

SYLLABUS

Statistics and Probability

EG 2207 SH

Year: II
Semester: II

Total: 4 hours/week
Lecture: 3 hours/week
Tutorial: 1 hours/week
Practical: hours/week

Course Description

This course deals with a practical knowledge of the principles and concept of probability and statistics and their application to simple engineering problems.

Course Objectives

After completing this course the students will be able to:

1. understand the principles and concept of probability to simple engineering problems
2. understand statistics and their application to simple engineering problems

Course Contents

Unit 1. Introduction	[4]
1.1. Origin and Definition of Statistics	
1.2. Importance and Scope of Statistics	
1.3. Limitation of statistics	
Unit 2. Basic Concept of Statistical Studies	[6]
2.1 Data, Type and Sources of Data	
2.2 Population and Sample	
2.3 Variables and Parameter	
Unit 3. Organizing a Raw Data	[6]
3.1. Classification of Data	
3.2. Meaning and Importance of Table	
3.3. Parts of Table	
Unit 4. Pictorial Representation of a Data Set	[5]
4.1. Introduction	

4.2.	Difference between Diagram and Graphs	
4.3.	Bar diagram, Histogram, Pie diagram, Steam leaf display	
4.4.	Graphical Representation of Data	
4.5.	Limitation of Diagram and Graphs	
Unit 5.	Summarizing a Data Set	8
5.1.	Introduction	
5.2.	Central Tendency (mean, median, and mode)	
5.3.	Variability of Dispersion (range, inter quartile range, and standard deviation)	
Unit 6.	Concepts of Probability	18
6.1.	Introduction of Probability	
6.2.	Definition of Probability	
6.3.	Basic Terms of Probability Theory	
6.4.	Counting Rule (permutation and combination)	
6.5.	Additive and Multiplicative Law of Probability	
Unit 7.	Theoretical Probability Distribution	41
7.1.	Random Variables	
7.2.	Binomial distribution	
7.3.	Poisson distribution	
7.4.	Normal Distribution	
Unit 8.	Bivariate Data Analysis	41
8.1.	Introduction	
8.2.	Correlation (Karl Pearson's Coefficient of Correlation)	
8.3.	Linear regression	

CONTENT

Unit 1: Introduction to Statistics	1-10
1.1 Introduction	1
1.2 Origin of Statistics	1
1.3 Meaning of Statistics	2
1.4 Definition of Statistics	2
1.5 Important and Scope of Statistics	4
1.6 Functions of Statistics	5
1.7 Use of Statistics	5
1.8 Important of Statistics in Engineering	8
1.9 Limitations of Statistics	9
Unit 2: Basic Concept of Statistical Studies	11-14
2.1 Introduction	11
2.2 Data	11
2.3 Types of Data	11
2.4 Sources of Primary and Secondary Data	12
2.5 Population and Sample	12
2.6 Parameter and Statistic	13
2.7 Parameter and Statistic	13
2.8 Variables	14
Unit 3: Organizing Raw Data	15-30
3.1 Introduction	15
3.2 Classification of Data	15
3.3 Frequency and Frequency Distribution	19
3.4 Cumulative Frequency Distribution	23
3.5 Meaning and Importance of Table	24
3.6 Tabulation of Data	25
3.7 Parts of Table	27

Unit 4:	Pictorial Representation of Data Set	31-52
4.1	Introduction	31
4.2	Graphical Representation of Data	31
4.3	Diagrammatic Representation of Data	38
4.4	Steam Leaf Display	45
4.5	Limitation of Diagram and Graph	48
4.6	Difference between Diagram and Graph	48
Unit 5:	Summarizing a Data Set	53-110
5.1	Introduction	53
5.2	Requisites of an Ideal Measure of Central Tendency	53
5.3	Measures of Central Tendency	54
5.4	Relation among three Central Tendencies	77
5.5	Partition Values	77
5.6	Variability of Dispersion	93
5.7	Various Measures of Variations	94
Unit 6:	Concepts of Probability	111-132
6.1	Introduction	111
6.2	Basic Terms of Probability	111
6.3	Tree Diagram	114
6.4	Classical or priori (or mathematical) Theory of Probability	114
6.5	Basic Laws of Probability Theory	116
6.6	Counting Rules (Permutation and Combination)	126
Unit 7:	Theoretical Probability Distribution	133-154
7.1	Introduction	133
7.2	Random Variables	133
7.3	Probability Distribution of a Random Variable	134
7.4	Binomial Distribution	137
7.5	Poisson Distribution	145
7.6	Normal Distribution	147

Unit 8: Bivariate Data Analysis **155-182**

8.1	Introduction	155
8.2	Correlation and Its Types	155
8.3	Method of Measuring Correlation	157
8.4	Co-variance	158
8.5	Karl Pearson's Coefficient of Correlation	161
8.6	Correlation Coefficient of Bi-variate Distribution	166
8.7	Properties of Correlation Coefficient	168
8.8	Interpretation of Correlation Coefficient	170
8.9	Line Regression	174
8.10	Regression Equations and Regression Coefficients	174
8.11	Difference between Correlation and Regression	180
	Bibliography	183
	Questions of CTEVT	184
	Area under Standard Normal Curve	192

Introduction to Statistics

1

Unit

1.1. Introduction

The word statistics seems to have been derived from the Latin word 'status' or the Italian word 'Statista' or the German word 'Statistik' each of which means a political state. In ancient times, the government used to collect the information regarding the population, property or wealth of the country for forming military and fiscal policies. But in the modern times, the scope of statistics has considerably enlarged. It is not only helpful to the state in administration, but also considerable use to the Economist, Scientist, Businessman, Research workers, Bankers, Engineering and so on.

1.2. Origin of Statistics

Statistics is the study of information expressing (representing) in terms of data which are usually symbolized by numbers. Thus, we can say that statistic is the beginning stage of the modern mathematics. The numbers and the data can be taken as the two sides of a coin. That is both are mutually begin and developed. Thus, they are the complementary of each other. In the beginning of the human society, the numbers were used as the symbols to keep the record of different informations to systemize their (or activities) behaviours. The problems of the society such as to keep the record of the number of family in the group, number of cattles, quantity of food and etc. Hence, as the numbers are developed, in this way in the same time the statistics is also originated and developed which takes place indifferent sectors of the human society. Thus it can be said that statistics is the oldest branch of modern mathematics.

In ancient time, the ruling community used to collect the informations regarding the members (population), property (needed to the community), number of fighters members (soldiers) to protect the community. In the last few centuries, it has become a part of mathematics and

the mathematicians like Pascal, De'moivre, Laplace , James Bernoulli, Gauss, Bays, Markoff, Euler's Lagrange etc. take interest in the development of the theory of probability as the theory of games and chance phenomena. In the nineteenth century statistics was mainly concerned with official statistics needed to the collection of informations on revenue population and the area of land under cultivation. Thus, the statistics has been gradually develop and it's field of application has been increasing. Now a days, it has broad application in every area of the present states in the world.

