3) = \frac{\textit{number of ways to roll a 3}}{\textit{total number of sides}} = \frac{1}{6}$$

$$P(4) = \frac{\textit{number of ways to roll a 4}}{\textit{total number of sides}} = \frac{1}{6}$$

$$P(5) = \frac{\textit{number of ways to roll a 5}}{\textit{total number of sides}} = \frac{1}{6}$$

$$P(6) = \frac{\textit{number of ways to roll a 6}}{\textit{total number of sides}} = \frac{1}{6}$$

$$P(\textit{even}) = \frac{\textit{number of ways to roll an even number}}{\textit{total number of sides}} = \frac{3}{6}$$

$$P(\textit{odd}) = \frac{\textit{number of ways to roll an odd number}}{\textit{total number of sides}} = \frac{3}{6}$$

Illustration 8.2 describes the difference between an outcome and an event.

A single outcome of this experiment is rolling a 1, or rolling a 2, or rolling a 3, etc. Rolling an even number (2, 4 or 6) is an event, and rolling an odd number (1, 3 or 5) is also an event.

In illustration 8.1, the probability of each outcome is always the same. The probability of landing on each colour of the spinner is always one-fourth.

In illustration 8.2, the probability of rolling each number on the die is always one-sixth. In both of these experiments, the outcomes are equally likely to occur.

Let us look at an experiment in which the events are not equally likely.

Illustration 8.3

A glass jar contains 6 red, 5 green, 8 blue and 3 yellow balls. If a single ball is chosen at random from the jar, what is the probability of choosing:

- *a red ball?*
- *a green ball?*
- *a blue ball?*
- *a yellow ball?*

Solution:

The possible outcomes of this experiment are red, green, blue and yellow.

$$P(red) = \frac{number\ of\ ways\ to\ choose\ red}{total\ number\ of\ balls} = \frac{6}{22} = \frac{3}{11}$$

$$P(green) = \frac{number\ of\ ways\ to\ choose\ green}{total\ number\ of\ balls} = \frac{5}{22}$$

$$P(blue) = \frac{number\ of\ ways\ to\ choose\ blue}{total\ number\ of\ balls} = \frac{8}{22} = \frac{4}{11}$$

$$P(yellow) = \frac{number\ of\ ways\ to\ choose\ yellow}{total\ number\ of\ balls} = \frac{3}{22}$$

The events in this experiment are not equally likely to occur. You are more likely to choose a blue ball than any other colour. You are least likely to choose a yellow ball.

Illustration 8.4

Choose a number at random from 1 to 5. What is the probability of each outcome? What is the probability that the number chosen is even? What is the probability that the number chosen is odd?

The possible outcomes of this experiment are 1, 2, 3, 4 and 5.

Solution:

$$P(1) = \frac{number\ of\ ways\ to\ choose\ 1}{total\ number\ of\ numbers} = \frac{1}{5}$$

$$P(2) = \frac{number\ of\ ways\ to\ choose\ 2}{total\ number\ of\ numbers} = \frac{1}{5}$$

$$P(3) = \frac{\text{number of ways to choose 3}}{\text{total number of numbers}} = \frac{1}{5}$$

$$P(4) = \frac{\text{number of ways to choose 4}}{\text{total number of numbers}} = \frac{1}{5}$$

$$P(5) = \frac{\text{number of ways to choose 5}}{\text{total number of numbers}} = \frac{1}{5}$$

$$P(\text{even}) = \frac{\text{number of ways to choose an even number}}{\text{total number of numbers}} = \frac{2}{5}$$

$$P(\text{odd}) = \frac{\text{number of ways to choose an odd number}}{\text{total number of numbers}} = \frac{3}{5}$$

The outcomes 1, 2, 3, 4 and 5 are equally likely to occur as a result of this experiment. However, the events even and odd are not equally likely to occur, since there are 3 odd numbers and only 2 even numbers from 1 to 5.

Know your progress

1. Three coins are tossed. Write the sample space? What is the probability of getting three heads?
2. In a school there are 100 science students 100 commerce students and 150 humanities students. If a leader is selected at random, what is the probability that he is a humanities student?
3. A die is marked with 1 on two faces, 5 on two faces and 6 on the remaining two faces. If it is rolled find the probability of getting an even number?
4. A number is selected from the natural numbers less than 50. What is the probability of
   (A) Getting an even number?
   (B) Getting a multiple of 10?
   (C) Getting a perfect square?

Activity

Cite 5 examples for equally likely events.

Let us summarise what we have learnt so far:

- The probability of an event is the measure of the chance that the event may occur as a result of an experiment.
- The probability of an event A is the number of ways an event A can occur divided by the total number of possible outcomes.
- The probability of an event A, symbolised by P(A), is a number between 0 and 1, inclusive, that measures the likelihood of an event in the following way:
  If P(A) >P(B) then event A is more likely to occur than event B.
  If P(A) = P(B) then events A and B are equally likely to occur.

Counting rules for determining the number of outcomes

To assign probabilities to experimental outcomes it is first necessary to identify and count them. The following are some important rules for counting the experimental outcomes.

Counting rule for multistep experiments

The counting rule for multistep experiments helps us to determine the number of experimental outcomes without counting them. The rule is stated as:

"If an experiment is performed in k stages with n_1 ways to accomplish the first stage, n_2 ways to accomplish the second stage, n_3 ways to accomplish the third stage... and n_k ways to accomplish the k^{th} stage, then the number of ways to accomplish the experiment is $n_1 \times n_2 \times n_3 \times \cdots \times n_k$ ways."

Suppose a person can take three routes from a city A to city B, four routes from city B to city C and three from city C to city D, then the number of possible routes from city A to city D, while he travel from A to B, B to C and C to D is,

(A to B) × (B to C) × (C to D) = 3 × 4 × 3 = 36 ways.

Tossing of two coins can be considered as two step experiment in which each coin can land in two ways: head (H) and tail (T). Hence we can have $2 \times 2 = 4$ outcomes.

Permutation and Combination

The word 'permutation' means 'arrangement' and 'combination' means 'group'. The word permutation refers to different arrangements of objects in a set. The number of observations which can be made from a group of things, irrespective of the order, it is called combination. Each combination of objects can be arranged in a number of ways. Precisely,

- If the order does matter it is a permutation.
- If the order does not matter it is a combination.

There are basically two types of permutation:

1. Permutations when repetition allowed.
2. Permutation without repetition.

Permutation with repetition

When we have n things to choose, we have n choices each time. When we choose r things out of n, the permutations are $n \times n \times n \times \cdots \times n(r \text{ times}) = n^r$

Consider a number lock as shown below:

There are 10 numbers to choose from $(0, 1, 2, \ldots, 9)$ and we choose 3 of them. We can have $10^3 = 1000$ permutations.

Permutations without repetition

In this case the number of available choices reduces each time. If we choose r things out of n, the available permutations are $n \times (n-1) \times (n-2) \times \cdots \times (n-r+1)$.

Factorial function

The factorial function (symbol: !) just means to multiply a series of descending natural numbers.

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

Examples:

- $4! = 4 \times 3 \times 2 \times 1 = 24$
- $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5040$
- $1! = 1$

Note: it is generally agreed that 0! = 1. It may seem funny that multiplying no numbers together gets you 1, but it helps to simplify a lot of equations.

Number of permutations of n different items taken r at a time is usually denoted by ${}^nP_r$ or $P(n,r)$ and the formula reduces to

$$P(n,r) = {}^nP_r = {}_nP_r = \frac{n!}{(n-r)!}$$

In particular, if $r = n$, we can have, ${}^nP_n = \frac{n!}{(n-n)!} = n!$ permutations

Permutations are designed to find out the total number of ways in which the event can be performed.

Combinations

A combination of n different things taken r at a time is denoted by ${}^nC_r$, or $C(n,r)$ or $\binom{n}{r}$.

When we make permutations of size r from n things, it may be noted that $r!$ Permutations are of the same combination. Therefore, ${}^nC_r$ becomes

$${}^{n}C_{r} = \frac{n \times (n-1)(n-2)\cdots(n-r+1)}{n \times (n-1) \times \ldots \times 2 \times 1} = \frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r!(n-r)!}$$

The formula reduces to,

$${}^{n}C_{r} = {}_{n}C_{r} = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

From a list of 10 questions 4 questions are to be answered. How many different selections can be made?

Number of selections= ${}^{10}C_{4} = \frac{10 \times 9 \times 8 \times 7}{1 \times 2 \times 3 \times 4} = 210$

It can be seen that the formula is symmetrical, that is, ${}^{n}C_{r} = {}^{n}C_{n-r}$

$$\frac{n!}{r!(n-r)!} = \binom{n}{r} = \binom{n}{n-r}$$

Consider ${}^{16}C_{3}$.

$$\binom{16}{3} = \binom{16}{16-3} = \binom{16}{13}$$

$$\frac{16!}{3!(16-3)!} = \frac{16!}{13!(16-13)!} = \frac{16!}{3!13!}$$

Combinations of n different objects taken any number of them at a time (the number of selections of some or all objects from n different objects) is

$${}^{n}C_{1} + {}^{n}C_{2} + {}^{n}C_{3} + \ldots + {}^{n}C_{n} = 2^{n} - 1$$

Combinations with restrictions

In certain cases we may have to include or exclude some particular items in when we choose a group of r out of n. When we have to include k particular items always, we will have ${}^{n-k}C_{r-k}$ combinations.

When we have to exclude k particular items always, we will have ${}^{n-k}C_{r}$ combinations.

Illustration 8.5

A committee of two persons is selected from two men and two women, what is the probability that the committee will have (a) no man? (b) one man? (c) two men?

Solution:

The total number of persons is $2+2=4$. *Out of these 4 persons, two can be selected in* ${}^4C_2 = \frac{4\times 3}{1\times 2} = 6$ *ways.*

(a) No men in the committee means there will be two women in the committee. Out of two women two can be selected in ${}^2C_2 = 1$ *way.* $({}^nC_n = 1)$

Therefore, P(no men) = $\frac{1}{6}$

(b) One man in the committee means that there will be one man and one woman. One man out of two can be selected in ${}^2C_1 = 2$ $({}^nC_1 = n)$ *ways and one woman out of two can be selected in* ${}^2C_1 = 2$ *ways. Together they can be selected in* ${}^2C_1 \times {}^2C_1 = 4$ *ways.*

Therefore, P(one man) = $\frac{4}{6} = \frac{2}{3}$

(c) Two men can be selected in ${}^2C_2 = 1$ *way.*

Hence P(two men)= $\frac{1}{6}$

Illustration 8.6

A bag contains 7 red and 9 blue balls. 3 balls are drawn together. What is the probability that (a) all are blue? (b) all are red? (c) one is red and two are blue? (d) two is red and one is blue?

Solution:

There are 16 balls in total.

(a) 3 balls from 16 can be drawn in ${}^{16}C_3 = \frac{16\times 15\times 14}{1\times 2\times 3} = 560$ *ways.*

3 blue balls can be drawn in ${}^9C_3 = \frac{9\times 8\times 7}{1\times 2\times 3} = 84$ *ways.*

P(all are blue) = $\frac{84}{560} = \frac{3}{20}$

(b) *3 red balls can be drawn in* $\frac{7\times6\times5}{1\times2\times3}=35$ *ways.*

P(all are red) $=\frac{35}{560}=\frac{1}{16}$

(c) *1 red ball and 2 blue balls can be drawn in* ${}^{7}C_{1}\times{}^{9}C_{2}=\frac{7}{1}\times\frac{9\times8}{1\times2}=252$ *ways.*

P(1 red and 2 blue) $=\frac{252}{560}=\frac{9}{20}$

(d) *2 red balls and 1 blue ball can be drawn in* ${}^{7}C_{2}\times{}^{9}C_{1}=\frac{7\times6}{1\times2}\times9=189$ *ways.*

P(1 red and 2 blue) $=\frac{189}{560}=\frac{27}{80}$

Know your progress

:

1. There are 20 boys and 10 girls in a class. How many different teams of 5 can be selected? If a team of 5 is selected at random, what is the probability that there are 3 boys and 2 girls in the team?
2. The workers of a factory are allowed to take two weekly off at their choice. Find the probability that they select Monday and Tuesday?
3. If you try to unlock a number lock of three rings with numbers from 0 to 9, what is the probability that you succeed? (Hint: It is a number, order is important.)

Algebra of events

The event 'not A' (complement of A)

The concept of complementary events is very important in probability theory. For instance when a die is rolled, the sample space consists of 1,2,3,4,5 and 6. The event A of getting odd numbers consists of the outcomes 1,3 and 5. The event of not getting an odd number is called the complement of A and in this case it consists of the outcomes 2,4 and 6. See the complements of the events listed below:

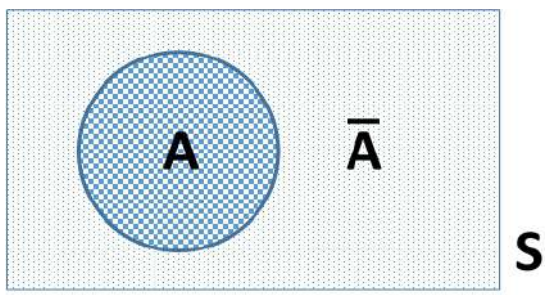
1. Rolling a die and getting 6.
2. Selecting a letter from English alphabet and getting a consonant.
3. Selecting a month and getting a month with 31 days.
4. Selecting a day of a week and getting Sunday.

Complements of the events given are:

1. Getting 1,2,3,4 or 5.
2. Getting a vowel.
3. Getting February, April, June, September or November.
4. Getting Monday, Tuesday, Wednesday, Thursday, Friday or Saturday.

The event 'not A' or complement of A is the set of outcomes in the sample space that are not included in the event. The complement of A is denoted by $\bar{A}$ (read A bar) or A'.

Venn diagram for complementary events



Rule for finding probability of complement of an event

$$P(A) + P(A') = 1$$

$$P(A) = 1 - P(A') \quad \text{or} \quad P(A') = 1 - P(A)$$

Therefore, if the probability of an event or its complement is known, then the other can be found by subtracting the same from 1.

Illustration 8.7

Two players Sangeetha and Haritha play a tennis match. It is known that the probability of Sangeetha winning the match is 0.6. What is the probability of Haritha winning the match?

Solution:

Let H and S be the events Haritha winning the match and Sangeetha winning the match respectively.

Probability of Sangeetha winning = 0.6 (given)

Probability of Haritha winning

$= 1 - P(S)$ [*as the events H and S are complementary*]

$= 1 - 0.6$

$= 0.4$

Illustration 8.8

Baby tosses two different coins simultaneously. What is the probability that he gets at least one head? **Solution:**

The possible outcomes are $(H,H), (T,T), (H,T)$ *and* (T,H). *There are 4 outcomes.*

$$\text{Probability of no head} = \frac{1}{4}$$

$$\text{Probability of at least one head} = 1 - \text{probability of no head}$$

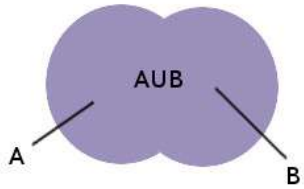
$$= 1 - \frac{1}{4} = \frac{3}{4}$$

It may be noted that in three outcomes, there is at least one head, therefore probability of at least one head is $\frac{3}{4}$

Event 'A or B'

With any two events A and B we can define a new event 'A or B' defined by the condition that "either event A or event B or both occur" or equivalently "at least one of the events A or B occurs". This event is denoted by (A or B) or $A \cup B$ (read A union B), and consists all outcomes, either in A or in B or in both.

Venn diagram for the event (A or B):

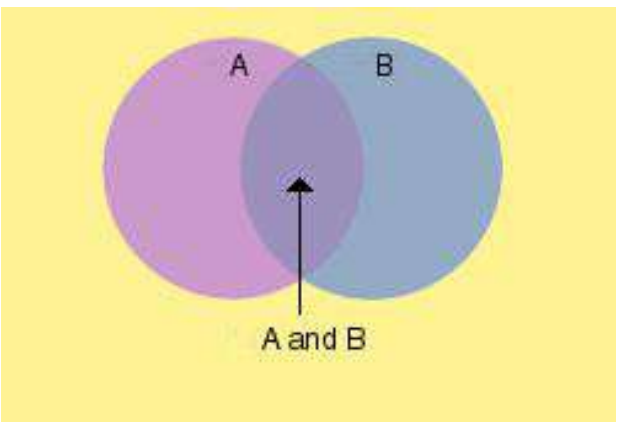


The shaded portion represents A ∪ B.

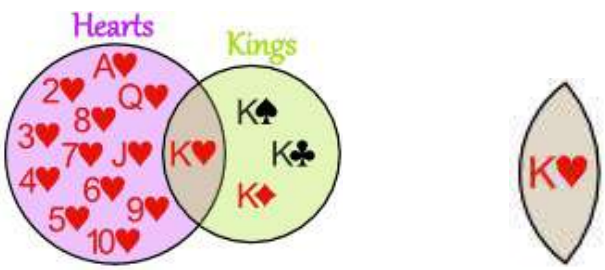
Event 'A and B'

Event 'A and B' defined by the condition that "both A and B occur" is another new event associated with the events. This event is denoted by 'A and B', A&B or A ∩ B (read A intersection B). Event 'A and B' consists of all outcomes common to both event A and event B.

Venn diagram for the event (A and B):



The shaded portion represents A ∩ B.



Hearts and Kings together is only the King of Hearts.

Algebra of events

1. A' → not A
2. (A or B) → atleast one among the events
3. (A and B) → all events (Simultaneous occurrence of A and B)
4. (A' and B') → neither A nor B
5. (A and B') → only A
6. (A and B') or (A' and B) → exactly one among A and B

8.4 Addition Rules for Probability

Many problems involve in finding probability of two or more events. For example, think of a large gathering of university students. One might wish to know, for a student selected at random, the probability that

1. The student is a commerce graduate.
2. The student is a female.
3. The student is a female and is also a commerce graduate.

Consider another example, at the same gathering there are Commerce graduates and Mathematics graduates. What is the probability that the student selected is a commerce graduate? In this case there are two possibilities:

1. The student is a Commerce graduate.
2. The student is a Mathematics graduate.

The difference between the two examples is that in the first case, the student selected can be female and commerce graduate at the same time. In the second case, the student selected cannot be both Commerce graduate and

Mathematics graduate. In the second case the events are mutually exclusive. In the first case the events are not mutually exclusive.

The probability of two or more events can be determined by addition rules. Addition rule for mutually exclusive events and for events those are not mutually exclusive are not the same. For mutually exclusive events A and B, the probability of (A or B) is obtained by adding the probability of A and probability of B.

Addition Rule 1

When two events A and B are mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B)$$

When the events A and B are not mutually exclusive, to obtain the probability of (A or B) we add the probability of A and probability of B and then subtract the probability of (A and B) from the sum. We subtract the probability of (A and B) to avoid the double counting of (A and B).

Addition Rule 2

When two events A and B are not mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

Illustration 8.9

A single 6-sided die is rolled. What is the probability of rolling a 2 or a 5?

Solution:

1. *The number rolled can be 2.*

2. *The number rolled can be 5.*

These events are mutually exclusive since they cannot occur at the same time. If two events A and B are mutually exclusive, we have

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(2) = \frac{1}{6}, \qquad P(5) = \frac{1}{6}$$

$$P(2 \text{ or } 5) = P(2) + P(5)$$

$$= \frac{1}{6} + \frac{1}{6}$$

$$= \frac{1}{3}$$

Illustration 8.10

A spinner has 4 equal sectors colored yellow, blue, green and red. Find the probability of landing on red or blue after spinning this spinner?

Solution:

$$P(red) = \frac{1}{4}, \qquad P(blue) = \frac{1}{4}$$

$$P(red \text{ or } blue) = P(red) + P(blue)$$

$$= \frac{1}{4} + \frac{1}{4}$$

$$= \frac{1}{2}$$

Illustration 8.11

A glass jar contains 1 red, 3 green, 2 blue, and 4 yellow balls. If a single ball is chosen at random from the jar, what is the probability that it is yellow or green?

Solution:

$$P(yellow) = \frac{4}{10}, \qquad P(green) = \frac{3}{10}$$

$$P(yellow \text{ or } green) = P(yellow) + P(green)$$

$$= \frac{4}{10} + \frac{3}{10}$$

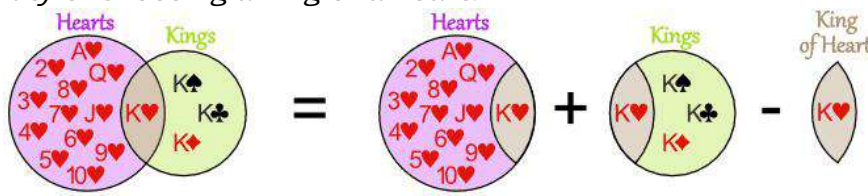
$$= \frac{7}{10}$$

In each of the three experiments above, the events are mutually exclusive. Let us look at some experiments in which the events are not mutually exclusive.

Illustration 8.12

A single card is chosen at random from a standard deck of 52 playing cards.

What is the probability of choosing a king or a heart?



Solution: .

$$P(\text{king or club}) = P(\text{king}) + P(\text{club}) - P(\text{king of clubs})$$
$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$
$$= \frac{4}{13}$$

$$P(\text{king or heart}) = P(\text{king}) + P(\text{heart}) - P(\text{king of hearts})$$
$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$
$$= \frac{4}{13}$$

Here the events are not mutually exclusive. We use Addition Rule 2: Where P(A and B) refers to the overlap of the two events.

Illustration 8.13

On New Year's Eve, the probability of a person having a car accident is 0.09. The probability of a person driving while intoxicated is 0.32 and probability of a person meeting with a car accident while intoxicated is 0.15. What is the probability of a person driving while intoxicated or meeting with a car accident?

Solution:

$$P(\text{intoxicated or accident}) = P(\text{intoxicated}) + P(\text{accident}) - P(\text{intoxicated and accident})$$
$$= 0.32 + 0.09 - 0.15$$
$$= 0.26$$

To find the probability of event A or B, we must first determine whether the events are mutually exclusive or non-mutually exclusive. Then we can apply the appropriate Addition Rule.

Addition Rule 1: When two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event.

$$P(A \text{ or } B) = P(A) + P(B)$$

Addition Rule 2: When two events, A and B, are non-mutually exclusive, there is some overlap between these events. The probability that A or B will occur is the sum of the probability of each event, minus the probability of the overlap.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

For any two events, A and B,

$$P(A \text{ or } B) \leq P(A)+P(B)$$
$$P(A \text{ and } B) \leq P(A)$$
$$P(A \text{ and } B) \leq P(B)$$

Know your progress

:

1. A person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit the target in 2 out of 4. If the probability of both hit the target is 0.5 find the probability of at least one hit the target?
2. A bag contains 30 balls numbered from 1 to 30. One ball is drawn at random. Find the probability that the number of the ball drawn will be a multiple of (a) 5 or 7 (b) 3 or 7 ?

8.5 Frequency approach to Probability

Probability from frequency distributions

A researcher asked 50 people if they liked the taste of a particular brand of coffee. The responses were “yes”, “no” and “undecided”.

Response	Frequency
Yes	30
No	16
Undecided	4

Probabilities of various categories can be computed now. For example, the probability of selecting a person who liked the taste is $\frac{30}{50}$ ie. $\frac{3}{5}$ since thirty out of 50 answered "yes".

Given the frequency distribution, the probability is computed as,

$$P(A) = \frac{\text{frequency of the class}}{\text{total frequency in the frequency distribution}}$$

In a sample of 100 persons, 42 had O-group blood, 44 had A-group blood, 10 had B-group blood 4 had AB-group blood. Find the probability that a person selected at random,

a) has O-group blood.

b) has A-group or B-group blood.

c) has neither A-group nor O-group blood.

d) does not have AB-group blood.

Solution:

Group	Frequency
A	44
B	10
AB	4
O	42

$$P(O) = \frac{42}{100}$$

$$\text{P(A or B)} = \frac{54}{100}\text{(frequencies of two classes added)}$$

$$\text{P(neither A nor O)} = \text{P(AB or B)}$$

$$= \frac{14}{100}$$

$$\text{P(not AB)} = 1 - \frac{4}{100} = \frac{96}{100}$$

For an experiment with equally likely outcomes, probabilities are identical to relative frequencies (or percentages)!

Statistical Regularity

When a coin is tossed once, it is a common knowledge that the probability of getting head is $\frac{1}{2}$. But what will happen when the coin is tossed 50 times? Will it results in 25 heads? Not all the time. One should expect about 25 heads if the coin is fair. But due to chance variation, 25 heads will not occur most of the time.

If the probability of getting a head is computed using a small number of trials, it is usually not exactly $\frac{1}{2}$. However if the number of trials increases, the probability of getting a head approaches the theoretical probability of $\frac{1}{2}$ if the coin is fair.

Suppose, the trials of an experiment is repeated n times and the event has occurred f times. The frequency ratio, $\frac{f}{n}$ approaches a constant for infinitely large values of n. This phenomenon is known as statistical regularity or law of large numbers.

Frequency definition or empirical definition of probability

The limiting value of for infinitely large n gives the probability of the event A.

$$P(A) = \lim_{n\to\infty} \frac{f}{n}$$

The difference between classical definition and empirical definition of probability is that classical probability assumes that certain outcomes are equally likely while empirical probability relies on actual experience to determine the likelihood of outcomes. In empirical probability one might conduct the experiment a large number of times and observe the relative frequencies to determine the probability of an outcome. Empirical probability is based on observations.

Activity

Collect data of birth registration for the last 20 years from the local body you belong to and compute the probability of male birth and female birth using data for 10 years and 20 years separately.

8.6 Axioms on Probability

For any event A attached to some random experiment, with a finite sample space S, we can associate a number P(A), known as the probability of the event A, subject to the following conditions.

Axiom 1 **Non-negativity**

For any event, A, $P(A) \geq 0$.

This axiom implies that, probability is never negative and its minimum is zero.

Axiom 2 **Certainty**

If S is the sample space, P(S) = 1.

This implies that, if we define any event, identical to the sample space, then any trial of the random experiment results in an outcome, favourable for the realisation of the event and occurrence of corresponding event is certain. Further this gives the maximum probability.

Axiom 3 **Additivity**

For any two events A_1 and A_2 such that they have no common elements, $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

That is, probability that least one of the events A_1 and A_2 having no common elementary outcomes is the sum of the probabilities.

Generalisation of Axiom 3:

$P(A_1 \cup A_2 \cup A_3 \ldots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$

where $A_1, A_2, A_3, \ldots A_k$ are having no elements in common.

8.7 Subjective Probability

Subjective probability uses a probability value based on an educated guess or employing opinions and inexact information. In subjective probability, a person or group makes an educated guess at the chance that the event will occur. This guess is based on the persons experience and evaluation of a solution. For example, a physician may say that there is a 30% probability that the patient may need a surgery based on his diagnosis. A sportswriter may say that there is 80% probability that Brazil will win the next FIFA world cup.

Let us sum up

In this chapter we introduced basic probability concepts and illustrated how probability analysis can be used to provide helpful information for decision making. We described how probability can be interpreted as a numerical measure of the likelihood occurrence of an event. We became familiar with the terminology of classical probability. In addition we saw the probability of events can be computed employing algebra of events, addition rule and counting rules. We also introduced empirical probability, subjective probability and axioms of probability.

Learning outcomes

After transaction of this unit, the learner:-

- recognises the amount of uncertainty that is involved before making important decisions.
- identifies random experiment, sample space, sample point, Events
- illustrates different approaches to probability.
- evaluates probability of events using classical definition of probability.
- evaluates joint probability of two events using addition rule.

Evaluation Items

Choose the correct answer from the given choice.

1. Which of the following is an experiment?
   a) Tossing a coin b) Rolling a single 6-sided die c) Choosing a ball from a jar d) All of the above.

2. Two coins are tossed together and its face is observed. Which of the following is an elementary outcome of this random experiment?
   a) Getting at least one Head b) Getting exactly one Head c) Getting utmost one Head d) Getting no Heads

3. What is the probability of choosing a vowel from the English alphabet?
   a) $\frac{21}{26}$ b) $\frac{5}{26}$ c) $\frac{14}{26}$ d) 0

4. A number from 1 to 11 (inclusive) is chosen at random. What is the probability of choosing an odd number?
   a) $\frac{1}{11}$ b) $\frac{6}{11}$ c) 0 d) 1

5. A day of the week is chosen at random. What is the probability of choosing a Monday or Tuesday?
   a) $\frac{5}{7}$ b) $\frac{2}{7}$ c) $\frac{1}{7}$ d) 0

6. In a pet store, there are 6 puppies, 9 kittens, 4 gerbils and 7 parakeets. If a pet is chosen at random, what is the probability of choosing a puppy or a parakeet?
   a) $\frac{21}{26}$ b) $\frac{5}{26}$ c) $\frac{13}{26}$ d) 0
7. A single 6-sided die is rolled. What is the probability of rolling a number greater than 3 or an even number?
   a) $\frac{5}{6}$ b) $\frac{2}{6}$ c) $\frac{1}{6}$ d) 0
8. P(A or B) = P(A) + P(B) means A and B are
   a) independent b) mutually exclusive c) equally likely d) none of these.
9. P(A or B) gives the probability of
   a) event A b) at least one event c) exactly one event c) event A and event B
10. When the outcomes of an experiment are equally likely, probability is identical with
    a) frequency b) relative frequency c) cumulative frequency d) total frequency

    Ans. 1-d, 2-d,3-b,4-b,5-b,6-c,7-a,8-b,9-b,10-b.

11. A coin is tossed twice, what is the probability that at least one tail occurs?
    Ans. $\frac{3}{4}$
12. A fair die is rolled. Find the probability of the following events:
    a) A prime number will appear.
    b) A number greater than or equal to 3 will appear.
    c) A number more than 6 will appear.
    d) A number less than 6 will appear.

    Ans. a) $\frac{3}{6}$ b) $\frac{4}{6}$ c) 0 d) $\frac{5}{6}$
13. A card is selected at random from a pack of 52 cards.
    a) How many points are there in the sample space?

b) Find the probability that the card is an ace of spades?

c) Find the probability that the card is a red card?

d) Find the probability that the card is an ace?

Ans. a) 52 b) $\frac{1}{52}$ c) $\frac{1}{2}$ d) $\frac{1}{13}$

14. There are 4 boys and 6 girls in a council. If one council member is selected at random, what is the probability that it is a girl?
Ans. $\frac{3}{5}$

15. Three coins are tossed once. What is the probability of getting:

a) 3 heads

b) 2 heads

c) At least two heads

d) At most two heads

e) No head

f) Exactly two tails

g) At most two tails.

Ans. a) $\frac{1}{8}$ b) $\frac{3}{8}$ c) $\frac{1}{2}$ d) $\frac{7}{8}$ e) $\frac{1}{8}$ f) $\frac{3}{8}$ g) $\frac{7}{8}$

16. If $\frac{2}{13}$ is the probability of an event A, what is the probability of event not A'?
Ans. $\frac{11}{13}$

17. A letter is chosen at random from the word **ASSASSINATION**' find the probability that the letter is (i) a vowel. (ii) a consonant.
Ans. (i) $\frac{6}{13}$ (ii) $\frac{7}{13}$

18. Check whether the following probabilities P(A) and P(B) are consistently defined?

i. P(A) = 0.5, P(B) = 0.7, P(A and B) = 0.6

ii. P(A) = 0.5, P(B) = 0.4, P(A or B) = 0.8

Ans. i) No, since P(A and B) >P(B) ii) Yes, since P(A or B) ≤ P(A) + P(B).

19. Fill the blanks in following table:

No.	P(A)	P(B)	P(A and B)	P(A or B)
i)	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{1}{15}$	
ii)	0.35		0.25	0.6
iii)	0.5	0.35		0.7

Ans. i)$\frac{7}{15}$ ii) 0.5 iii) 0.15.

20. Given P(A) = $\frac{3}{5}$ and P(B) = $\frac{1}{5}$. Find P(A or B) if A and B are mutually exclusive events.
Ans. $\frac{4}{5}$

21. If E and F are events such that P(E) = $\frac{1}{4}$, P(F) = $\frac{1}{2}$ and P(E and F) = $\frac{1}{8}$. Find (i) P(E or F) and (ii) P(not E and not F)?
Ans. i)$\frac{5}{8}$, ii)$\frac{3}{8}$

22. A and B are events such that P(A) = 0.42, P(B) = 0.48, P(A and B) = 0.16. Determine (i) P(not A) (ii) P(not B) and P(A or B) ?
Ans. i) 0.58 ii) 0.52 iii) 0.74

23. In school 40% students study music and 30% students study martial arts. 20% students study both music and martial arts. If a student is selected at random from the school, what is the probability that the student will be studying music or martial arts?
Ans. 0.5

24. An entrance test is graded on the basis of two examinations. The probability of a randomly chosen student passing the first examination is 0.8 and passing the second examination is 0.7. If the probability of passing at least one of the examination is 0.95, what is the probability that a student

passes both?

Ans. 0.55

25. In a driving test the probability that a candidate will pass both the H- test and 8-test is 0.5 and passing neither is 0.1. if the probability of passing the H-test is 0.75, what is the probability of passing the 8-test?

Ans. 0.65

26. From a commerce class of 60 students, 30 opted for B. Com, 32 opted for BBA, and 24 opted both. If one of these student is selected at random, find the probability that

    i. The student opted for B.Com or BBA?

    ii. The student has opted for neither B.Com nor BBA?

    iii. The student has opted for BBA but not B.Com?