1.3. Meaning of Statistics

Statistics is the word which is just influence from the Latin word 'statis' or the Italian word 'statista' or the German word 'statistik' whose typical meaning is political state. It is originated from two different sectors, first from political states and that of secondly from game of chances. These two sectors have two relations (disciplines and characteristics), primarily descriptive and secondly essentially analytical. First, is related to the collection of data and the second is concerned with the concept of chance (or probability). Now a days, the meaning of the statistics is defined as the collection of data, classification of data, tabulation of data, presentation of data, analysis of the data and the conclusion with planning for the remedy of concern problem raised under study (discussion).

1.4. Definition of Statistics

The word statistics is used in two senses- singular and plural. When the word statistics is used in singular, it means the science of state. i.e. the body of principles and techniques used in collection, presentation, analysis and interpretation of statistical data. And when the word statistics is used in plural; it refers to fact and figure i.e. the statistical data itself such as statistics of population, national income, unemployment and so on. The term statistics has been defined by the various writers both in singular as well as in plural sense.

Definition according to singular sense

Croxton and **Cowden** have given a very simple and concise definition of statistics. In their view, " statistics may be defined as the collection, presentation, analysis and interpretation of numerical data." This definition clearly points out four stages in a statistical investigation, namely-

- (a) *Collection of data:* It is the first step in a statistical investigation. Utmost care must be exercised in collecting data because they form the foundation of statistical analysis. If the collected data are faulty, the conclusions drawn can never be reliable.
- (b) *Presentation of data:* After the data have been collected the next step is the presentation of data. There are two different modes in which the collected data may be presented; Diagrams and Graphs.
- (c) *Analysis of data:* After collection and presentation, the next step is that of analysis. The purpose of analyzing data is to dig out information useful for decision-making. Methods used in analyzing the presented data are numerous, ranging from simple observation of the data to complicated, sophisticated and highly mathematical techniques. However, in this text only the most commonly used methods of statistical analysis are included, such as measure of central tendency, measure of dispersion, correlation etc.
- (d) *Interpretation of data:* The fourth and the last step in statistical investigation is interpretation of data. By interpretation of data we mean drawing appropriate conclusions from the data that has been collected and analyzed. The interpretation of data is a difficult task and necessitates a high degree of skill and experiences. If the data are not properly interpreted, the whole object of inquiry may be defeated and fallacious conclusions be drawn. Correct interpretation will lead to a valid conclusion of the study and thus can aid one in taking suitable decisions.

According to the **Prof. Ya-Lun-Chou**, "Statistics is a method of decision-making in the face of uncertainty on the basis of numerical data and calculated risks". This definition highlights the role of statistical techniques in the decision-making in the face of uncertainty.

Definition according to plural sense:

When the word statistics is used in plural, it refers to facts and figures. In plural sense also several definitions have been given by various statisticians. Here we will discuss the definition given by **Prof. Horace Secrist**. He has defined statistics as "By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes numerically expressed,

enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other".

From the above definition, statistics must conform to the following characteristics:

- (a) It should aggregates of facts.
- (b) It should be affected by multiplicity of causes to a marked extent.
- (c) It should be numerically expressed.
- (d) It should be enumerated or estimated according to reasonable standards of accuracy.
- (e) It should be collected in a systematic manner.
- (f) It should be collected for a pre-determined purpose.
- (g) It should be capable of being placed in relation to each other.

1.5. Important and Scope of Statistics

As we have seen in foregoing pages, statistics simplifies complexity. It studies relationships. In this ways, statistics enlarges human experience and thus helps to find the solution of many difficulties for which human being faces now a days. By adopting the techniques of description, comparison and correlation statistics helps in improving human knowledge as well. This is why it is often said that in today's context statistics has been indispensable. It is because of the following reasons:

- (a) It is useful in the administration of various states.
- (b) It is useful to the businessman.
- (c) It is useful to the economists.
- (d) It is useful to the research workers.
- (e) It is useful in planning.
- (f) It is useful to the bankers.
- (g) It is useful to the insurances companies and etc.
- (h) It is useful in supervision.
- (i) It is useful in the physical, biological and other sciences.

1.6. Functions of Statistics

Facts and figures are collected, classified, tabulated, analyzed and interpreted in order to derive conclusions and formulate theories, policies. Test the accuracy and prepare the framework within which one has to work in accomplishing a task and achieving the desired results. All these are done by statistics. The functions of statistics are as follows:

- (a) It presents facts in a definite form.
- (b) It simplifies complexity.
- (c) It facilitates comparison.
- (d) It helps in formulating and testing hypothesis.
- (e) It helps in prediction (forecasting).
- (f) It helps in formulation of suitable policies.
- (g) It helps other sciences in research and conducting tests etc.

1.7. Use of Statistics

As we have seen in foregoing pages, statistics simplifies complexity. It studies relationships. In this way, statistics enlarges human experience and thus helps in the solution of many difficulties which mankind faces today. By adopting the techniques of description, comparison and correlation statistics helps in improving human knowledge as well. Thus, statistics plays a vital role in every field of human activity. Statistics helps in determining the existing position of per capita income, unemployment, population growth rates, housing, medical facilities, etc., in a country. Statistics holds a central position in almost every field including industry, commerce, trade, physics, chemistry, economics, mathematics, biology, psychology, astronomy, etc. So, the application of statistics is very wide. Now we shall discuss some important fields in which statistics is commonly applied.

Statistics and business

Statistics plays an important role in business. A successful businessman must be very quick and accurate in decision making. He knows what his customers want; he should therefore know what to produce and sell and in what quantities. Statistics helps businessmen to plan production according to the taste of the customers, and the quality of the products can also be checked more efficiently by using statistical methods. Businessmen can make correct decisions about the location of business, marketing of the

products, financial resources, etc. Thus, it can be seen that all business activities are based on statistical information.

Statistics and economics

Economics largely depends upon statistics. National income accounts are multipurpose indicators for economists and administrators, and statistical methods are used to prepare these accounts. The relationship between supply and demand is studied by statistical methods; imports and exports, inflation rates, and per capita income are problems which require a good knowledge of statistics. In economics research, statistical methods are used to collect and analyze the data and test hypotheses. Thus, there is inseparable relation of statistics with the study of economical variables.

Statistics and banking

Statistics plays an important role in banking. Banks make use of statistics for a number of purposes. They work on the principle that everyone who deposits their money with the banks does not withdraw it at the same time. The bank earns profits out of these deposits by lending it to others on interest. Bankers use statistical approaches based on probability to estimate the number of deposits and their claims for a certain day.

Statistics and administration

Statistics is essential to a country. Different governmental policies are based on statistics. Statistical data are now widely used in making all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the cost of living, and statistical methods will be used to determine the rise in the cost of living. The preparation of federal and provincial government budgets mainly depends upon statistics because it helps in estimating the expected expenditures and revenue from different sources. So statistics are the eyes of the administration of the state.