Ans. i) $\frac{19}{30}$ ii) $\frac{11}{30}$ iii) $\frac{4}{30}$

27. A box containing 10 red balls, 20 blue balls and 30 green balls. 5 balls are drawn from the box. What is the probability that

    i. All will be blue?

    ii. At least one will be green?

Ans. i) $\frac{^{20}C_5}{^{60}C_5}$ ii) $1 - \frac{^{30}C_5}{^{60}C_5}$

28. 4 cards are drawn from a well-shuffled deck of 52 cards. What is the probability of obtaining 3 spades and one diamond?

Ans. $\frac{^{13}C_3 \times {}^{1}3C_1}{^{52}C_4}$

29. The following table gives the number of days stayed by maternity patients in a nursing home.

Days stayed	Number of patients
3	15
4	32
5	56
6	19
7	5
Total	127

Find the probability that,

i. A patient stayed exactly 5 days

ii. A patient stayed less than 6 days

iii. A patient stayed at most 4 days

iv. A patient stayed at least 5 days

Ans. i) $\frac{56}{127}$ ii) $\frac{103}{127}$ iii) $\frac{47}{127}$ iv) $\frac{80}{127}$

30. The particulars of the members of golden jubilee steering committee members of a school are given below:

Sl. no	Name	Sex	Age in years
1	Alis	F	28
2	Hari	M	30
3	Faisal	M	41
4	Mohan	M	33
5	Haritha	F	46

A person is selected at random from this group to act as the spokesperson. What is the probability that the spokesperson will be either male or over 35 years?

Ans. $\frac{4}{5}$

Introduction

In the last chapter we discussed the basic concepts of probability. Classical approach is used by mathematicians to help determine probability associated with experiments in which all possible cases are equally likely to occur. For example, the classical approach specifies that the probabilities of head and tail in the flip of balanced coin are equal. Since the sum of the probabilities must be 1, the probability of heads and the probability of tails are both 50%. Similarly, the six possible outcomes of the toss of a balanced die have the same probability; each is assigned a probability of $\frac{1}{6}$. In some experiments, it is necessary to develop mathematical ways to count the number of outcomes.

9.1 Meaning of Conditional Probability

Suppose Pradeep is participating in a Quiz Programme. For a question, 4 choices are given and he is not sure about the right answer. Clearly, the probability that his answer is correct is $\frac{1}{4}$. But suppose he knows that two of the choices are wrong. What is the probability of choosing a right answer? Clearly, the probability is $\frac{1}{2}$ as the possible number of outcomes is reduced to 2. Here the event is conditioned according to given additional information.

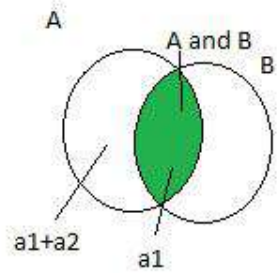
Let us consider another random experiment of drawing a card from a well shuffled pack of cards. Let the card drawn is red. What is the probability that it is also an ace ? We know that out of 52 cards, there are 26 red cards and 4 aces. Of the 4 aces, 2 are red. As the drawn card is red, the number of possible outcomes reduces to 26. Hence the required probability becomes $\frac{2}{26}$ or $\frac{1}{13}$.

9.2 Definition of Conditional Probability

The probability of an event B given that A has already occurred is denoted by $P(B|A)$, read as probability of the event B given that A has already occurred

or simply probability of B given A. A rule for finding conditional probability is derived as follows.

Let there be n outcomes in the sample space. Suppose a_1 cases favour 'A and B' and $a_1 + a_2$ cases favour A, of which a_2 cases favour the occurrence of 'only A'.



Then $P(B|A)$ is defined as

$$P(B|A) = \frac{a_1}{a_1 + a_2} = \frac{\frac{a_1}{n}}{\frac{(a_1 + a_2)}{n}} = \frac{P(A \text{ and } B)}{P(A)}, \text{ provided } P(A) \neq 0$$

Thus,

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}, \text{ provided } P(A) \neq 0$$

Similarly,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \text{ provided } P(B) \neq 0$$

The general question usually we ask is as follows: suppose we know that a certain event B has occurred. How does this affect the probability of some other event A ? This type of questions is addressed by the above mentioned conditional probabilities.

Illustration 9.1

A die is rolled once and it is known that the number shown up is greater than '4'. We have to find probability of the event that the die show up an even number.

Let A - Number shown is even. That is $A = \{2,4,6\}$ and B Number shown is greater than 4. That is $B = \{5,6\}$. Then A and $B = \{6\}$

The required probability is therefore, $P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{(1/6)}{(2/6)} = \frac{1}{2}$

Illustration 9.2

Consider the random experiment of tossing a coin twice. Let A be the event that at least one of the tosses result in head and B be the event that the first toss results in head. Here $A = \{HT, TH, HH\}$ and $B = \{HT, HH\}$ Also A and $B = \{HT, HH\}$,

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{(2/4)}{(3/4)} = \frac{2}{3}$$

Activity

Consider the random experiment of tossing a die. Let the face shown by the die is odd. What is the probability that the shown face is 3 ?

Activity

Consider the random experiment of throwing a die twice. Let A be the event of getting a total of 8 and B be the event that the die shows a number between 3 and 5 (both inclusive) in both throws. What is the probability of B given A ?

9.3 Multiplication Theorem

Sometimes we may need to find the probabilities of two or more events happening together. By definition of conditional probability, for two events A and B,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \quad P(B) \neq 0$$

and

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}, \quad P(A) \neq 0$$

By re-arranging the formula, we have

The probability of occurrence of both A and B as follows

$$P(A \text{ and } B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

This result is termed as *multiplication theorem*.

9.4 Independent and Dependent Events

Let there be 2 blue and 3 red balls in a bag. If one ball is drawn randomly from the bag, what is the probability of getting a blue ball?

Clearly, the probability is $\frac{2}{5}$.

Suppose two balls are drawn one by one without replacing the first ball. Consider the following two situations:

- if you got a **red** ball first, then the probability of a blue ball next is $\frac{2}{4}$
- if you got a **blue** ball first, then the probability of a blue ball next is $\frac{1}{4}$

See how the chances change each time. Each event **depends on** what happened in the previous event, and are called **dependent**.

Consider the above two situations if the first ball is replaced back. Both probabilities are same and is $\frac{2}{5}$. Here the happening of the second event is not affected by the previous event, and are called **independent**.

- With Replacement: the events are Independent (the chances don't change)
- Without Replacement: the events are Dependent (the chances change)

Two events are said to be *independent* if the occurrence of one event do not affect the occurrence of the other. Otherwise the events are said to be *dependent*

Examples of independent events are:

- Choosing a 3 from a deck of cards, replacing it, and then choosing an ace as the second card.

- Getting head both times when a coin is tossed twice.

If $P(A|B) = P(A)$ and $P(B|A) = P(B)$, then A and B are independent.

- Independence is also termed as stochastic independence.

Activity

Write similar examples for identifying independent and dependent events

Multiplication Theorem for Independent Events

If A and B are independent, multiplication theorem becomes

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Activity

Extend the multiplication theorem to three or more independent events.

Illustration 9.3

A dresser drawer contains five pairs of socks each with one of the following colours: blue, brown, red, white and black. Each pair is folded together in a matching set. You reach into the sock drawer and choose a pair of socks without looking. Now, You replace this pair with another pair.What is the probability that you will choose the red pair of socks both times ?

Solution.

$$\begin{aligned} P(\text{red}) &= \frac{1}{5} \\ P(\text{red and red}) &= P(\text{red}) \cdot P(\text{red}) \\ &= \frac{1}{5} \cdot \frac{1}{5} \\ &= \frac{1}{25} \end{aligned}$$

Illustration 9.4

A coin is tossed and a single 6-sided die is rolled. Find the probability of landing on the head side of the coin and rolling a 3 on the die.

Solution. Then the Probabilities will be as follows

$$\begin{aligned} P(\text{head}) &= \frac{1}{2} \\ P(3) &= \frac{1}{6} \\ P(\text{head and } 3) &= P(\text{head}) \cdot P(3) \\ &= \frac{1}{2} \cdot \frac{1}{6} \\ &= \frac{1}{12} \end{aligned}$$

Illustration 9.5

Consider the random experiment of drawing two cards from a well shuffled pack of cards one by one. What is the probability that both are kings if

(i) the first card is replaced back

(ii) the first card is not replaced back

Solution. Let A be the event of drawing a King first, and B, the event of drawing a King second.

(i) If cards are taken with replacement, then the events are independent. $P(A) = \frac{4}{52}$ and $P(B) = \frac{4}{52}$.

$$P(A \text{ and } B) = P(A) \cdot P(B) = \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$

(ii) If the first card is not replaced, then the events are dependent. $P(A) = \frac{4}{52}, P(B|A) = \frac{3}{51}$

And so,

$$P(A \text{ and } B) = P(A) \times P(B|A) = \frac{4}{52} \times \frac{3}{51} = \frac{12}{2652} = \frac{1}{221}$$

Illustration 9.6

A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black

Solution. Let A be the event that a black ball is drawn in the first attempt and B, the event of drawing a black ball in the second attempt.

$$P(A) = \frac{3}{5+3} = \frac{3}{8}$$

Probability of drawing the second black ball given that the first ball drawn is black

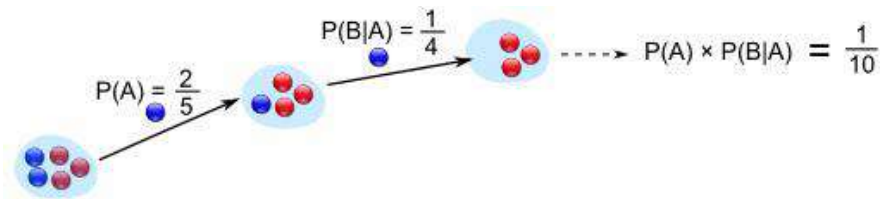
$$P(B|A) = \frac{2}{5+2} = \frac{2}{7}$$

The probability that both balls drawn are black is given by

$$P(A \text{ and } B) = P(A) \times P(B|A) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$

Illustration 9.7

Consider the figure given below, The probability of getting 2 blue balls if the first ball is not replaced before the second draw is computed below.



Solution. The probability of getting 2 blue balls if the first ball is replaced before the second draw is $P(A \text{ and } B) = P(A) \cdot P(B) = \frac{2}{5} \times \frac{2}{5} = \frac{4}{25}$

Illustration 9.8

A box contains a total of 100 CDs that are manufactured with two machines. Of them, 60 were manufactured with machine 1. Of the total of CD's 15 are defective. Of the 60 CD's that were manufactured with machine 1, 9 are defective.

Solution. Let B be the event that a randomly selected CD is defective and let A be the event that a randomly selected CD was manufactured with machine 1. Are the events B and A are independent ?

$$P(B) = \frac{15}{100} = 0.15$$
$$P(B|A) = \frac{9}{60} = 0.15$$

Since $P(B) = P(B|A)$; A and B are independent.

Illustration 9.9

The probability that a problem of statistics will be solved by a student X is $\frac{1}{3}$ and a Student Y is $\frac{2}{3}$. What is the chance that the problem will be solved ?

Solution. The problem will be solved if either X or Y or both solves the problem. Let A be the event that X solves the problem and B the event that Y solves the problem. By addition theorem of probability,

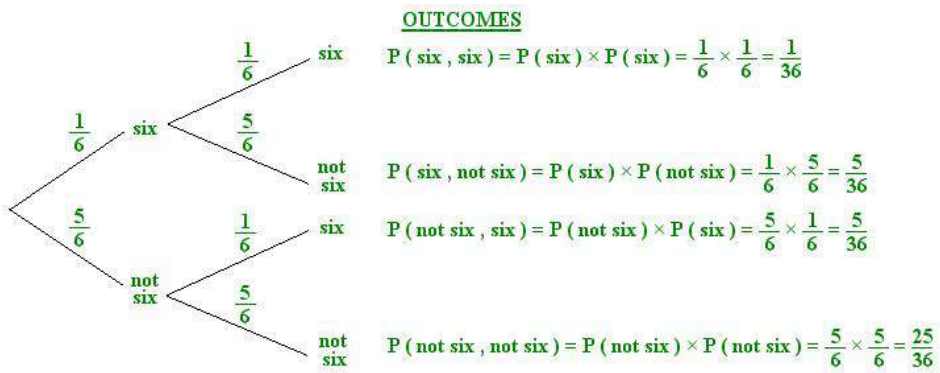
$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= P(A) + P(B) - P(A)P(B) \quad [\text{since } A \text{ and } B \text{ are independent}] \\ &= \frac{1}{3} + \frac{2}{3} - \frac{1}{3} \times \frac{2}{3} \\ &= 1 - \frac{2}{9} \\ &= \frac{7}{9} \end{aligned}$$

Illustration 9.10

Two fair dice are rolled. Introduce a new idea of probability tree diagram to determine the probability of obtaining:

(a) two sixes, (b) no sixes, (c) exactly one six.

Solution.



So, the answers are as follows:

(a) $P(\text{two sixes}) = P(\text{six, six}) = \frac{1}{36}$

(b) $P(\text{no sixes}) = P(\text{not six, not six}) = \frac{25}{36}$

To find the probability of "exactly one six", we need to combine the outcomes for (six, not six) and (not six, six), as they both have "exactly one six". We combine these outcomes by adding their probabilities:

(c) $P(\text{exactly one six}) = P(\text{six, not six}) + P(\text{not six, six}) = \frac{10}{36}$

See the end probabilities, the total of end probabilities will be one. Here the sample space is splitting up to different sections or divisions so that probabilities can be easily found out from the known probabilities.

Activity

Try to draw tree diagram on the following example for with and without replacement.

A box contains 20 DVDs, 4 of which are defectives. Two DVD s are selected at random from the box what is the probability that

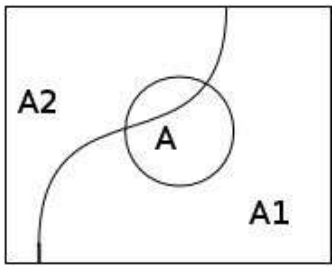
(i) both are defective

(ii) no defectives and

(iii) exactly one defective.

9.5 Total Probability Theorem.

Suppose there are 50 students in a class. Of them, 30 are boys and the remaining are girls. 15 students are members of music club.9 of them are boys. Suppose we have to find the probability that a randomly selected student is a member of music club.



Let A be the event that the student is a member of music club and A_1 and A_2 the events that the selected student is a boy and girl respectively. A can happen in the following two mutually exclusive and exhaustive ways.

(i) The student from music club is a girl.

(ii) The student from music club is a boy.

$$
\begin{aligned}
P(A) &= P(A \text{ and } A_1) \text{ or } P(A \text{ and } A_2) \\
&= P(A_1)P(A|A_1) + P(A_2)P(A|A_2), \text{ using multiplication theorem in each term} \\
&= \frac{30}{50} \times \frac{9}{30} + \frac{20}{50} \times \frac{6}{20} \\
&= \frac{270}{1500} + \frac{120}{1000} \\
&= \frac{900}{3000} \\
&= \frac{3}{10}
\end{aligned}
$$

The above explained example leads to the general idea of total probability theorem.

If A_1 and A_2 are two mutually exclusive and exhaustive events and A is any other event which can occur along with A and B, then total probability theorem states that

$$P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2)$$

Let us consider the case for 3 events.

Suppose in a factory out of 100 bolts produced, 50 are produced by machine A, 30 are produced by machine B and 20 by machine C. The number of defectives bolts from these 3 machines be 9, 6 and 4 respectively. Suppose we have to find the probability that a randomly selected bolt is defective.

Let A_1, A_2, A_3 be the events that the randomly selected bolt is produced by machine A, B and C respectively. Let A be the event that the selected bolt is defective. The event A can happen in the following 3 mutually exclusive and exhaustive ways.

(i) The defective bolt is produced by machine A

(ii) The defective bolt is produced by machine B

(iii) The defective bolt is produced by machine C

$$\begin{aligned}
P(A) &= P(A \text{ and } A_1) \text{ or } P(A \text{ and } A_2) \text{ or } P(A \text{ and } A_3) \\
&= P(A \text{ and } A_1) + P(A \text{ and } A_2) + P(A \text{ and } A_3) \\
&= P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3) \\
&= \frac{50}{100} \times \frac{9}{50} + \frac{30}{100} \times \frac{6}{30} + \frac{20}{100} \times \frac{4}{20} \\
&= \frac{9}{100} + \frac{6}{100} + \frac{4}{100} \\
&= \frac{19}{100} \\
&= 0.19
\end{aligned}$$

The Total Probability Theorem for 3 events can be stated as follows:

If A_1, A_2 and A_3 are three mutually exclusive and exhaustive events and A is any other event which can occur along with A_1, A_2 and A_3, then

$$P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3)$$

Similarly the total probability theorem extended to n events can be stated as follows:-

If $A_1, A_2, A_3, \ldots, A_n$ are n mutually exclusive and exhaustive events and A is any other event which can occur along with $A_1, A_2, A_3, \ldots, A_n$, then

$$P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3) + \cdots + P(A_n).P(A|A_n)$$

Illustration 9.11

One bag has 4 white balls and 3 black balls and a second bag contains 3 white balls and 5 black balls. One ball is drawn from first bag and placed without noticing the colour in the second bag. What is the probability that a ball now drawn from second bag is black?

Solution.

Bag - I	Bag - II
$4W$	$3W$
$3B$	$5B$

Let A_1 be the event that white ball is transferred from bag-I to bag-II and A_2 be the event that black ball is transferred from bag-I to bag-II.

$$P(A_1) = \frac{4}{7}, \quad P(A_2) = \frac{3}{7}$$

Let 'A' be probability the finally a black ball is drawn from the second bag. Then

$$P(A|A_1) = \frac{5}{9}, \quad P(A|A_2) = \frac{6}{9}.$$

Now from the total probability theorem we get,

$$P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2) = \frac{4}{7} \cdot \frac{5}{9} + \frac{3}{7} \cdot \frac{6}{9} = \frac{38}{63}$$

Illustration 9.12

Suppose a doctor is going to meet his patients. From previous experience it

is expected that the probability for his arrival, if he chooses train, bus and scooter, is $\frac{3}{10}, \frac{1}{5}$ and $\frac{1}{10}$ respectively. If he prefer to train, the probability for his late arrival is $\frac{1}{4}$. If he prefer to bus, the probability that his late arrival is $\frac{1}{3}$ and if he prefer scooter, the probability of his late arrival is $\frac{1}{12}$. What will be the probability for his late arrival?

Solution. Let A_1 be the event the doctor come by train, then $P(A_1) = \frac{3}{10}$,
A_2 be the event the doctor come by bus, then $P(A_2) = \frac{1}{5}$,
A_3 be the event the doctor come by scooter, then $P(A_3) = \frac{1}{10}$.
Let A be event that the doctor is late.
Given that $P(A|A_1) = \frac{1}{4}$; $P(A|A_2) = \frac{1}{3}$; $P(A|A_3) = \frac{1}{12}$. Then

$$\begin{aligned} P(A) &= P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3) \\ &= \frac{3}{10} \times \frac{1}{4} + \frac{1}{5} \times \frac{1}{3} + \frac{1}{10} \times \frac{1}{12} \\ &= \frac{3}{20} \end{aligned}$$

Activity

- Draw tree diagram for the Illustration 9.12
- Suppose we have two hats: one has 4 red balls and 6 green balls, the other has 6 red and 4 green. We toss a fair coin, if heads, pick a random ball from the first hat, if tails from the second. What is the probability of getting a red ball ?

9.6 Bayes' Theorem

Consider the problem related to bolt manufacturing factory we discussed in total probability theorem. In that case, we found out the probability of the selected bolt being defective.

Suppose a randomly selected bolt is defective. It is interesting to know the probability that it is produced by machine A ? Bayes' theorem gives a solution for this.

We have to find $P(A_1|A)$. We have $P(A_1|A) = \frac{P(A \text{ and } A_1)}{P(A)}$. By total probability theorem,

$$P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3)$$

$$\text{and } P(A \text{ and } A_1) = P(A)P(A_1|A)$$

So

$$\begin{aligned} P(A_1|A) &= \frac{P(A_1)P(A|A_1)}{P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3)} \\ &= \frac{\frac{50}{100} \times \frac{9}{50}}{\frac{50}{100} \times \frac{9}{50} + \frac{30}{100} \times \frac{6}{30} + \frac{20}{100} \times \frac{4}{20}} \\ &= \frac{\frac{9}{100}}{\frac{9}{100} + \frac{6}{100} + \frac{4}{100}} \\ &= \frac{\frac{9}{100}}{\frac{19}{100}} \\ &= \frac{9}{19} \end{aligned}$$

This is Bayes' theorem for three events. The table given below simplifies the computation of probability using Bayes' theorem.

Elementary Events	Prior Probability $P(A_i)$	Conditional Probability $P(A\|A_i)$	Joint Probability $P(A_i)P(A\|A_i)$	Posterior Probability $P(A_i\|A)$
A_1	$\frac{50}{100}$	$\frac{9}{50}$	$\frac{9}{100}$	$\frac{9}{19}$
A_2	$\frac{30}{100}$	$\frac{6}{30}$	$\frac{6}{100}$	$\frac{6}{19}$
A_3	$\frac{20}{100}$	$\frac{4}{20}$	$\frac{4}{100}$	$\frac{4}{19}$
Total	1		$\frac{19}{100}$	1

From the table, it is clear that $P(A_1|A) = \frac{9}{19}$

The probabilities $P(A_1), P(A_2), P(A_3)$ are called prior probabilities. These probabilities are known before the random experiment is conducted. $P(A_1|A), P(A_2|A), P(A_3|A)$

are called posterior probabilities. These probabilities are based on evidence or experimentation

Bayes' theorem can be generalised as follows.

Suppose $A_1, A_2, A_3, \ldots, A_n$ are n mutually exclusive and exhaustive events and A is any event which can occur with at least one of the events $A_1, A_2, A_3, \ldots, A_n$.

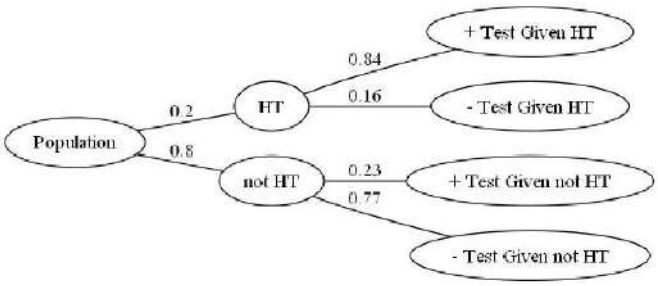
$$P(A_i|A) = \frac{P(A_i)P(A|A_i)}{P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + \cdots + P(A_n)P(A|A_n)}$$

for any $i = 1, 2, 3, \ldots, n$

Bayes' probability is also known as inverse probability. Bayes theorem is named after British Mathematician Thomas Bayes (1702-1761)

Illustration 9.13

Consider the following example, Assume 20% of adults are hypertensive (HT) Assume a testing machine diagnoses 84% of HT adults correctly and 23% of not HT adults



A Person is selected at random and for him, the test is positive. What is the probability that he is coming from the category hypertensive ?

Solution. By applying Baye's Theorem we have

P(The Person coming from Hypertensive category given that the test is positive)

$$= \frac{0.84 \times 0.2}{0.84 \times 0.2 + 0.8 \times 0.23}$$
$$= 0.48$$

Activity

Consider the problem of screening for breast cancer. A doctor discovers a lump in a womans breast during a routine physical exam. The lump could be cancerous. Without performing any further tests, the probability that the lump is a cancer is 0.01. A mammogram is a test that, on average, is correctly able to establish whether a lump is benign or cancerous 90% of the time. What is the probability that the lump is cancerous if the test result from a mammogram is positive ?

Illustration 9.14

Consider the illustration 9.10. If the question is "If he arrives at late, what is the probability that he comes by bus?"

Solution. In this case we have to apply Baye's theorem. That is, we have to find $P(A_2|A)$ which will be the ratio of P(he is coming late by bus) and P(he is late by any of modes of travel).

That is, P(he is coming late by bus) $= P(A_2) \times P(A|A_2) = \frac{1}{3} \times \frac{1}{5} = \frac{1}{15}$. Therefore $P(A_2|A) = \frac{4}{9}$.

Illustration 9.15

In a multiple choice question with four choices an examinee answers by guessing or copying or from his own knowledge. The probability that he makes a guess is $\frac{1}{3}$ and the probability that he copies the answer is $\frac{1}{6}$. The probability that his answer is correct given that he copied it is $\frac{1}{8}$. Find the probability that he actually knew the answer to the question, given that he correctly answered it ?

Solution.

A_1 : Examinee guesses the answer to the question

A_2 : Examinee copies the answer to the question

A_3 : Examinee knows the answer to the question

A : The answer is correct

Then $P(A_1) = \frac{1}{3}, P(A_2) = \frac{1}{6}, P(A_3) = 1 - \left[(P(A_1) + P(A_2)\right] = \frac{1}{2}$. Since the question is multiple choice with 4 choices, $P(A|A_1) = \frac{1}{4}, P(A|A_2) = \frac{1}{8}$ (Given) and $P(A|A_3) = 1$. Then

$$P(A_3|A) = \frac{P(A|A_3) \times P(A_3)}{P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3)}$$
$$= \frac{24}{29} \quad \text{(By substitution of probability values)}$$

Activity

Suppose 8 men out of 100 and 30 women out of 10000 are colour blind. A colour blind person is chosen at random. What is the probability of his being male ? (Assume males and females to be in equal number)

Let us sum up

- Conditional probability is the probability of an event conditioned according to given additional information.
- Probability of an event A given that B has happened is $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, P(B) \neq 0$. Similarly, $P(B|A) = \frac{P(A \text{ and } B)}{P(A)}, P(A) \neq 0$.
- Multiplication theorem of probability defines the probability of simultaneous occurrence of two or more events.

  For two events A and B, $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$
- Two events are said to be independent if the occurrence of one event doesnt affect the occurrence of the other, otherwise they are said to be dependent.
- For two independent events A and A, $P(A \text{ and } B) = P(A)P(B)$
- **Total probability theorem :** Let $A_1, A_2, A_3, \ldots, A_n$ are n mutually exclusive and exhaustive events and A is any other event which can occur along with at least one of the events $A_1, A_2, A_3, \ldots, A_n$. Then,

$$P(A) = P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + P(A_3)P(A|A_3) + \cdots + P(A_n)P(A|A_n)$$

- Bayes' theorem gives the posterior probability of an event. ie., if $A_1, A_2, A_3, \ldots, A_n$ are n mutually exclusive and exhaustive events and A is any other event which can occur along with at least one of the events $A_1, A_2, A_3, \ldots, A_n$, then

$$P(A_i|A) = \frac{P(A_i)P(A|A_i)}{P(A_1)P(A|A_1) + P(A_2)P(A|A_2) + \cdots + P(A_n)P(A|A_n)}$$

for any $i = 1, 2, 3, \ldots, n$

Learning outcomes

After transaction of this unit, the learner:-

- explains the meaning and concept of conditional probability.
- recognises multiplication theorem of probability and solves the problems.
- differentiates independent and dependent events.
- identifies total probability theorem.
- illustrates Baye's theorem and solves problems.

Evaluation items

1. Given that $P(A) = \frac{1}{5}, P(B) = \frac{2}{3}$ and $P(A \cup B) = \frac{4}{5}$. Find $P(A|B)$ and $P(B|A)$
2. A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.
3. A bag contains 10 white and 6 black balls. 4 balls are successively drawn out and not replaced. What is the probability that they are alternately of different class ?
4. In a product testing procedure, each radio on an assembly line must pass through two inspection points before being packed for shipment. The probability of detection of a defective radio of the first inspection point is 0.7 and at the second inspection point is 0.8. What is the probability of packing a defective radio packed for shipment ?

5. The probability of X, Y and Z becoming managers are $\frac{4}{9}, \frac{2}{9}$ and $\frac{1}{3}$ respectively. The probabilities that a bonus scheme will be introduced if X, Y and Z become managers are $\frac{3}{10}, \frac{1}{2}$ and $\frac{4}{5}$ respectively.

   (i) What is the probability of introduction of the bonus scheme?

   (ii) If the bonus scheme is introduced, what would probability be the manager appointed was Z ?

6. A class consists of 25 students, out of whom 5 are girls. In the class, 2 girls and 5 boys are ranke holders in the previous examination. If a student is selected at random and is found to be a rank holder, then what is the probability that the student selected is a girl ?

7. In an examination, 30% students have failed in Mathematics, 20% have failed in Chemistry and 10% have failed in both Mathematics and Chemistry. A student is selected at random. What is the probability that the student has failed in Mathematics if it is known that he has failed in Chemistry ?

8. A class consists of 80 students, 25 of them are girls and 55 are boys. 10 of them are rich And the remaining are poor. 20 of them are fair complexioned. What is the probability of selecting a fair complexioned rich girl ?

9. Three groups of workers contain 3 men and one woman, 2 men and 2 women and one man and 3 women respectively. One worker is selected at random from each group. What is the probability that the group selected consists of 1 man and 2 women ?

10. Three bags contain 6 red, 4 black; 4 red, 6 black and 5 red and 5 black balls respectively. One of the bags is selected at random and a ball is drawn from it. If the ball drawn is red, find the probability that it is drawn from the first bag ?

11. A bag contains 4 white and 3 black balls. Two draws of 2 balls are successively made. What is the probability of getting 2 white balls at first draw and 2 black balls at second draw when balls drawn at first draw were replaced ?

12. A number selected randomly from each of the two sets

    1 2 3 4 5 6 7 8

    2 3 4 5 6 7 8 9

    What is the probability that the sum of the numbers equal to 9?

13. A machine part is produced by three factories A, B and C. Their production is 25, 35 and 40 percent respectively. Also the percentages of defective machines manufactured by three factories are 5, 4 and 3 respectively. A part is taken at random and is found to be defective. What is the probability that the selected part belongs to the factory B ?

14. From a pack of 52 cards two cards are drawn, the first is replaced before the second is drawn. Find the probability that the first one is diamond and the second is a king.

15. Among the workers in a factory, only 30% receive bonus. Among those receiving bonus only, 20% are skilled. What is the probability of a randomly selected worker who is skilled and receiving bonus ?

16. If A and B are two events such that $P(A) = \frac{1}{3}, P(B) = \frac{3}{4}$ and $P(A \cup B) = \frac{11}{12}$. Find $P(A|B)$ and $P(B|A)$

17. A Bag contains 5 red and 3 black balls. Another bag had 4 red and 5 black balls. One of the bags is selected at random and a draw of 2 balls is made from it. What is the probability that one of them is red and the other is black. Also draw tree diagram.

18. Urn I contains three green and five red balls. Urn II contains two green and one red and two yellow balls. We select an Urn at random and then draw one ball at random from the Urn. What is the probability that we obtain a green ball ?

19. An Urn contains 10 white and 3 black balls, while another Urn contains 3 white and 5 black balls. Two balls are drawn from the first urn and put into the second urn and then a ball is drawn from the latter. What is the probability that it is a white ball ?

20. The contents of 3 Urns are

Urn I	1 White	2 Red	2 Black Balls
Urn II	3 White	1 Red	1 Black Balls
Urn III	3 White	3 Red	3 Black Balls

Two balls are chosen from a randomly selected urn. If the balls are drawn are 1 white and 1 red ball, what is the probability that they come from Urn II

Answers:

$1)P(A|B) = \frac{1}{10}, P(B|A) = \frac{1}{3}$ $2)\frac{3}{28}$ $3)\frac{45}{364}$ $4)0.06$

$5)$ $(i)0.511$ $(ii)0.52$ $6)\frac{2}{7}$ $7)\frac{1}{2}$ $8)\frac{5}{512}$

$9)\frac{13}{32}$ $10)\frac{\frac{6}{10}\times\frac{1}{3}}{\frac{6}{10}\times\frac{1}{3}+\frac{4}{10}\times\frac{1}{3}+\frac{5}{10}\times\frac{1}{3}}$ $11)\frac{2}{49}$

$12)\frac{7}{64}$ $13)0.36$ $14)\frac{1}{52}$ $15)0.06$

$16)\frac{2}{9}\text{and}\frac{1}{2}$ $17)\frac{275}{504}$ $18)\frac{31}{80}$

$19)\frac{10C_2}{13C_2}\times\frac{5}{10}+\frac{3C_2}{13C_2}\times\frac{3}{10}+10C_1\times\frac{3C_1}{13C_2}\times\frac{4}{10}$ $20)\frac{2}{5}$

Introduction

In chapter 2, we discussed census and sample survey. The term population is used to mean the totality of items in an investigation. The population may be finite or infinite. If each and every unit in the population is considered in the enquiry, it is called census or complete enumeration. As the population in most enquires are quite large, complete enumeration is not practical or feasible. In such cases, a representative part of the population is taken into consideration. This method is called sampling. Each unit in the population is called a sampling unit. The list of all sampling units of the population is called sampling frame. It is basically, a list from which you can choose your sample

Activity

Look at the following situations. What is your population? What is a sampling unit? How would you develop a sampling frame?