Statistics and medical science

Statistical techniques are extremely useful in the study of medical science. In medical science, the statistical tools for the collection, presentation and analysis of observed factual data relating to the causes and incidence of diseases are of paramount importance. For example, the factual data relating to pulse rate, body temperature, Hb%, rate of transpiration, blood pressure, heart beats, weight etc; of the patients greatly help the doctor

for the proper diagnosis of the disease; statistical papers are used to study heart beats through electrocardiogram (E.C.G.).

Statistics and agricultural science

Statistics is indispensable to agricultural science. In agricultural science, we deal with the variabilities within plant and animal species growing and reproducing in different environment. An emphasis is placed on mastering a wide use of statistical methods in order to allow the students to apply these techniques in many fields of agricultural science like: crops production, vegetable production, horticulture, fruit growing and production, plant protection, livestock, veterinary medicine etc.

Crops and animals are planted, grown and harvested. This is affected by variation in temperature, rainfall, day length, soil characteristics, competition from other plants and animals, challenges of diseases, insects or storms, and different management practices. Statistics have been a fundamental part of agricultural research, in keeping up with providing food to an ever growing population.

Agricultural scientists are using statistical procedures adapted from engineers, ecologists, and economists in order to deal with the growing problems of analysis and interpretation of more and more data. Problems associated with sampling will continue to grow, especially when dealing with rare or highly variable events, such as insect counts, microorganisms, mycotoxins, or specie metabolites. The statisticians provide usable, efficient, understandable, user-friendly procedures to analyze and interpret more and more complex data.

Statistics and research

Statistics is indispensable tools in research work. Most of the advancement in knowledge has taken place because of experiments regarding collecting, organizing, presenting, analyzing and interpreting quantitative data gathered from the related field of study. For example, experiments about crop yields and other related variables such as different types of fertilizers, types of soils, irrigations etc are design and analysed with the help of statistical methods. Statistical techniques are extremely useful in the study of the growth of animals under the different diets and environments. It is impossible to understand the meaning and implications of most of the research findings related to the quantitative data without statistical methods.

Statistics and Engineering

The field of statistics deals with the collection, presentation, analysis and use of data to make decisions, to solve problems and design products and processes. Thus, statistics is a science that helps us make decisions and draw conclusions about the phenomena. Engineers solve problems constantly and repetitively that's why the need of statistics in the Engineering domain is of a huge value. Engineering statistics combines engineering and statistics using scientific methods for analyzing data. Engineering statistics involves data concerning manufacturing processes such as: component dimensions, tolerances, type of material, and fabrication process control.

1.8. Important of Statistics in Engineering

There are many methods used in engineering analysis and they are often displayed as histograms to give a visual of the data as opposed to being just numerical. The important of statistics in engineering are as follows:

- (a) Design of Experiments is a methodology for formulating scientific and engineering problems using statistical models. The protocol specifies a randomization procedure for the experiment and specifies the primary data-analysis, particularly in hypothesis testing. In a secondary analysis, the statistical analyst further examines the data to suggest other questions and to help plan future experiments. In engineering applications, the goal is often to optimize a process or product, rather than to subject a scientific hypothesis to test of its predictive adequacy. The use of optimal (or near optimal) designs reduces the cost of experimentation.
- (b) Quality control and process control use statistics as a tool to manage conformance to specifications of manufacturing processes and their products.
- (c) Time and methods engineering use statistics to study repetitive operations in manufacturing in order to set standards and find optimum (in some sense) manufacturing procedures.
- (d) Reliability engineering which measures the ability of a system to perform for its intended function (and time) and has tools for improving performance.
- (e) Probabilistic design involving the use of probability in product and system design

- (f) System identification uses statistical methods to build mathematical models of dynamical systems from measured data. System identification also includes the optimal design of experiments for efficiently generating informative data for fitting such models.

The role of Statistics in Engineering is indispensable. All practicing engineers realize that soon after getting into the workplace. Starting from designing a product, making a finished one and making it work, at every step an engineer needs help of Statistics in some form or other. From rudimentary statistics of data collection to advance probability in Quantum mechanics and Space science, Aeronautics and Civil, Mechanical and Electrical, statistics becomes a tool that engineers use frequently.

1.9. Limitations of Statistics

Despite the usefulness of statistics in many fields, the following are the important limitations of the science of statistics:

- (a) Statistics does not deal with individual measurement since it deals with aggregates of facts, the study of individual measurements lies out side the scope of statistics.
- (b) Statistics deals only with quantitative characteristics. The qualitative characteristics like honesty, efficiency, intelligence, blindness and deafness cannot be studied directly.
- (c) Statistical laws are not exact. Since the statistical laws are probabilistic in nature, inference based on them are only approximate.
- (d) Statistical results are true only on an average. The conclusions obtained statistically are not universally true. They are true only under certain conditions.
- (e) Statistics can be misused. The greatest limitation of statistics is that it is liable to be misused. For example, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions.

Basic Concept of Statistical Studies

2.1 Introduction

This unit gives the basic idea of data, its types, idea of collecting data from different sources (i.e. concept of primary and secondary data). To differentiate between population and sample, parameter and statistics. Also it gives the basic concept of variables and its types which are very important for the study of statistics.

2.2 Data

From the very definitions of Statistics, it is an aggregate of facts that can numerically be expressed. These types of facts are known as data. For drawing any statistical inference we need to collect and compile them. In the present unit we will discuss about the collection, classification, tabulation and diagrammatical and graphical representation of numerical data. The details of each are discussed below in the next articles.

2.3 Collection of Data

For any statistical investigation, the first attempt is to collect the data. Utmost care must be given while collecting data because it is only on the data collection that the super-structure of statistical analysis and decision-making is built. If the data collected are reliable, correct decision can be taken otherwise one may reach to the conclusions, which may mislead. Before collecting the data for a given statistical enquiry, it is essential to examine carefully the following points:

- (a) Objectives and scope of the enquiry
- (b) Sources of information
- (c) Method of data collection
- (d) Units of data collection
- (e) Degree of accuracy aimed at in the final results.

2.4 Types of Data

Depending on the sources, statistical data are classified under two categories and they are primary data and secondary data.

Primary Data: The data that are originally collected by an investigator for the first time for the purpose of statistical enquiry are taken as primary data. For example, data obtained in a population census by the Ministry of Population and Environment in every ten years is primary data.

Secondary Data: The data that are not originally but rather obtained from some published or unpublished sources are known as secondary data. For example, the reports and publications made by Central Bureau of Statistics are primary for that organization but secondary data for those who use it.

2.5 Sources of Primary and Secondary Data

Sources of Primary Data

Primary data may be collected by any of the following methods.

- (a) Direct personal interviews
- (b) Indirect oral interviews
- (c) Information from correspondents
- (d) Mailed questionnaire method
- (e) Questionnaire sent through enumerators

Direct personal interviews: Under this method of collecting data, there is a face to face contact with the persons from whom the information is to be obtained. The interviewer asks them questions pertaining to the survey and collects the desired information.