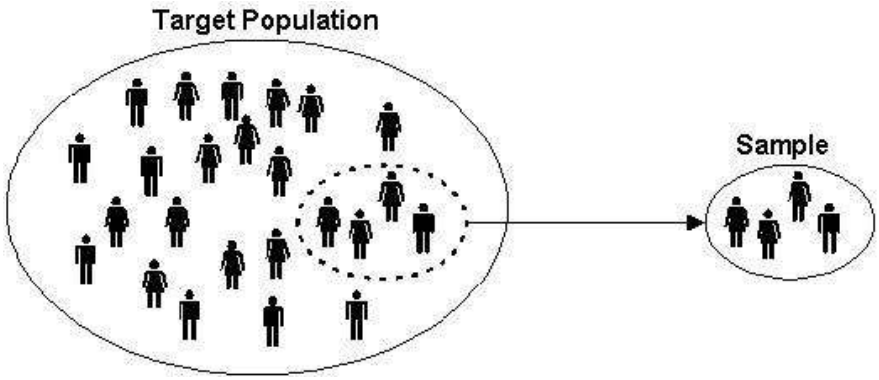
- Monalisa health club wants to conduct a survey to see the facilities that customers expect?
- A chocolate manufacturing company has introduced a new variety of chocolates. They want to check the quality of those chocolates.

Census Survey

The traditional method of acquiring knowledge about an aggregate of individuals is to enumerate them all. Census of population of a country and study of total agricultural production of a country are examples of complete enumeration. The major statistics of agriculture production, distribution of labour force and unemployment have all been based on census approach.

Surveying male smokers between 20 and 40 age who had surgery for throat cancer in a particular hospital during last year, surveying number of printing mistakes in a book, surveying the number of accidents reported in a locality etc. are also examples for census study.

Figure 10.1:



A pioneer in census taking was Pierre-Simon de Laplace. In 1780, he developed the Laplace method of estimating the population of a country. The principle behind his method was to take a census of a few selected communities and determine the ratio of the population to the number of births in these communities. This number would be used to multiply the number of births in the entire country to estimate the number of citizens in the country.

Know your progress

From chapter 2, recall the situations where census and sampling can be implemented.

10.1 Need and Importance of Sampling

In homes, we take out one or two rice grains (any other food item) from the cooking pan to examine whether the grains are fully cooked or not. In clinical laboratory, a few drops of blood are collected to test to know whether the blood has any abnormalities or not.

Whatever observed in the few drops of blood is true for whole blood of the body. In a bulb manufacturing company, one tests the life of few bulbs and comes to a conclusion about the average life of bulbs in the whole lot. Similarly, instead of examining the whole population, which may be difficult or impossible to do, one can examine a representative 3 part of the population. This is called sample. The process of drawing samples is called sampling. These examples reveal that sampling is an age old practice. Now-a-days, sampling methods are extensively used in socio-economic surveys to know the living condition, cost of living index etc of a class of people.

Let us consider the studies to get the following information

- The percentage of Keralites who have access to internet
- Opinion about selection of places for a study tour with the students of your class.
- Any opinion poll on various political leaders of Kerala

In first and third case, it is obvious that interviewing more than 3 crore people is not possible. The process is costly, time consuming and requires a lot of trained investigators. Due to these reasons it is convenient to use sampling method. But in second case, data may be collected through census

10.2 Census and Sampling - Advantages and Disadvantages

Table 10.1: Census:- Advantages and Disadvantages

Advantages	Disadvantages
(i) 100% perfect.	(i) If the population is infinite, the study is impossible.
(ii) For an enquiry, if all the units in the population are to be inspected, census is the only method	(ii) Census method requires more time, money, trained persons etc
(iii) The data obtained by census method may be used for further investigations	(iii) If the units are destroyed in the course of inspection, census is not at all desirable.

Table 10.2: Sampling:- Advantages and disadvantages

Advantages	Disadvantages
(i) Saves time and money	(i) Improper sampling technique may lead to misleading results
(ii) When population is too large or items are destructive in nature, sampling alone can be adopted (eg. Testing the strength of chalks)	(ii) If information is required from each and every unit of population, sampling is inadequate (eg: Leprosy test in school children)
(iii) More accuracy is expected	
(iv) The expected error also can be estimated	

Activity

Find out situations where sampling is preferred to census and vice versa

Errors in survey

The results obtained from statistical studies may not be free from errors. The errors involved in various stages of the study of data may be broadly classified under two heads viz (i) sampling error and (ii) non-sampling error.

10.3 Sampling and Non-Sampling Errors

Sampling errors are seen in sample surveys due to the fact that only a part of the population is used for enquiry. Clearly, sampling errors are absent in census. For example, the estimate of average income of people in certain region obtained on the basis of sampling will not be equal to the true average income.

Sampling errors cannot be completely eliminated but may be minimized by choosing a proper sample of adequate size and a proper sample survey design. Sampling error decreases as sample size increases. Sampling errors can be detected, measured and controlled. Sampling errors arise due to the following

(i) Lack of clarity about the coverage of the population

(ii) Faulty selection of the sample

(iii) Inadequate sample size

(iv) Inappropriate questionnaire

(v) Errors due to substitution

Errors other than sampling errors in a survey are called non-sampling errors. Non sampling errors arise at various stage of observation and processing of data, presentation and printing of tabulated results and are thus present in both census and sampling. Thus the data obtained in complete enumeration, although free from sampling errors would still be subject to non sampling errors. Data obtained in a sample survey would be subject to both sampling and non sampling errors. These errors can be minimized by choosing proper sampling designs, employing efficient investigators and better sampling. Non sampling errors usually increase with increase in sample size. Non sampling errors may arise due to

1. Irrelevant responses to questions
2. Errors in printing and publication of results
3. Errors in data processing

Know your progress

Distinguish between sampling and non sampling errors

Methods of Sampling

Consider a study about the spending habits of students in a class of 60. A sample of 10 students is selected for the study. The investigator can select the sample according to convenience. This type of samping is termed as **Non Probability Sampling**. On the other hand the investigator can select the sample randomly, in which each member of the population has some specified probability of being included in the sample. This method is termed as **Probability Sampling**.

10.4 Non Probability Sampling

In non probability sampling, members are selected from the population in some non- random manner. These include convenience sampling, judgment sampling and quota sampling. These methods are subjective.

Convenience Sampling

A convenience sample is obtained by selecting convenient population units. A sample obtained from readily available lists such as automobile registrations, telephone directories, etc. is a convenience sample, if the sample is drawn according to the convenience of the investigator. The results obtained by this method will not be a representative of the population. This method is very popular in online research and Traditional ″man on the street″ interviews conducted frequently by the visual media.

Judgment Sampling

The investigator exercises his judgment in the choice and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics under investigation. For example, if a sample of ten students is to be selected from a class of sixty for analyzing the spending habits of students, the investigator would select 10 students who, in his opinion, are

representative of the class. In this sampling the sample is selected with definite purpose in view. If the investigator is experienced and skilled and the sampling carefully applied, then the judgment samples may yield valuable results. This method is also very useful when you need to reach a targeted sample quickly.

Quota Sampling

In this method quotas are set up according to some specified characteristics such as several income groups. Within the quota the selection of sample items depends on personal judgment. For example in an income survey, the interviewers may be told to interview 100 people living in certain area in which 60 are housewives, 25 are regular employees and 15 are businessmen. Within these quotas the interviewer is free to select the people to be interviewed. This method often used in public opinion studies and personal interviews and people are systematically according to some fixed quota.

Activity

Find out similar situations where non probability sampling is appropriate

10.5 Probability Sampling

Probability sampling is the scientific method of selecting samples according to some laws of chance in which each unit in the population has some definite pre-assigned probability of being included in the sample. This method is purely objective.

Different types of probability sampling includes

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Random Sampling
4. Cluster Sampling

5. Multistage Sampling

There are many situations which demanded probability sampling or non probability sampling or a combination of both. Search them out.

10.6 Simple Random Sampling (SRS)

Simple Random Sampling is a probability sampling in which each unit in the population has an equal chance of being included in the sample. In this case the sampling units are selected at random. Simple random sampling overcome the drawbacks of non-probability sampling viz favouritism, subjectiveness etc. This method is applicable when population is homogeneous.

There are two types of Simple Random Sampling - Simple Random Sampling Without Replacement (SRSWOR) and Simple Random Sampling With Replacement (SRSWR). Suppose you are going to buy orange from a fruit shop. You are selecting five oranges one by one from a basket of oranges without replacing the selected ones. This type of sampling in which all units have an equal chance of being included in the sample is called as simple random sampling without replacement. If the sampling is done by replacing the selected unit it is called simple random sampling with replacement. If a population consists of N units and a sample of n units to be taken, the possible number of samples in SRSWOR is ${}^{N}C_{n}$ and in SRSWR is N^{n}.

Illustration 10.1

If a population consists of 5 numbers 2,3,6,8 and 11

Consider all simple random samples of size 2 that can be drawn

1. *with replacement*
2. *without replacement*

Samples: Using SRSWOR

(2,3), (2,6), (2,8), (2,11), (3,6), (3,8), (3,11), (6,8), (6,11), (8,11)

${}^5C_2 = 10$ *samples*

Samples : Using SRSWR

(2,2), (2,3), (2,6), (2,8), (2,11), (3,2), (3,3), (3,6), (3,8),(3,11), (6,2), (6,3), (6,6), (6,8), (6,11), (8,2), (8,3), (8,6), (8,8), (8,11), (11,2), (11,3), (11,6), (11,8), (11,11)

$5^2 = 25$ *samples*

Activity

Solve the following questions.

1. Suppose we have 5 cards numbered from 1 to 5 and two cards are to be selected, write all possible samples using (i) SRSWOR and (ii) SRSWR
2. A bag contains 10 balls, how many samples of size 3 can be taken in (i) SRSWOR and (ii) SRSWR.

Methods of Sample Selection - SRS

Random samples can be obtained by any of the following methods

(i) Lottery Method

(ii) Random Number Table Method

Lottery Method

The Simplest method of selecting a simple random sample is the lottery method. Suppose we want to select n candidates out of N. We assign the numbers serially starting from 1 to N. Write these numbers (1 to N) on N slips. These slips are made as homogeneous as possible in shape, size, colour etc. These slips are folded and put in a bag and shuffled thoroughly and then n slips are drawn one by one. The n candidates corresponding to the numbers on the selected slips will constitute a random sample. For example, suppose we have to select five students out of 50 to visit an old age home. We assign numbers from 1 to 50 to the students. 50 identical slips are made for these students. These slips folded and put in a box and shuffle thoroughly. Then five slips are

drawn. Suppose the numbers drawn are 44, 6, 28, 39 and 25. Then the students bearing these numbers are selected for visiting the home.

Random Number Table Method

The limitation of lottery method is that it is quite time consuming if the population is large. The most practical and inexpensive method of selecting a random sample consists of the use of Random Number Tables. The random number table are in such a way constructed that each of the digits 0,1,2,3,4,5,6,7,8,9 appears approximately in the same frequency. The digits are also independent. The method of drawing a random sample by this method consists of the following steps.

Let N be the population Size with k digits and n be the Sample Size to be drawn.

- Identify the N units in the population with the numbers from 1 to N
- Select at random, any page of the table and pickup the successive k digit numbers in any row or column or diagonal at random until we get n number of units.
- Discard numbers which are greater than N.
- The population units corresponding to the numbers selected constitute the random sample

Commonly used Random Number Tables are Tippetts Random Number Table, Fisher and Yates Table, Kendall and Babington Smith Table, Rand Corporation table, C.R. Rao, Mitra and Mathai Table. Random Number Generating Programmes are available in Internet, Computer and Calculator.

Know your progress

Explain the selection procedure of a sample of 20 units from a population containing 80 units using random number table.

Activity

Create similar situations in your class where lottery and random number table method can be applied and solve the same using these methods.

10.7 Systematic Sampling

A sampling method in which one unit is selected at random and the remaining units are selected at an interval of predetermined length is called systematic sampling.

Suppose we want to select a systematic sample of 8 units out of 48 units. To do this we first find the sampling interval $k = \frac{48}{8} = 6$. The first unit in the sample is selected by a random number r between 1 and 6. Let it be 3. Then the third unit will be selected to the sample. There after every sixth unit will be selected automatically into the sample. Hence the resulting systematic sample will contain the units with the following serial numbers

3, 9, 15, 21, 27, 33, 39,45

If the population contains 48 items and a sample of 8 items is to be taken, the selection of every 6^{th} $(48 \div 8)^{th}$ item will give the required sample. The first entry (random start) is determined by selecting a number at random between 1 and 6. If the first item obtained in this manner is 3^{rd} then 9^{th}, 15^{th}, 21^{st}, 27^{th}, 33^{rd}, 39^{th} and 45^{th} items will be picked up. This type of sampling in which n samples are taken out of N units and $k = \frac{N}{n}$. A random start is selected from 1 to k. Let it be i, where $1 \leq i \leq k$ then i^{th} , $(i+k)^{th}$,$(i+2k)^{th}$,... comprising of n items are included in the sample with sampling interval as k. This type of sampling is called Systematic Sampling. These type of sampling is done when a complete list of the population is available. In the above example, $k = 6$, the sampling interval and $i = 3$ is the random start. The pictorial representation of above example is given below

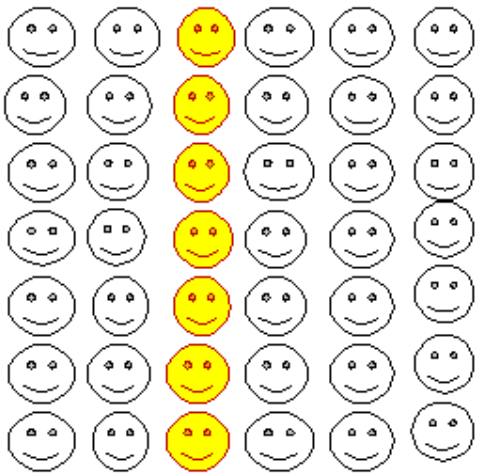


Illustration 10.2

In a class there are 100 students with Roll Numbers from 1 to 100. It is desired to take sample of 10 students. $K = \frac{100}{10} = 10$. *From 1 to 100 roll numbers, the first student between 1 and* k *ie., 1 and 10, will be selected at random and then we will go on taking every* k^{th} *student. Suppose the first student comes out to be* 4^{th}*, the sample would then consist of the following Roll Numbers. 4, 14, 24, 34, 44, 54, 64, 74, 84 and 94.*

10.8 Stratified Random Sampling

Simple Random Sampling is suitable for homogeneous population. When the population is heterogeneous, it is first subdivided into non overlapping exhaustive homogeneous subgroups. These subgroups are called strata . From each stratum, units are selected at random. The number of items taken from each subgroup may be in proportion to its size. This type of sampling is called stratified random sampling. This method is applied so that units within each group are as homogeneous as possible and the group means are as widely different as possible.

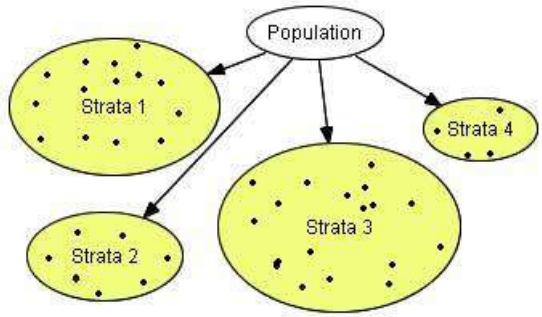


Illustration 10.3

Consider a population which consists of males and females who are smokers or non smokers. The researcher wants to include in the sample, people from all groups-that is, males who smoke, males who do not smoke, female who smoke and female who do not smoke. To accomplish their selection, the researcher divides the population into four subgroups and then selects a random sample from each sub group. This method ensures that the sample is representative on the basis of the characteristics of gender and smoking.

10.9 Cluster Sampling

If we are interested in obtaining the income data in a city, the whole city may be divided into different blocks (clusters) and a Simple Random Sample of required number of blocks is drawn. The individuals of these selected blocks constitute the Cluster Sample. In Cluster Sampling, the total population is divided into some recognizable subdivisions which are termed as clusters and a Simple Random Sample of these clusters is drawn. These clusters are examined completely. This sampling procedure is called Cluster Sampling.

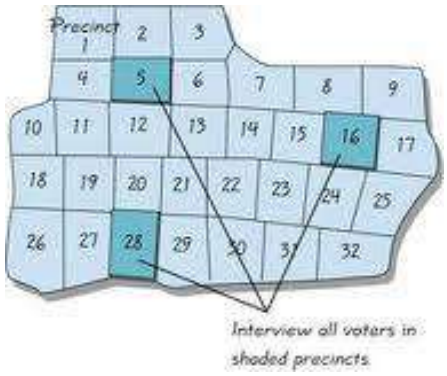


Illustration 10.4

Population	*All school students in the District*
Clusters	*Each school in the district*
Obtain SRS of clusters	*Four schools from the district*
Sample	*Every student in the four schools*

10.10 Multi-Stage Sampling

Selection of a sample of households from a particular State can be done through different stages. The first stage units may be districts, second stage units may be villages in the selected districts and third stage units may be households in the Villages, which are the ultimate units.

State
↓
Districts (First stage)
↓
Villages (Second stage)
↓
Households (Third stage)

Such type of sampling is called Multi-stage Sampling. As the name indicates, multistage sampling refers to a sampling technique which is carried out in various stages. Multi stage sampling consists of sampling first stage units by

some suitable method of sampling. From among the selected first stage units, a sub sample of secondary stage units is drawn by some suitable method of sampling which may be same as or different from the method used in selecting first stage units. Further stages may be added to arrive at a sample of desired sampling units. If the sampling is done only in two stages, it is called Sub-sampling.

Illustration 10.5

Suppose we have to study about the behaviour pattern of marketing of a product in households of a country. Divide the country into different States, States are divided into districts. Districts are divided into cities and towns. These are again divided into wards, and households are selected and study about the marketing of product.

Illustration 10.6

First Stage sampling unit for national surveys are often administrative districts, urban districts or parliamentary constituencies. Within the selected first stage unit one may go direct to the final sampling units, such as individuals, households or addresses, in which case we have a two-stage sample. It would be more usual to introduce intermediate sampling stages, i.e. administrative districts are sub-divided into wards, then polling districts etc.

Probability Sampling Methods and Strategies

Type of Sampling	Selection Strategy
Simple	Each member of the population has an equal probability of being selected.
Systematic	Each member of the population is either assembled or listed, a random start is designated, then members of the population are selected at equal intervals
Stratified	Each member of the population is assigned to a group or stratum, then a simple random sample is selected from each stratum
Cluster	Each member of the population is assigned to a group or cluster, then clusters are selected at random and all members of each selected cluster are included in the sample.
Two stage Sampling	Each member of the population is divided into sub groups, a sample of these groups are selected at random and then a sample of members of each selected subgroups are included in the sample
Multistage sampling	The above stage is extended to multi-levels

Let us sum up

Population is the collection of all observations about which conclusions are to be made. Sample is a part of population. While collecting and processing the data, there may arise two types of errors- sampling error and non-sampling error. Sampling can be done by using non probability or probability sampling. Some of the methods of non probability sampling are convenience sampling, judgment sampling and quota sampling. Some of the methods of probability sampling are simple random sampling - with or without replacement, stratified sampling, systematic sampling, cluster sampling and multistage sampling.

Learning outcomes

After transaction of this unit, the learner:-

- illustrates Census and sampling and their advantages and disadvantages.
- recognises probability and non probability sampling.
- identifies sampling and non sampling errors.
- differentiates SRSWOR and SRSWR, methods of simple random sampling - lottery method and random number table method.
- describes different kinds of sampling - simple random sampling, systematic sampling, stratified random sampling, cluster sampling and multistage sampling.

Evaluation Items

1. Census study involves ———

   (a) 50% subjects of the population

   (b) Each and every subject comprising the population

   (c) Any Number of subjects

   (d) None of the above

2. If a doctor wants to assess the efficacy of a drug on the patients of gastroentitis, then which sampling procedure should he follow?

   (a) Simple random sampling with replacement

   (b) Simple random sampling without replacement

   (c) Judgment sampling

   (d) None of the above

3. From a well shuffled pack of cards, a card is drawn blindly. Its colour is noted and replaced. This process is continued 5 times. This type of sampling is known as

   (a) Sampling with replacement

   (b) Sampling without replacement

   (c) Convenience sampling

(d) Non random sampling

4. Sample study is inevitable because:

   (a) It is not possible to study an infinite population

   (b) It is not possible to test all units of the population if they are perished under observation

   (c) A population study requires too much time and rewources

   (d) All the above

5. There are more chances of Non sampling errors than sampling errors in case of

   (a) Studies of large sample

   (b) Inefficient investigators

   (c) Complete enumeration

   (d) All the above

6. Which type of sampling technique is used in following situations

   (a) Trees in a forest

   (b) Houses in blocks

   (c) Entries in a register which are in serial order

7. Which of the following sampling designs will be categorized as non probability sampling?

   (a) Quota sampling

   (b) Convenience sampling

   (c) Judgment sampling

   (d) All the above

8. Errors other than sampling errors are termed as —

9. The sampling procedure in which population is first divided into homogeneous groups and then a sample is drawn from each group is called—

10. Stratification is appropriate when population is —

11. A sample consists of— of population

12. If a10Cn investigator selects districts from a State, Panchayat Samities from districts and farmers from Panchayat Samities, then such a sampling procedure is known as—

13. Suppose there are 10 students in your class. You want to select 3 out of them. How many distinct samples are possible?

14. Discuss how you would you use the lottery method to select 3 students out of 25 in your class using simple random sampling with replacement and without replacement

15. Explain the procedure for selecting a random sample of 10 students out of 60 in your group by using random number tables

16. Do the errors in sample studies are always greater than that of complete enumeration? Justify your answer

17. A population consists of four numbers 3,7,11 and 15. consider all simple random samples of size 2 that can be drawn (i) with replacement & (ii) without replacement from this population

18. Suggest three situations when sampling is more suitable than census

19. Distinguish between sampling errors and Non sampling errors

20. If a survey is conducted to estimate the crop production in villages and on farms, which type of sampling is preferred?

21. For each of the following sampling plans, suggest methods of sampling plans

(a) A librarian wants to estimate the proportion of the damaged books in the library. He decide to select a book per shelf as sample by measuring 12 inches from the left edge of each shelf and selecting the book in that location

(b) Political surveyors visit 200 houses to collect the details of eligible voters in each house whom they intend to vote for

22. Give three situations where non sampling errors arise.

Answers:

1) b 2) c 3) a 4) d 5) d

6 (a) systematic sampling (b) cluster sampling (c) systematic sampling

7) d 8) non sampling errors 9) stratified random sampling 10) hetrogeneous

11) representative part 12) multi-stage sampling 13 $10C_3$

Glossary

Arithmetic Mean	:	The sum of observations divided by the number of observations.
Bar Diagram	:	Diagrammatic representation of data using bars proportional to the frequencies
Bivariate Frequency Table	:	Frequency distribution of a bivariate data in rows and columns
Box plot	:	Box plot is the graphical representation of data based on its quartiles. It is also known as box and whisker plot.
Census	:	Data collected from each and every unit of the populations
Central Tendency	:	Tendency of the observations in a data to cluster around a central value is called central tendency.
Chronological Classification	:	Classification based on time.
Classification	:	Arrangement of items according to some attributes
Cluster Sampling	:	Choosing a cluster of items as a unit.
Coefficient of QD	:	Relative measure of dispersion based on Quartiles.
Coefficient of variation	:	Relative measure of dispersion based on standard deviation.
Combined mean	:	The mean of a combined group two or more sets taken together.
Conditional probability	:	Probability of an event conditioned by another event

Continuous variables	:	Variables take any values within a specified range.
Covariance	:	It indicates strength of linear relationship between two variables.
CSO	:	Central Statistical Office
Cumulative Frequency	:	Number of observations less than or greater than a particular value.
Cumulative Frequency Table	:	Tabular representation of cumulative frequencies
Data	:	Any measurement, result, fact or observation which gives information.
Deciles	:	The values of a data which divide the distribution into ten equal parts are called deciles.
Dichotomy	:	Classification into two disjoint groups
Discrete variable	:	Variables take countable number of values.
Enumeration	:	The process of data collection by enumerator.
Enumerator	:	The person deputed by the investigator to collect data from field.
Equally likely events	:	Two or more events having an equal chance of occurrence.
Event	:	Subset of a sample space.
Frequency	:	Number of repetitions of an observation
Frequency curve	:	Joining the points of a frequency polygon by a freehand smoothed curve
Frequency polygon	:	Graphical device for understanding the shapes of distributions.
Frequency Table	:	Tabular representation of frequencies

Geographical Classification	:	Classification based on location.
Geometric Mean	:	GM is the n th root of the product of n observations in a data.
Harmonic Mean	:	HM of a number of observations is the reciprocal of the AM of the reciprocals of the observations.
Histogram	:	Graphical representation of frequency distribution using adjacent vertical bars.
Independent events	:	Occurrence of one event does not affect the occurrence of the other
Investigator	:	The person authorised to make investigation.
ISI	:	Indian Statistical Institution
Kurtosis	:	Measure of Peakedness
Lepto Kurtic	:	Highly peaked curve.
Manifold Classification	:	Classification by considering more than one attribute at a time
Mean deviation	:	Arithmetic mean of the absolute deviations of observations from their average
Median	:	The middlemost observation in the data which divides the distribution into two equal parts, when the data is arranged in ascending or descending order.
Meso Kurtic	:	Curve which is moderately peaked.
Mode	:	Mode of a data is the value that is repeated most often in the data.
Moments	:	Represent a convenient and unifying method of summarising certain descriptive statistical measures

Multistage Sampling	:	Sampling in various stages.
Mutually exclusive events	:	Events which cannot occur together.
Non Sampling Error	:	Errors other than sampling error.
NSSO	:	National Sample Survey Office
Ogives	:	Curves obtained by plotting cumulative frequencies.
Percentage Frequency	:	Frequency in terms of the percentage of the total frequency
Percentage Frequency Table	:	Tabular representation of percentage frequencies
Percentiles	:	The values of a data which divide the distribution into hundred equal parts are called deciles.
Pie diagram	:	Circle divided in to various segments proportional to the frequencies.
Platy Kurtic	:	Curve which is flat topped.
Population	:	All elements whose characteristics are being studied
Probability	:	A numerical measure of the possibility of an event.
Probability Sampling	:	All units have specified probability of being included in the sample
Qualitative Classification	:	Classification based on the quality
Qualitative data	:	Data which can be observed but cannot be numerically measured.
Quantitative Classification	:	Classification based on quantity
Quantitative data	:	Variables which can be numerically measured.
Quartile deviation	:	Half of the difference between third quartile and first quartile
Quartiles	:	The values of a data which divide the distribution into four equal parts are called deciles.

Random experiment	:	Experiment having more than one possible result.
Range	:	Difference between the highest and lowest values.
Relative Frequency	:	Ratio of frequency to the total frequency
Relative Frequency Table	:	Tabular representation of relative frequencies
Sample	:	Representative part of the population
Sample space	:	The set of all possible outcomes of a random experiment.
Sampling	:	Studying the population by using samples.
Sampling Error	:	Errors due to sampling.
Scatter plots	:	Diagrammatic representation of bivariate data
Simple event	:	The basic possible outcome of a random experiment.
Simple Random Sampling	:	All units have equal chance or probability of being included in the sample
Skewness	:	Lack of symmetry
Standard deviation	:	Positive square root of the arithmetic mean of the squares ofdeviations of the observations from their arithmetic mean
Statistical Investigation	:	Collection, organization, analysis and interpretation of data according to well defined procedure.
Stratified Sampling	:	Sampling by dividing the population into strata.
Symmetric distribution	:	Data distributed equally on either sides of the mode.

Systematic Sampling	:	Sampling by systematic manner.
Tabulation	:	Presentation of data in rows and columns.
Variance	:	Square of standard deviation
Weighted AM	:	The AM that assign a weight to each observation on its importance related to other is called weighted AM.

References

1. Principles of Statistics, Dr. S M Shukla and Dr.Sahai, Sahitya Bhavan Publications, Delhi
2. Mathematics and Statistics for Economics, G.S Monga, Vikas Publishing House pvt ltd
3. Fundamentals of Mathematical Statistics, S.C Gupta & V K Kapoor, Sultan Chand & sons Educational Publishers.
4. Fundamentals of Statistics, D N Elhance, Veena Elhance & B.L. Agarwal, Kitab Mahal Publishers.
5. Statistical Methods, S. P. Gupta, Sultan Chand & Sons, New Delhi.
6. Applied General Statistics, Frederick E Croxton, Dudley J Cowden, Sidney Klein, Prentice Hall India.
7. Fundamentals of Mathematical Statistics: S C Gupta, V K Kapoor , Sulthan Chand & sons, New Delhi
8. Business Statistics , Naval Bajpai , Pearson Educational Publications
9. Practical Statistics ,R.S.N Pillai & Bagavathi
10. Programmed Statistics, B L Agarwal, New Age Publishers, Delhi
11. Elementary Statistical Methods, S.P. Gupta ,Sultan Chand & sons Publishing co.
12. Introduction to Statistics, R.P.Hooda
13. Elementary Statistics A step by step Approach, Allan G Bluman, McGraw Hill Publishers.
14. Elementary Statistics and Indian Economic development -T.R. Jain,V.K.Ohri, VK Publishers Delhi.

15. Statistics for Management and Economics, Gerald Keller & Brian Warrack, Eastern Economy Edition

16. Statistics for Management, Richard.J.Levin & David S Rubin, Eastern Economy Edition

17. Statistics, David Freedman, Robert Pisani & Roger Purves,w.w. Norton & Company Inc , Viva Books Pvt Ltd, Delhi

18. Probability and Statistics for Engineers,G S S Bhishma Rao, SCITECH Publishers.

19. Schaums Outlines, Statistics Murray R Spiegal & Larry J Stephens, Metric Editions,Schaums Publishing Company, New York

20. Head First Statistics, Dawn Griffiths, Shroft Publishers and Distributors Pv. Ltd.

21. Statistics an Introduction, Robert D Mason , Douglas A Lind, and William G Marchal, Harcourt Brace Jovanovich Inc.

.1 Appendix-1

Many softwares are available in the market for statistical data analysis. Spreadsheet, SPSS, Statistica, Minitab, R etc.. are some examples. Among these R is free software programming language and software environment for statistical data analysis and graphics. Some of the R – Codes are given below.

Frequency Distribution of Qualitative Data

Example 1

Following are school types of 54 schools in an education district.

A C CC H HH C C D DDDDD E A AA B BBB C E EEEE G G A AB A A

D D E F FFF G D D A A B G GGG H

Obtain the frequency distribution of school types.

R Codes

```
>school=c(”A”,”C”, ”C”, ”C”, ”H”, ”H”, ”H”, ”C”,
”C”, ”D”, ”D”, ”D”, ”D”, ”D”, ”D”, ”E”, ”A”, ”A”,
”A”, ”B”, ”B”, ”B”, ”B”, ”C”, ”E”, ”E”, ”E”,
”E”,”E”,”G”, ”G”, ”A”, ”A”, ”B”, ”A”, ”A”, ”D”,
”D”,”E”, ”F”, ”F”, ”F”, ”F”, ”G”, ”D”,”D”, ”A”,
”A”, ”B”, ”G”, ”G”, ”G”, ”G”, ”H”) # Enter the raw
data as a vector
>table(school) # Prepare the frequency
distribution for school #type
```

Output:

School	A	B	C	D	E	F	G	H
	10	6	6	10	7	4	7	4

Remark: To get a fancy output one can use the R Codes:

```
>mytable=table(school) # Prepare the frequency
table

>cbind(mytable) Prepare a fancy frequency
table(try!)
```

Bar Diagrams

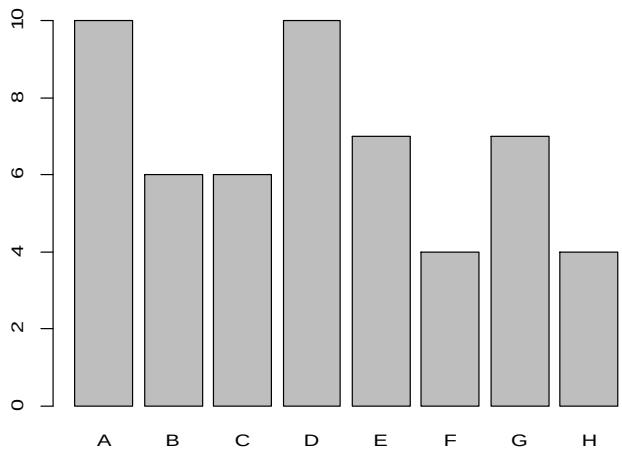
Example 2(Simple Bar Diagram)

Draw the bar graph of the variable school types in example 1.