Indirect oral interviews: Under this method, the investigator contacts third parties called witness capable of supplying the necessary information. This method is generally adopted in case when the informants hesitate to give information directly.

Information from correspondents: Under this method, the investigator appoints local agents or correspondents in different places to collect information. These agents collect the information and supply it to the central office, where this information is further processed.

Mailed questionnaire method: Under this method, a list of questions pertaining to the enquiry is prepared and is sent by post to the various persons from whom the information is to be collected. The informants are expected to mail it back after duly filling in all the questions contained in the questionnaire. This is the most popular method of collecting primary data.

Questionnaire sent through enumerators: Under this method, the questionnaire is sent to the informants through the enumerators and not by post. The enumerators fill up the questionnaires by interviewing the informants. The enumerators ask the answers to the questions contained in questionnaire and fill up the questionnaire themselves.

Sources of Secondary Data

The sources of secondary data can broadly be classified under two heads and they are published sources and unpublished sources.

Published sources: Published data are found from various sources such as

- Official publications of different ministries and departments of the government e.g. National Planning Commission, Central Bureau of Statistics and so on.
- Reports and publications of organizations like Nepal Chamber of Commerce, Nepal Rastra Bank, Nepal stock Exchange and so on.
- Reports of economists, research scholars, Universities and so on.
- Official publications of foreign government or international bodies like United Nations Organization (UNO), International Labor Organization (ILO) etc.

Unpublished sources: All statistical material is not always published. There are various sources of unpublished data such as records maintained by various government and private offices, studies made by research institutions, scholars etc.

2.6 Population and Sample

The whole mass of observation (data) collecting from different sources is called the population (or universe). That is the total number of data collected, are called the population about which we need to study to find a concrete conclusion. But it will not always be convenient (or possible) to examine each member of the entire population. Such as, it is not practical to test all the mangoes in the trees. For that we test few of them whether they are ready to sell on market or not. Here, the total numbers of mangoes in the trees is the population and the few number of the mangoes which are kept to test from the tree is called the sample members. Now the randomly selected items (pieces) from the whole observation are called the sample pieces or simply it is called sample. And the process of selecting some of the sample pieces from the whole mass of observation is called the sampling. Thus, in statistics, the whole mass of the objects (or observations) is called population (or universe) and sample is a part of the population. Hence, in the study of statistics the population refers not only to people but it refers to the whole mass of collecting observations over which an analysis is made to find out a concrete conclusion to remove any types of problems (or errors).

2.7 Parameter and Statistic

The statistical measures like, mean, median, standard deviation, moments etc. which are computed from population (or universe) are called

the parameters. Parameter, in statistics is any numerical quantity that characterizes a given population or some aspect of it. When parameters are present, the definition actually defines a whole of universe, one of for every valid set of values of the parameters. Again the statistical measures computed from the sample data are called the statistics. The notation (symbols) that are used for population parameter and sample statistic are as below:

	Population	Samples
Size	N	n
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	p	p

2.8 Variables

Any quantity or quality liable to show variation from one individual to the other in the same population is known as variable. An individual observation of any variable is known as variate. A variable may be either discrete or continuous.

Discrete variable: A discrete variable is that which cannot take all values and there is a gap between one value and the other. Number of persons in a family, number of books in a library are discrete variable. The discrete variable may take any integer value from 0 to ∞ .

Continuous variable: These variables which can take all the possible values (integral as well as fractional) in a given specified range are called continuous variable. For example, the weight and height of human being is a continuous variable because it may take any value. Height of students in a class may be 110cm, 112.5cm, 112.7cm, 115cm, 117cm, 117.3cm and so on.

EXERCISE - 2

1. What is a data? Write its types.
2. What is primary data? Discuss about the various method of collecting primary data.
3. What is secondary data? Discuss about the sources of secondary data.
4. Distinguish between primary and secondary data. Explain with examples.
5. Distinguish between primary and secondary data. What are the sources of secondary data? [CTEVT 2072/2074/2075]
6. Differentiate between population and sample with suitable examples.
7. What are parameter and statistic? Also write their different symbols.
8. Discuss about the variables in statistics with their different types.



the parameters. Parameter, in statistics is any numerical quantity that characterizes a given population or some aspect of it. When parameters are present, the definition actually defines a whole of universe, one of for every valid set of values of the parameters. Again the statistical measures computed from the sample data are called the statistics. The notation (symbols) that are used for population parameter and sample statistic are as below:

	Population	Samples
Size	N	n
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	p	p

2.8 Variables

Any quantity or quality liable to show variation from one individual to the other in the same population is known as variable. An individual observation of any variable is known as variate. A variable may be either discrete or continuous.

Discrete variable: A discrete variable is that which cannot take all values and there is a gap between one value and the other. Number of persons in a family, number of books in a library are discrete variable. The discrete variable may take any integer value from 0 to ∞ .

Continuous variable: These variables which can take all the possible values (integral as well as fractional) in a given specified range are called continuous variable. For example, the weight and height of human being is a continuous variable because it may take any value. Height of students in a class may be 110cm, 112.5cm, 112.7cm, 115cm, 117cm, 117.3cm and so on.

EXERCISE - 2

1. What is a data? Write its types.
2. What is primary data? Discuss about the various method of collecting primary data.
3. What is secondary data? Discuss about the sources of secondary data.
4. Distinguish between primary and secondary data. Explain with examples.
5. Distinguish between primary and secondary data. What are the sources of secondary data? [CTEVT 2072/2074/2075]
6. Differentiate between population and sample with suitable examples.
7. What are parameter and statistic? Also write their different symbols.
8. Discuss about the variables in statistics with their different types.



Organizing Raw Data

3.1 Introduction

Raw data are collected from two sources first by an investigator for the first time for the purpose of statistical inquiry which are called primary data. Secondly the raw data are collected from some published or unpublished sources. At first the collected data are in rough form and need to present it in a systematic and scientific way. For that the raw data are classified and tabulated in simple and lucid way. Classification means, process of arranging the data into different groups of classes according to their similarities (characteristic). It is the first step in tabulation become the items (data) with common similarities must be keep together before the data are presented in table. The process and rules of classification and tabulation are discussed below with examples.

3.2 Classification of Data

In the last few topics we described the process of data collection. The collected data in any statistical investigation are so voluminous and huge. So the next step is to organize it in a simple and compact form which will highlight the important characteristics of the data, facilitate comparisons and render it suitable for further statistical analysis and interpretations.