R Codes

```
>school
>mytable=table(school)
>barplot(mytable)
```

Output:



Remark: To get a fancy output one can use the R Codes:

```
>mytable=table(school)

>barplot(mytable,xlab="School Type",ylab="No.of
Schools", main="BAR DIAGRAM",
col=c("red", "yellow", "green","violet",
```

```
"orange","blue", "pink", "cyan")) # Try!
```

Example 3(Multiple and Subdivided Bar diagrams)

Following table shows the number of students admitted in different faculties in a university indifferent years:

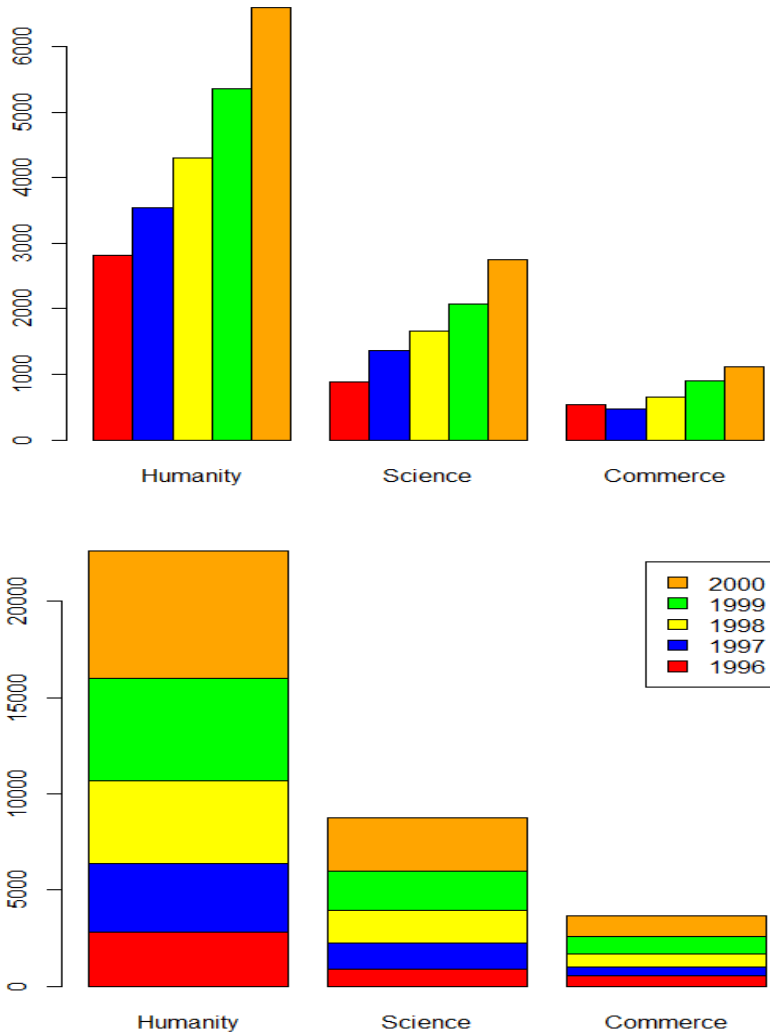
Sl.Number	Year	Humanity	Science	Commerce
1	1996	2810	890	540
2	1997	3542	1363	471
3	1998	4301	1662	652
4	1999	5362	2071	895
5	2000	6593	2752	1113

R Codes

```
>no.stud=matrix(c(2810,890,540,3542,1363,471,4301,
1662,652,5362,2071,895,6593,2752,1113), byrow=T,
ncol=3)

# Entries( that is number of students) into matrix
form; enter #values by row and number of columns
is 3;

>rownames(no.stud)=c("1996","1997","1998","1999","
2000")

# define row names

>colnames(no.stud)=c("Humanity","Science","Commerc
e")

 # define column names

>no.stud     # Print table in the matrix form

>barplot(no.stud,
col=c("red","blue","yellow","green","orange"),

legend = rownames(no.stud))
```

```
# Gives subdivided bar diagram with legends as row
names.

>barplot(no.stud,beside=T,col=c("red","blue","yell
ow","green","orange")) # Gives multiple bar
diagram in which bars side by #side.
```

Outputs



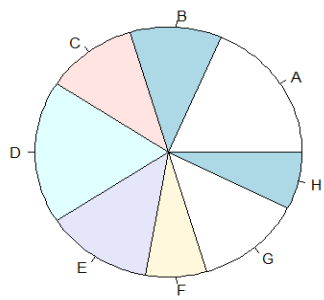
Pie Diagram

Example 4(For qualitativedata)

Draw Pie diagram for the school types in example 1.

```
>school #print the values of the qualitative
variable ‘school’

>mytable=table(school)

>pie(mytable)   # pie diagram for the frequency
table of #school.

>pie(mytable,label=c(“A”,”B”,”C”,”D”,”E”,”F”,”G”,”
H”))
```

Output



Example 5 (For quantitative data)

The salesof an appliance in 5 cities in October, 2013 is given in lakhs(Rs.) is given bellow:

Kannur	Kozhikode	Cochi	Kollam	Thiruvanathapuram
78.5	98.75	135,75	65.5	82.45

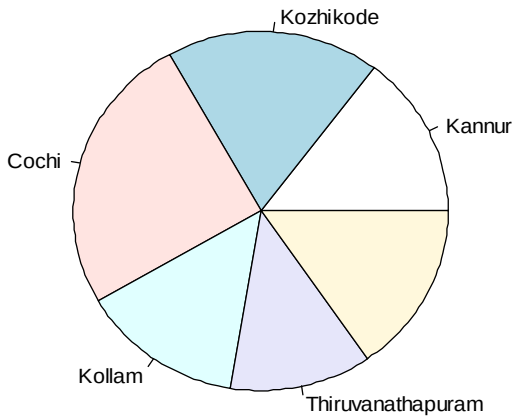
Draw the pie diagram for the sales data

Answer

R Codes

```
>sales.data=c(78.5,  98.75,  135,75,  65.5, 82.45)
>names=c("Kannur"," Kozhikode"," Cochi", "Kollam"
, "Thiruvanathapuram")
>pie(sales.data,label=names)
```

Output



Frequency Distribution of Quantitative variables

Example6 : The daily numbers computer stoppages are observed over 30 days at a school computing center. Prepare the frequency distribution and draw the line diagram for the following data.

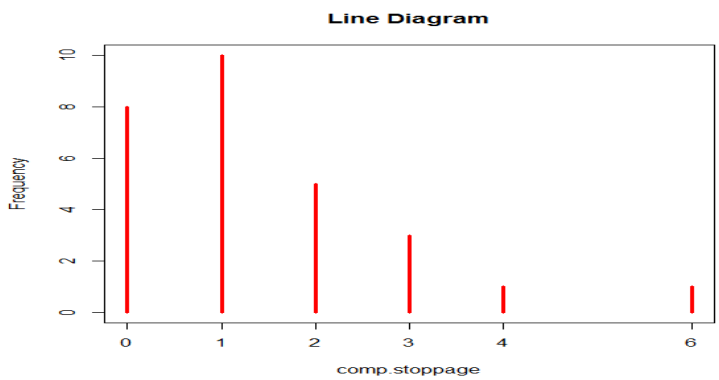
Daily Numbers of Computer Stoppages

3	1	1	0	1	0	1	1	0	2	2
0	0	0	1	2	1	2	0	0	1	6
4	3	3	1	2	1					

R Codes

```
>comp.stoppage=c(3,1,1,0,1,0,1,1,0,2,2,0,0,0,1,2,1
,2,0,0,1,6,4,3,3,1,2,1)

# Enter the raw data as a vector

> stoppage=table(comp.stoppage)

#Prepare the frequency table(Try the table
output!)

>plot(stoppage, type = "h", col = "red",
lwd=3,ylab=Frequency main="Line Diagram")

#Draw the line diagram;type=”h” adds vertical
lines and lwd=3 #decides the thickness of vertical
lines.
```

Output



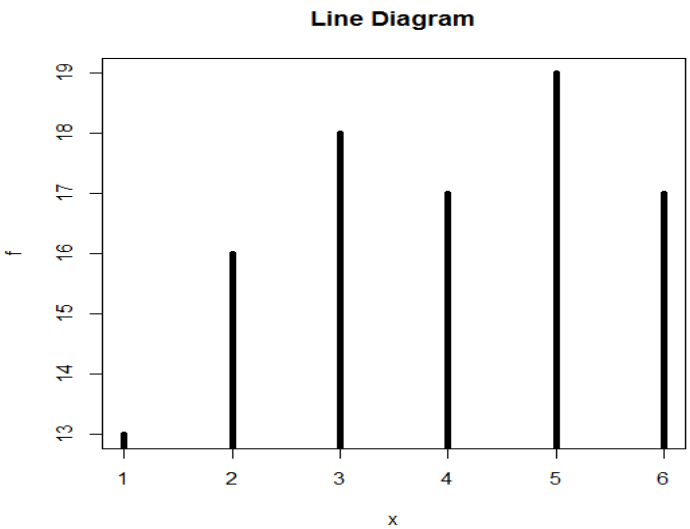
Example7.(When values and frequencies are directly given)

The number (X) obtained when a die is tossed 100 times. is tabulates as

X:	1	2	3	4	5	6
f:	13	16	18	17	19	17

R Codes

```
>x=c(1,2,3,4,5,6)

>f=c(13,16,18,17,19,17)

>plot(x,f,type="h",lwd=4,xlab="x",ylab="f",main="L
ine Diagram")
```



Histogram

Example 8

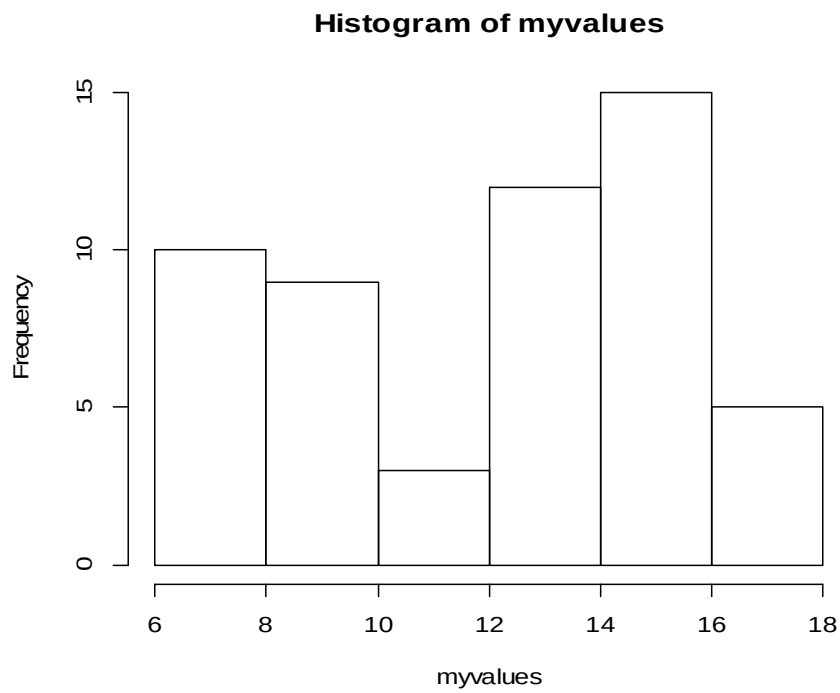
Obtain the Histogram for the data

8	16	13	16	15	16	17	16	12	18	13	15
	15	14	14	15	15	14	10	12	15	15	8
	6	9	8	9	6	14	14	15	10	14	6
	13	17	10	13	17	10	10	15	6	10	8
	14	6	13	12	10	8	16	15	17		

R Codes

```
>myvalues=c(8,16, 13,16,  15, 16,  17, 16,  12, 18,  13,
15,  15, 14,  14, 15,  15, 14,  10, 12,  15,  15,  8,  6,
9,  8,  9,  6, 14, 14,  15, 10, 14,  6,  13, 17, 10, 13,  17, 10,  10,
15,  6, 10,  8, 14,  6,  13, 12,  10,  8, 16, 15, 17)
```

```
>hist(myvalues) # Default output; break points set
automatically
```



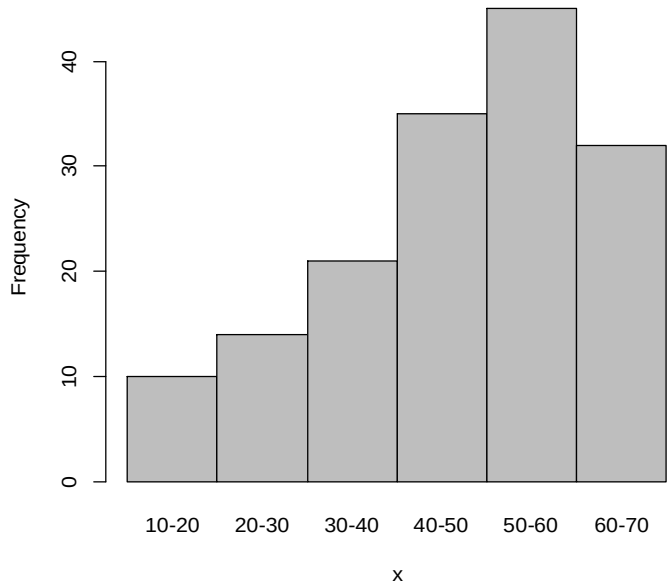

```
>hist(myvalues,breaks=c(5,6.5,8,9.5,11,12.5,14,15.
5,17,18.5),col="red") # Fancy; break points are
manualy set.(Try!)
```

Example 8(a) Draw the histograms for the following frequency distributions

1. Class: 10-20 20-30 30-40 40-50 50-60 60-70
   Freq.: 10 14 21 35 45 32

R Codes

```
>x=c("10-20","20-30","30-40","40-50","50-
60","60-70")
> f=c(10,14,21,35,45,32)
>
barplot(f,names=x,space=0,xlab=”x”,ylab=”Frequ
ency”) # It #is a trick of drwing bardiagram
with 0 space between bars.
```



2. Midvalue: 5 10 15 20 25 30
   Frequency: 1 8 14 21 18 32

R Codes

```
>x=c("5","10","15","20","25","30")
>f=c(1,8,14,21,18,32)
>barplot(f,names=x,space=0)# Try!
```

Probability Curve (probability Density Plot)

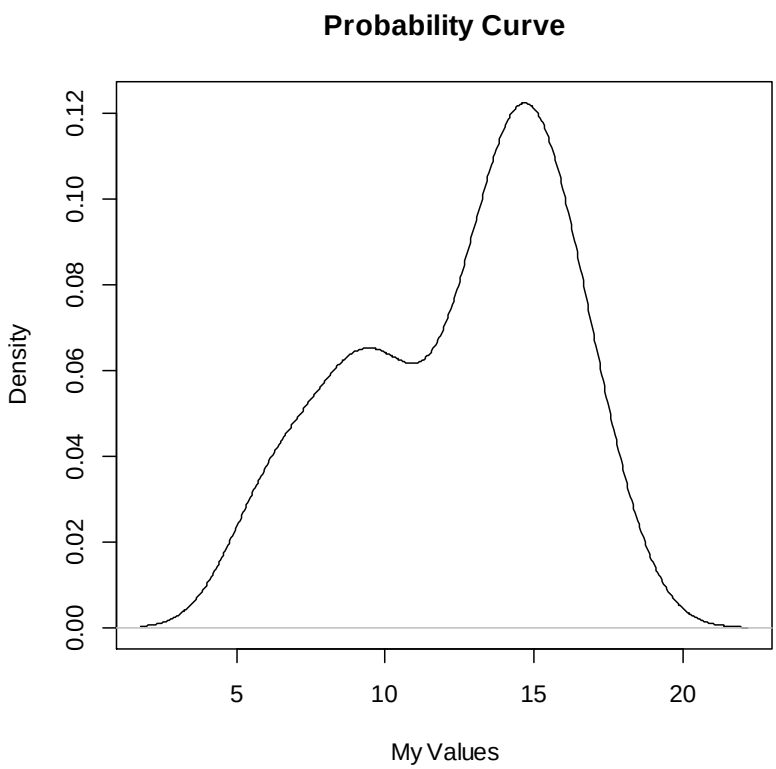
Example 9 Draw the probability curve for the data given in example 8

R Codes

```
>myvalues=c(8,16, 13,16,  15, 16,  17, 16,  12, 18,  13,
15,  15, 14,  14, 15,  15, 14,  10, 12,  15,  15,  8,  6,
9,  8,  9,  6, 14, 14,  15, 10, 14,  6,  13, 17, 10, 13,  17, 10,  10,
15,  6, 10,  8, 14,  6,  13, 12,  10,  8, 16, 15, 17)
```

```
>plot(density(myvalues))# default output(Try!)
```

```
>plot(density(myvalues),xlab="My
Values",main="Probability Curve")# Output shown
```



Scattar Diagram

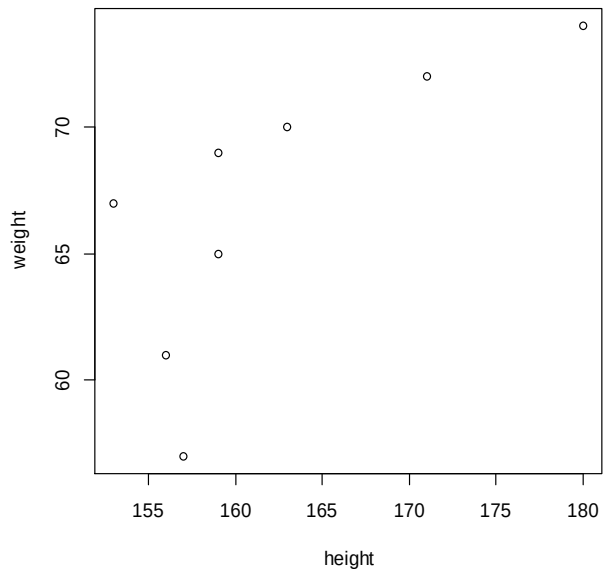
Example10 Draw Ascattar diagram to the bivariate data given bellow and comment on the plot.