To know the important characteristics of the data, the heterogeneous mass should be changed in to condense and homogeneous form. This process is known as the classification of data. According to Secrist "Classification is the process of arranging data in to sequence and groups according to their common characteristics or separating them into different but related parts." Thus, classification impresses upon the arrangement of the data into different classes, which are to be determined depending upon the nature, objectives and scope of the enquiry. For example, the number of students registered in CTEVT during the academic year 2005-06 may be classified on the basis of any one of the following criterion:

- (a) Sex
- (b) Age
- (c) The district to which they belong
- (d) Religion
- (e) Different faculties like Pharmacy, General Medicine, Oral Health etc.
- (f) Height
- (g) Weight
- (h) Institutions (Colleges) and so on.

Objectives of classification

Process of classification is carried out with the following objectives:

- (a) To condense the data in such a manner that points of similarities and dissimilarities can easily be brought out.
- (b) To make the study and comparison easier.
- (c) To pinpoint the most significant features of the data at a glance.
- (d) To simplify the complexion of the raw data and make it to possible to draw statistical inferences.
- (e) To help the drafting of the final report.
- (f) To make the proper use of the collected data and prepares the basis of tabulation.

Rules for classification

Classification is one of the most important techniques for the statistical treatment and analysis and numerical data. There are no hard and fast rules for the classification. The following general guiding principles may be observed for good classification.

- (a) It should be unambiguous. The classes should be rigidly defined so that they should not lead to any ambiguity.
- (b) It should be exhaustive and mutually exclusive. The classification must be exhaustive in the sense that each and every item in the data must belong to one of the classes. Further the various classes should be mutually disjoint or non-overlapping so that an observed value belongs to one and only one of the classes.
- (c) It should be stable. In order to make meaningful comparisons of the results, the same pattern of classification should be adopted throughout the analysis and also for further enquiries on the same subject.

- (d) It should be suitable for the purpose. The classification must be kept with the objectives of the enquiry.
- (e) It should be flexible. A good classification should be flexible in that it should be adjustable to the new and changed situations and conditions.

Types of classification

Broadly, the collected data can be classified on the following four basis:

- a. Geographical - Area wise (Regional, cities, districts etc.)
- b. Chronological - On the basis of time
- c. Qualitative - On the basis of some character or attribute
- d. Quantitative - In terms of magnitude

(a) Geographical classification

In this type of classification data are classified on the basis of geographical or locational differences between the various items, like countries, cities, regions, zones, districts, areas etc. For example, the yield of food grains in Nepal may be presented region wise in the following manner:

Region wise production of Food Grains: 2004-05

Region	Total Food Grains (In Kg/hect)
Eastern Development Region	170
Central Development Region	155
Western Development Region	160
Mid Western Development Region	145
Far Western Development Region	130

(b) Chronological classification

In this type of classification, the data are classified on the basis of differences in time. For example, the HIV infected peoples in Nepal may be presented with respect to the time in the following way:

HIV Infected People in Nepal

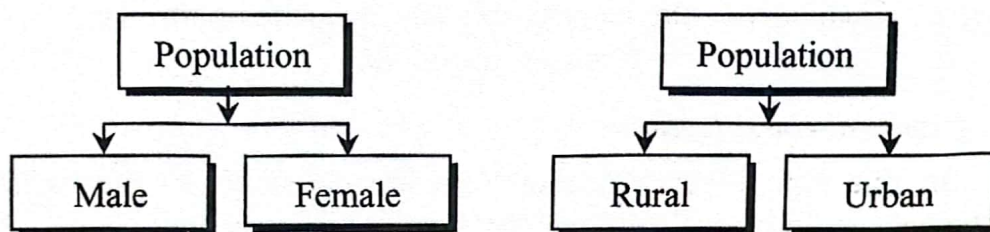
Year	HIV Infected People In hundred
2000	15
2001	18
2002	21
2003	27
2004	39

(c) Qualitative classification

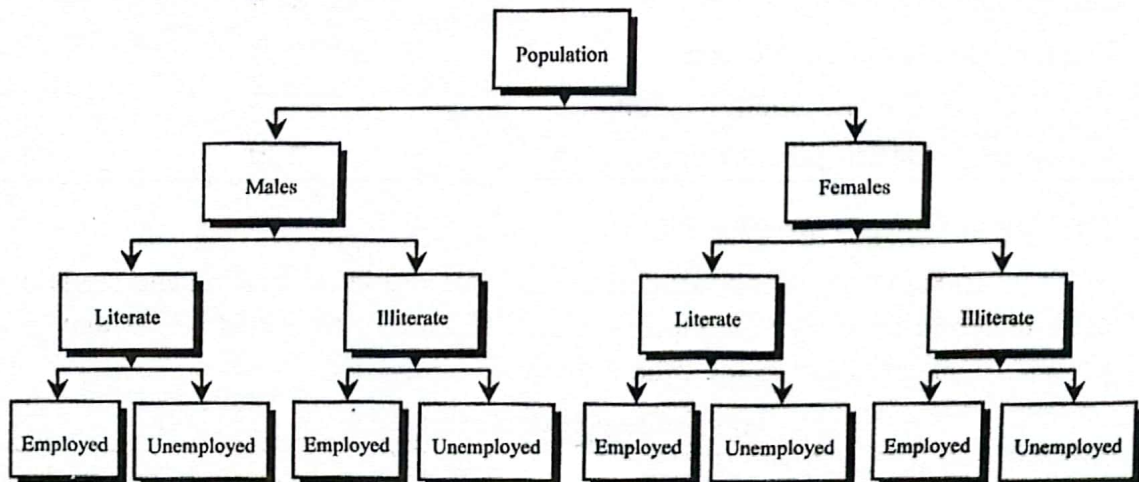
In this type of classification, the data are classified according to some qualitative phenomena which can not be measured quantitatively like honesty, beauty, intelligence, employment, occupation, sex, literacy, religion etc.

In qualitative classification the data are classified according to the presence or absence of the attributes in the given units. There may be two types of qualitative classification.

(1) *Simple classification*: Here the object of classification can be put under two columns. For example, human population may be classified into two categories as males/females, married/unmarried, literate, illiterate, rich/poor etc.



(2) *Manifold classification*: It is the classification in which more than one attribute is studied simultaneously. For example, we may first divide the population into male and females on the basis of the attribute 'sex'; each of these classes may be further subdivided into literate and illiterate on the basis of the attribute 'literacy'; and so on. An example of manifold classification is given below:

**(d) Quantitative classification**

In this type of classification the collected data are classified on the basis of some characteristics which is capable of quantitative measurement like age, height, weight, pulse rate, blood pressure, RBC, WBC, HB % and so

on. For example, chickens of a poultry farm may be classified on the basis of their growth rate. Quantitative data may also be classified into two types.

(1) *Discrete*: Values of variate vary by infinite jumps. For example, the oxygen consumption of rat (*Rattus*) of different weight groups was measured as 500 cc/h/100ml, 600 cc/h/100ml, 620 cc/h/100ml, 680 cc/h/100ml and so on.

(2) *Continuous*: Values of variate do not exhibit any breaks or jumps. For example, the students of a college may be classified according to weight as follows:

Weight in kg.	No. of students
10-20	17
20-30	11
30-40	22
40-50	15
50-60	10

3.3 Frequency and Frequency Distribution

The number of times in which the variate value is repeated is known as frequency. The arrangement of the value of variable and their corresponding frequencies is called the frequency distribution. The table in which the value of variable and their corresponding frequencies are arranged is called frequency distribution table. If the frequency distribution consists of only one variable then it is called univariate frequency distribution. If it consists two variable then it is called bivariate frequency distribution. Univariate frequency distribution are of three types.

Individual frequency distribution (Individual Series)

Discrete frequency distribution (Discrete series)

Continuous frequency distribution (Continuous series)

(a) Individual frequency distribution

In this frequency distribution, the individual value of variable is listed. This distribution is used when the number of observation is small and all observations are distinct. For example the temperature of a certain city in the first week of August is given as follows.

Day	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Tem (°c)	34	35	33	37	31	33	32

(b) Discrete frequency distribution

Discrete frequency distribution is also called ungrouped frequency distribution. In this distribution we write the variable and their corresponding number of repetitions called frequency from the set of observations of any experiment. While writing the frequency we use the tally marks. For example, we make following frequency distribution from 25 women according to their number of child.

3, 2, 3, 4, 2, 2, 3, 4, 5, 1, 1, 2, 3, 4, 3, 2, 2, 2, 1, 4, 3, 2, 2, 3, 1.

Now, let us make the frequency distribution table

No. of children	No. of women (f)	
(X)	Tally marks	f
1		4
2	 	9
3	 	7
4		4
5		1
		$\Sigma f = 25$

While using tally marks we use the symbols |, ||, |||, |||| for 1, 2, 3, 4 and we write ||||| for 5.

The discrete frequency is used if the number of distinct variate value is not large i.e. if number of district variate value is between 5 and 10. In above example the different values that X has taken is only 5.

(c) Continuous frequency distribution

When the set of observations contains large number of distinct values or if variation of the data is very large then the distribution of the data in discrete series will be much complicated. In this case, we form certain group of variate value called classes. Then we count the number of variate value lying in side the class as the frequency of that class.

For example the mark of 28 students in a test is given:

95,	72,	05,	09,	11,	18,	25,	29
21,	33,	55,	67,	73,	27,	44,	49
93,	82,	86,	33,	51,	63,	45	32
33,	39,	42,	89				

We make the continuous frequency distribution as follows

Class	Tally marks	Frequency
0-10		2
10-20		2
20-30		4
30-40	 	5
40-50		4
50-60		2
60-70		2
70-80		2
80-90		3
90-100		2

While making continuous frequency distribution we mainly use two types of classes.

1) *Exclusive classes*

In exclusive classes the value of variable which is equal to the lower limit of the class belongs to that class and the value equal to upper limit of the class doesn't belong to that class.

In exclusive classes, the upper limit of each class is equal to the lower limit of preceding class. The exclusive classes are also known as overlapping classes. The frequency distribution where the exclusive classes are used is called exclusive continuous series.

An example of exclusive continuous series is given here.

Class	Frequency
0-10	2
10-20	7
20-30	3
30-40	9
40-50	11

2) *Inclusive classes*

In inclusive classes, the values of variable equal to the lower limit and upper limit belong to that class. The frequency distribution where the

inclusive classes are used is called inclusive continuous frequency distribution. In inclusive classes there is a gap between the classes. An example of inclusive frequency distribution is as shown below:

Number of member	No. of family
1-2	10
3-4	13
5-6	19
7-8	7
9-10	3

While making continuous frequency distribution we usually use the exclusive classes. If the given classes are given as inclusive type then we can change them into exclusive classes by introducing **correction factor**. It is denoted by C_f and is calculated by

$$C_f = \frac{\text{lower limit of second class} - \text{upper limit of first class}}{2}$$

After calculating the value of C_f we subtract C_f from lower limit of each class and we add C_f to upper limit of each class to get the exclusive series.

Example 1.

Change the following frequency distribution into exclusive type

Rain fall in mm	No. of days
10-19	12
20-29	7
30-39	10
40-49	4
50-59	2

Solution:

$$\text{Here, Correction factor (y)} = \frac{20-19}{2} = \frac{1}{2} = 0.5$$

Thus the above inclusive series can be reduced to the following exclusive distribution we get.

Class	No. of days
9.5-19.5	12
19.5-29.5	7
29.5-39.5	10
39.5-49.5	4
49.5-59.5	2

While making the continuous frequency distribution we should remember the following points.

- The number of classes should be between 5 and 15.
- The class interval should be of size 5 or multiple of 5.
- All classes should have equal interval.
- The lower limit of the first class must be 0 or 5 or multiple of 5.
- Exclusive distribution should be made.
- Class frequency should be obtained by using tally marks.

3.4 Cumulative Frequency Distribution

The frequency distribution in which frequencies are added is called cumulative frequency distribution. There are mainly two types of cumulative frequency distributions

- Less than cumulative frequency distribution.
- More than cumulative frequency distribution.

(a) Less than cumulative frequency distribution

To make less than cumulative frequency distribution, we arrange the classes in the ascending order of magnitude and we add the frequencies from the top. Instead of writing classes the word less than or below are used to the upper limit of the classes.

(b) More than cumulative frequency distribution

In this case, the first step is arranging the data in their ascending order of magnitude and we add the frequency from bottom. Instead of writing classes we use the word 'above' or 'more than' with the lower limit of each of the classes.

Example 2.

From the following frequency distribution, make less than and more than cumulative frequency distribution.

Marks	No. of students
0-10	3
10-20	7
20-30	11
30-40	5
40-50	8

Solution:

The less than cumulative frequency distribution table is

Marks	No. of students
Less than 10	3
Less than 20	10
Less than 30	21
Less than 40	26
Less than 50	34

Next, the more than cumulative frequency distribution of the given frequency distribution is

Marks	No. of students
More than 0	34
More than 10	31
More than 20	24
More than 30	13
More than 40	8

3.5 Meaning and Importance of Table

In statistics, the first step is the collection of data from different Sources which can be collect using various rules (such as direct personal interviews, indirect oral interviews, mailed questionnaire method, published sources etc.). Then data are classified into different groups according to their characteristics and the objectives of study. After classified, the next step is to present the data into tabular form. The tabulation is the process of arranging the data in definite manner into rows and columns form. Rows are horizontal arrangements and that of columns are the vertical arrangements of the collected data. The purpose of tabulation is to simplify the presentation of data and to make easier to compare between the related informations. It is the last step of collection of data and is a stepping stone to analyze and interpret of the figure. The important of table are as below.

- (a) It helps the raw (or haphazard) data to present in systematic and scientific way.
- (b) It helps to identify the number of units belonging to different classified groups.
- (c) It makes easier to compare among the similar types of data together.
- (d) It is the simple way of presenting the raw data in three types of series (like individual, discrete and continuous series).
- (e) It helps to obtain the different types of statistical measurements. (Like: mean, medium, mode, partition values, deviations, moments etc).
- (f) It helps to present the facts into simpler form.
- (g) It facilitates comparison of data with different characteristics.
- (h) I makes easier for further statistical treatment or analysis of data etc.
- (i) It helps to analyze among data and to find concrete conclusion in short time.