Height(Cm): 157 159 163 156 171 180 153 159

Weight(Kg): 57 65 70 61 72 74 67 69

R Codes/Outputs

```
>height=c(157,159,163,156,171,180,153,159)
>weight=c(57,65,70,61,72,74,67,69)
>plot(height,weight)
```



Descriptive Statistics

Example10 Find the common measures of central tendencies and dispersions for the data in example 9(myvalues)

R Codes/Outputs

```
>mean(myvalues) # Arithematic Mean
[1] 12.46296
>median(myvalues) # Median
[1] 13.5
>sd(myvalues) # Standard Deviation
[1] 3.457084
var(myvalues)
[1] 11.95143
```

Quantiles(Quartiles,Deciles,Percentilesetc)

Example 11 Find the three quartiles, third and seventh deciles and 14th, 23rd and 72th percentiles for the data in example 9(myvalues)

R Codes/Outputs

```
>quantile(myvalues) # Default is minimum, Q1, Q2,
Q3 and maximum

     0%     25%     50%       75%       100%
     6.0    10.0    13.5      15.0      18.0

>quantile(myvalues,probs=c(.3,.7,.14,.23,.72))#
deciles and #percentiles of specified order.

     30%    70%     14%       23%       72%
     10     15      8         10        15
```

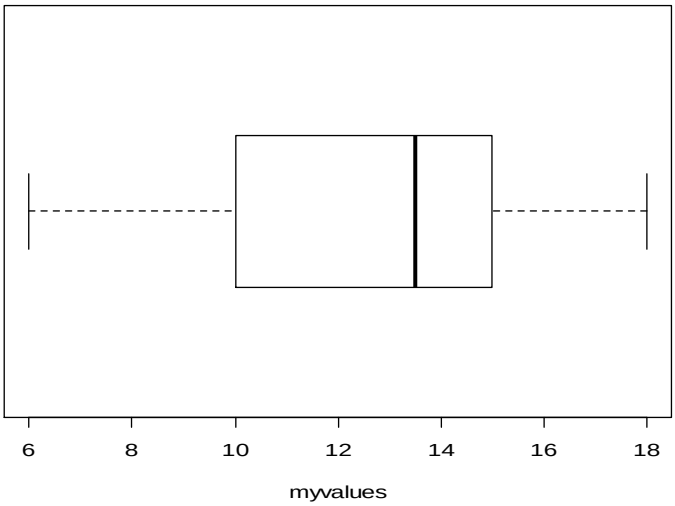
Box and Whisker Plot

Example 12 Draw the box and whisker plot for the data in example 9(myvalues)

R Codes

```
>boxplot(myvalues,horizontal=T,xlab="myvalues")
```

Output



Try the code: `>boxplot(myvalues)`

Range and inter Quartile Range

Example 13 Find the range and inter quartile range for the data in example 9(myvalues)

R Codes/Outputs

```
>range(myvalues) # Wrong code!! See the output
[1]  6 18
>range=max(myvalues)-min(myvalues)
>range
[1] 12
>IQR(myvalues)# Inter Quartile Range; IQR=Q3-Q1
[1] 5
```

.2 Appendix-2

CENSUS- 2011

POPULATION AND DECADAL GROWTH RATE

As per 2011 Provisional Population Figures, Rural Population in Kerala is 17,455,506. Out of this 8,403,706 are males and 9,051,800 are females whereas urban population in this state is 15,932,171. Out of this, 7,617,584 are males and 8,314,587 are females. The decadal decline of rural population was –25.96%, whereas the urban population has grown by 92.72%.

TRENDS IN RURAL AND URBAN CLASSIFICATION OF POPULATION IN KERALA

The State has now 52.30 percent rural population in 2011 Census as against 74.04 per cent in 2001 Census. The huge growth in urban population during the past decade 2001-2011(92.72 %) could be attributed squarely to the manifold increase in number of Towns in the State between 2001 & 2011 from 159 to 520. 47.72% of the total population of Kerala are from Urban. Ernakulam is the most urbanised district (68.07%) and Wayanad (3.87%) is the least urbanised district of the State.

POPULATION IN THE AGE GROUP 0-6 AGE

Total Population in the age group 0-6 is 3,322,247. Out of this males are 1,695,935 and females are 1,626,312. Rural Population in the age group 0- 6 is 1,747,512. Males are 891,668 and females are 855,844. Urban Population of the age group 0- 6 is 1,574,735. Out of this 804,267 are males and 770,468 are females. Percentage of rural population in the age group 0- 6 to the total rural population is 10.01 and that of urban is 9.88

SEX RATIO (FEMALES PER 1000 MALES)

The Overall sex ratio of Kerala is 1084 females per 1000 males. Whereas, sex ratio of rural area is 1077 and that of urban area is 1091.

CHILD SEX RATIO (0-6 YEARS)

Child sex ratio in respect of 0-6 age population in Kerala is 959. In rural areas it is 960, whereas, sex ratio of 0-6 age population in urban areas is 958.

LITERACY

Total number of literates in Kerala is 28,234,227and total literacy rate is 93.91%. Among these, literates in Rural area is 14,595,727 and that in Urban area is 13,638,500 .The numbers of male literates in Rural area is 7,158,427 and the number of male literates in Urban area is 6,597,461. Female literates in rural areas are 7,437,300 and that in urban area is 7,041,039. Literacy rate in the rural area is 92.92% and that of urban area is 94.99 %. The gender gap in literacy in rural area of the State is found to be 4.55%; whereas that in urban area is 3.5%

Figures at a glance and Data Sheet showing district level break up of Rural & Urban of the state is attached.

(See overleaf)

Table-1

Distrubution of Population, Decadal Growth Rate, Sex-Ratio and Population Density

State/ District Code	State/ District	Population 2011			Percentage decadal growth rate of population		Sex- Ratio* (Number of Females per 1000 Males)		Population density per sq. km.	
		Persons	Males	Females	1991-01	2001-11	2001	2011	2001	2011
1	2	3	4	5	6	7	8	9	10	11
32	**Kerala**	**3,33,87,677**	**1,60,21,290**	**1,73,66,387**	+9.43	+4.86	1058	1084	819	859
01	Kasaragod	13,02,600	6,26,617	6,75,983	+12.37	+8.18	1047	1079	604	654
02	Kannur	25,25,637	11,84,012	13,41,625	+6.98	+4.84	1090	1133	812	852
03	Wayanad	8,16,558	4,01,314	4,15,244	+16.14	+4.60	995	1035	366	383
04	Kozhikode	30,89,543	14,73,028	16,16,515	+9.89	+7.31	1057	1097	1228	1318
05	Malappuram	41,10,956	19,61,014	21,49,942	+17.09	+13.39	1066	1096	1021	1158
06	Palakkad	28,10,892	13,60,067	14,50,825	+9.88	+7.39	1066	1067	584	627
07	Thrissur	31,10,327	14,74,665	16,35,662	+8.66	+4.58	1092	1109	981	1026
08	Ernakulam	32,79,860	16,17,602	16,62,258	+9.35	+5.60	1019	1028	1012	1069
09	Idukki	11,07,453	5,51,944	5,55,509	+7.03	-1.93	993	1006	259	254
10	Kottayam	19,79,384	9,70,140	10,09,244	+6.86	+1.32	1025	1040	885	896
11	Alappuzha	21,21,943	10,10,252	11,11,691	+5.39	+0.61	1079	1100	1492	1501
12	Pathanamthitta	11,95,537	5,61,620	6,33,917	+3.84	-3.12	1094	1129	468	453
13	Kollam	26,29,703	12,44,815	13,84,888	+7.38	+1.72	1069	1113	1038	1056
14	Thiruvananthapuram	33,07,284	15,84,200	17,23,084	+9.76	+2.25	1060	1088	1476	1509

**For calculation of sex ratio, others have been considered as males.*

.4 Appendix-4

Table-3

Sex-Ratio for State and Districts : 1901-2011

State/ District Code	State/ District	*Sex-ratio (Number of females per 1000 males)											
		1901	1911	1921	1931	1941	1951	1961	1971	1981	1991	2001	2011
1	2	3	4	5	6	7	8	9	10	11	12	13	14
32	**Kerala**	**1004**	**1008**	**1011**	**1022**	**1027**	**1028**	**1022**	**1016**	**1032**	**1036**	**1058**	**1084**
01	Kasaragod	1060	1053	1050	1040	1039	1046	1026	998	1020	1026	1047	1079
02	Kannur	1060	1079	1121	1106	1110	1074	1048	1033	1040	1049	1090	1133
03	Wayanad	805	815	786	804	835	838	903	922	949	966	995	1035
04	Kozhikode	1009	1022	1038	1032	1044	1019	1007	1004	1020	1027	1057	1097
05	Malappuram	1017	1020	1037	1059	1062	1055	1057	1041	1052	1053	1066	1096
06	Palakkad	1042	1057	1069	1079	1079	1085	1077	1056	1056	1061	1066	1067
07	Thrissur	1004	1009	1051	1075	1082	1105	1093	1081	1100	1085	1092	1109
08	Ernakulam	985	990	969	994	994	1008	999	988	997	1000	1019	1028
09	Idukki	839	842	850	834	875	909	914	937	963	975	993	1006
10	Kottayam	965	969	947	966	966	987	988	991	1001	1003	1025	1040
11	Alappuzha	986	987	986	997	1003	1022	1026	1025	1043	1051	1079	1100
12	Pathanamthitta	986	987	949	975	986	996	1011	1019	1056	1062	1094	1129
13	Kollam	987	988	989	1006	1013	997	996	1000	1022	1035	1069	1113
14	Thiruvananthapuram	996	990	981	1003	1017	1010	1005	1008	1030	1036	1060	1088

**For calculation of sex ratio, others have been considered as males.*

Table-4

Population in the Age-Group 0-6, Number of Literates and Literacy Rate for State and Districts : 2011

State/ District Code	State/ District	Total Population			Population in age group 0-6			Number of literates			Literacy rate#		
		P	M	F	P	M	F	P	M	F	P	M	F
1	2	3	4	5	6	7	8	9	10	11	12	13	14
32	**Kerala**	**3,33,87,677**	**1,60,21,290**	**1,73,66,387**	**33,22,247**	**16,95,935**	**16,26,312**	**2,82,34,227**	**1,37,55,888**	**1,44,78,339**	**93.91**	**96.02**	**91.98**
01	Kasaragod	13,02,600	6,26,617	6,75,983	1,49,280	76,149	73,131	10,36,289	5,17,031	5,19,258	89.85	93.93	86.13
02	Kannur	25,25,637	11,84,012	13,41,625	2,65,276	1,35,189	1,30,087	21,56,575	10,22,972	11,33,603	95.41	97.54	93.57
03	Wayanad	8,16,558	4,01,314	4,15,244	89,720	45,776	43,944	6,49,186	3,30,093	3,19,093	89.32	92.84	85.94
04	Kozhikode	30,89,543	14,73,028	16,16,515	3,23,511	1,64,800	1,58,711	26,34,493	12,76,384	13,58,109	95.24	97.57	93.16
05	Malappuram	41,10,956	19,61,014	21,49,942	5,52,771	2,81,958	2,70,813	33,28,658	16,08,229	17,20,429	93.55	95.78	91.55
06	Palakkad	28,10,892	13,60,067	14,50,825	2,88,366	1,46,947	1,41,419	22,32,190	11,19,360	11,12,830	88.49	92.27	84.99
07	Thrissur	31,10,327	14,74,665	16,35,662	2,89,126	1,48,428	1,40,698	26,89,229	12,86,141	14,03,088	95.32	96.98	93.85
08	Ernakulam	32,79,860	16,17,602	16,62,258	2,89,281	1,48,047	1,41,234	28,61,509	14,27,572	14,33,937	95.68	97.14	94.27
09	Idukki	11,07,453	5,51,944	5,55,509	1,00,107	51,132	48,975	9,28,774	4,74,988	4,53,786	92.20	94.84	89.59
10	Kottayam	19,79,384	9,70,140	10,09,244	1,68,563	86,113	82,450	17,45,694	8,59,038	8,86,656	96.40	97.17	95.67
11	Alappuzha	21,21,943	10,10,252	11,11,691	1,86,022	95,556	90,466	18,63,558	8,95,476	9,68,082	96.26	97.90	94.80
12	Pathanamthitta	11,95,537	5,61,620	6,33,917	91,501	46,582	44,919	10,70,120	5,03,171	5,66,949	96.93	97.70	96.26
13	Kollam	26,29,703	12,44,815	13,84,888	2,38,062	1,21,481	1,16,581	22,42,757	10,76,509	11,66,248	93.77	95.83	91.95
14	Thiruvananthapuram	33,07,284	15,84,200	17,23,084	2,90,661	1,47,777	1,42,884	27,95,195	13,58,924	14,36,271	92.66	94.60	90.89

#Literacy rate is the percentage of literates to total population aged 7 years and above.

.6 Appendix-6

Random Number Table

4139 8518 2724 7450 8584 7837 0080 0695 3538 7663 7943 9759 1656 0654 6315 9332
2938 6809 0507 9413 7383 3472 5207 5313 4993 5954 3070 1722 3111 8945 1442 1295
4393 5100 5634 4032 8838 1763 0334 7277 3792 4245 8197 3685 4566 7236 9225 3258
5848 3391 0761 5995 0293 0054 5461 1895 5247 2536 3324 8304 3365 5527 4352 7877
7302 1682 5888 0614 9092 8345 0588 3859 6702 0827 8452 0267 4820 3818 9479 9840
6102 9973 1015 2577 0547 6636 8371 8477 8157 9118 3579 4886 6275 2109 4606 4459
7557 8264 8798 7196 2002 4927 3498 0441 6956 7409 1361 6849 7730 0400 9733 6422
9011 3899 3925 9159 0801 3218 8625 2404 8411 3044 6488 1468 6529 8691 7516 1041
7811 2190 9052 1122 2256 1509 3752 7022 9866 1335 1615 3431 7984 6982 2643 3004
9266 0481 4179 5741 3711 9800 8879 8986 8665 9626 6743 8050 9439 2617 7770 7623
0721 8772 9306 7704 2510 8091 4006 3604 0120 7917 1870 0013 8238 0908 2897 9586
9520 7063 7089 2323 3965 6382 1789 5568 1575 6208 9652 1976 9693 9199 8024 1549
0975 5354 2216 4286 5420 4673 6916 0186 0374 4499 4779 6595 1148 7490 3151 6168
2430 3645 7343 8905 6875 2964 2043 2150 1829 2790 9906 8558 9947 5781 0934 8131
1229 1936 2470 0868 5674 1255 7170 4113 3284 1081 5033 3177 1402 4072 6061 2750
2684 0227 7597 2831 7129 9546 2297 8732 2083 9372 0161 5140 2857 2363 1188 4713
6756 8144 3979 6476 6328 9426 0133 0667 9065 3872 6796 5367 2310 8826 9279 3231
8719 9599 2270 1603 8292 0881 5769 5794 1028 2671 5087 0494 4273 0281 4914 8358
3338 8398 0561 9386 2910 9680 4059 0921 2991 4126 3378 5621 8892 1736 3205 3485
5301 9853 8852 4513 4874 1135 2350 6048 7610 5581 1669 0748 0855 0535 1496 8612
9920 1308 4487 9640 9492 2590 0641 1176 9573 4380 9960 5875 5474 1990 9787 3739
1883 0107 2778 4767 1456 1389 8932 8958 4192 5835 8251 3658 7437 3445 8078 1522
3846 1562 1069 9894 3419 2844 7223 4085 6155 7290 6542 8785 2056 2244 6369 6649
8465 3017 9360 5021 8037 4299 5514 9212 0774 8745 4833 3912 4019 3699 4660 1776
0428 1816 7651 2804 0001 3098 3805 4339 2737 7544 3124 9039 5982 5154 2951 6903
5047 3271 5942 7931 4619 4553 2096 9467 4700 8999 1415 4166 0601 3953 1242 2030
7010 4726 4233 3058 6583 6008 0387 4594 9319 0454 9706 1949 2564 5408 9533 9813
1629 3525 2524 8185 1201 4807 8678 2376 1282 9253 5341 7076 7183 6863 7824 4940
3592 4980 0815 3312 3165 6262 6969 7503 5901 0708 3632 2203 9146 5662 6115 0067
5555 6435 9106 1095 5128 7717 5260 2630 7864 2163 1923 7330 3765 7117 4406 5194
0174 5234 7397 6222 9747 9172 3551 7758 2483 0962 0214 2457 5728 8572 2697 0321
2137 6689 5688 1349 1710 7971 1842 2885 4446 2417 8505 0240 0347 0026 0988 5448
7308 3996 3930 1473 8843 4077 8376 0792 3543 9978 3330 0619 1407 6387 4103 3609
2435 5959 5385 9764 3970 8696 9831 9083 1326 1941 4784 8910 9190 1006 2902 1900
7562 0578 4184 8055 1753 0659 8630 4718 6453 6560 6239 7201 4317 2969 4357 0191
2689 2541 5639 6346 6880 5278 0085 3009 1580 8523 5039 5492 9444 4932 5812 8482
0471 4504 7094 4637 2007 7241 1540 1300 6707 3142 6493 3783 4571 9551 4611 6773