3.6 Tabulation of Data

Tabulation is the process of orderly arrangement of data into series of rows and columns, where they can be read into two dimensions. Tabulated data can be more easily understood than those which are not tabulated. Tables make it possible for the analyst to present a huge mass of data in a detailed and orderly manner within the minimum space. It facilitates the detection of errors and omissions and also helps to compare two or more series.

Types of table

All types of table can be reduced to two major types.

- (a) Simple and complex table
- (b) General and summary tables.

(a) Simple and complex table

In simple table, only one characteristic is shown. It is also known as one way table. For example, the age and number of patients of tuberculosis is shown here

Age	0-10	10-20	20-30	30-40	40-50
No. of patients	5	3	1	10	27

In complex table, two or more than two characteristic are shown. If there are two co-ordinate factors, the table is called double table. If the number of co-ordinate factor is three then it is called triple table. If the number of co-ordinate factor is more than three then it is called multiple or manifold table.

For example,

Age	No. of patients	
	Male	Female
0-10	2	3
10-20	2	1
20-30	1	0
30-40	5	5
40-50	20	7

This is an example of two way table. If the male and female groups are further classified then we say it as three way table.

(b) General and summary table

These contain detailed information and are not constructed for specific discussion. This kind of table is rarely used for biological researches. For example in 2005, there were 83 fish ponds in far eastern region of Nepal out of which 6 were run by government. These produced fishes of 5 species in quintal as below.

Private: Labeo-52, Wallage-8, Mystus-10, H. Fossilis-2, Catela -5

Government: Labeo-2, Mystus-1, H. Fossilis-1 Catela-2.

We can frame general and summary table of above data as follows:

Fish pounds and fish production			
Species	Private ponds	Govt. ponds	Total
Labeo	52	0	52
Wallago	8	2	10
Mystus	10	1	11
H. fossilis	2	1	3
Catela	5	2	7
Total	77	6	83

3.7 Parts of Table

The parts of table depends upon the nature of the data and the purpose of the investigation. But, in general, a good statistical table contains the following parts.

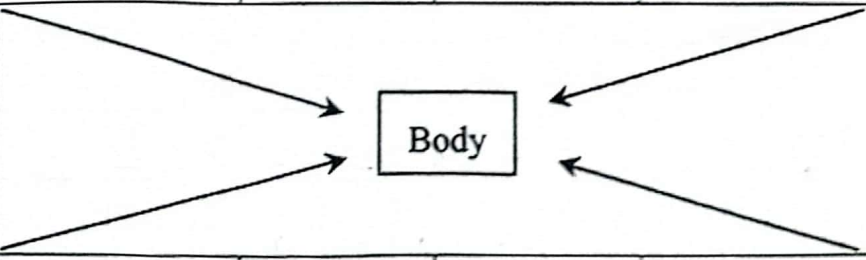
- (a) Table number
- (b) Title
- (c) Head notes
- (d) Captions and stubs (Heading of each columns and heading of each rows)
- (e) Body of the table
- (f) Foot note.
- (g) Source note.

A format of a table is given here.

Table No.

Title

Head notes (If any)

Captions (Title)					
Stubs	Sub heads		Sub heads		Total
	Column head	Column head	Column head	Column head	
					
Total					

Foot note:

Source note:

Requisites of a good table

- (a) The table should be simple and compact so that it is readily comprehensible.
- (b) A table should be complete and self explanatory. It should have suitable title, head note, captions and stubs and foot notes.

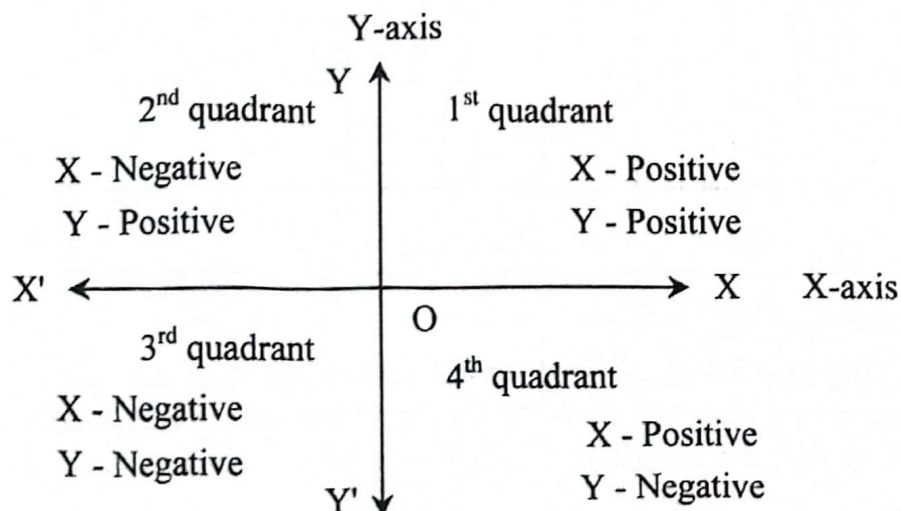
Pictorial Representation of Data Set

4.1 Introduction

After collecting the data, we classify and tabulate them. But the representation of data in table only is not always sufficient to understand and analyze the data. When their size increase the tabular representation become confusing and unsufficient. To avoid these difficulties we present the data in graphs and diagrams. Graphs and diagrams are the visual form of presentation of data where comparisons can be made between two or more phenomenon. In graphic and diagrammatic representation, we use geometric figure like points, lines, rectangles, bars and circles instead of numerical value. The characteristic of the observations can be obtained by inspection in graphs and diagram.

4.2 Graphical Representation of data

For graphical presentation of data, some basic knowledge of graph is required. In graphical method we use graph paper where we draw two straight lines one horizontal and the other vertical line. The horizontal line is called abscissa (X-axis) and the vertical line is called ordinate (Y-axis) the meeting point is called origin.



In X-axis, the points right to origin is taken as positive and left to origin is taken as negative. Similarly in Y-axis, the points above and below the origin is taken as positive and negative. The axes of X and Y divide the whole plane into four parts called quadrants. The name of quadrants and the sign of X and Y on them is shown above.

To represent the data in graph we first change them into grouped data i.e. in the frequency distribution table. The grouped data so obtained can be represented graphically in any one of the following ways.

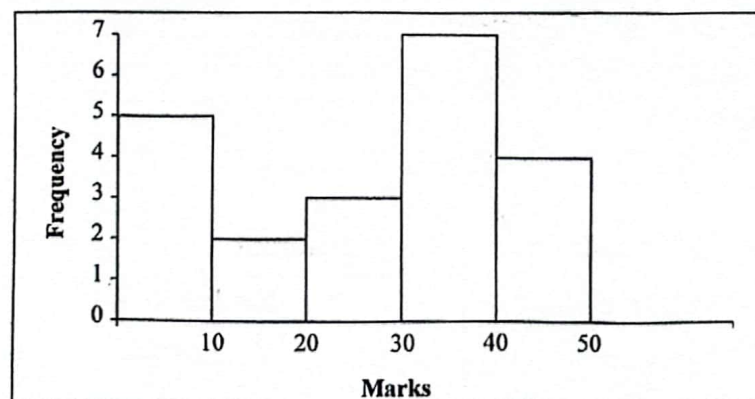
- (a) Histogram
- (b) Frequency polygon
- (c) Frequency curve
- (d) Cumulative frequency curve or Ogive.

(a) Histogram

This graph is used for continuous frequency distribution. The width of the class interval is plotted along x-axis and the corresponding frequencies of each class is represented by the length of adjacent rectangles. While constructing histogram we first make the distribution as having length of each class interval equal. For example, draw the histogram of following:

Mark	0-10	10-20	20-30	30-40	40-50
Frequency	5	2	3	7	4

HISTOGRAM REPRESENTING THE MARKS OBTAINED BY STUDENTS



(b) Frequency polygon

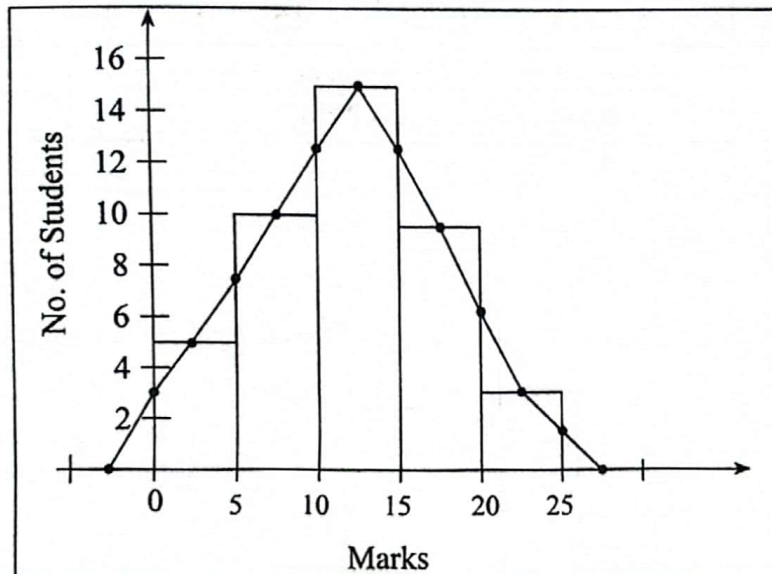
This graph is mainly used for discrete frequency distribution. To obtain the frequency polygon, we plot the value of variable along x-axis and their corresponding frequencies along y-axis. The points so obtained are joined by straight lines.

We can construct the frequency polygon for the continuous distribution also. In this case we join the mid point of top of each rectangle of the histogram.

The frequency polygon with histogram of the data is given below.

Marks	0-5	5-10	10-15	15-20	20-25
No. of Students	5	10	15	9	3

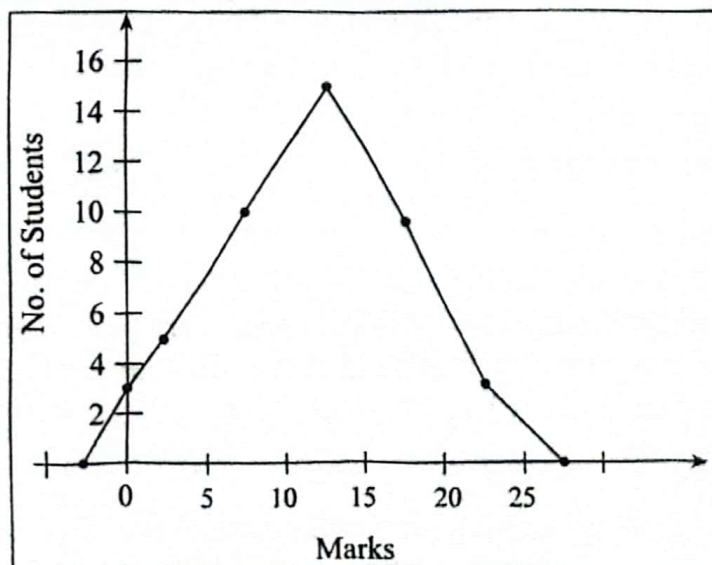
**FREQUENCY POLYGON (WITH HISTOGRAM)
REPRESENTING MARKS OBTAINED BY STUDENTS**



Without drawing histogram we construct frequency polygon by plotting mid point of each class along x-axis and their corresponding frequencies along y-axis.

The frequency polygon without histogram can be constructed from the above example as shown in the following graph.

**FREQUENCY POLYGON (WITH HISTOGRAM)
REPRESENTING MARKS OBTAINED BY STUDENTS**

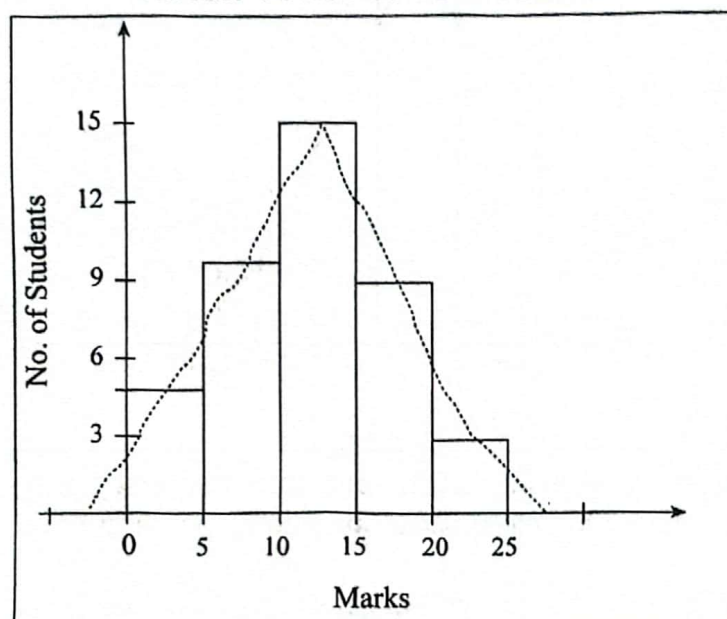


(c) Frequency curve

A smooth free hand curve drawn through the vertices of a frequency polygon is known as frequency curve. The area enclosed by frequency polygon and frequency curve above axis must be equal. So the end points of frequency polygon and frequency curve should join the points on x-axis. Here is an example showing frequency curve.

Marks	0-5	5-10	10-15	15-20	20-25
No. of Students	5	10	15	9	3

**FREQUENCY CURVE REPRESENTING
MARKS OBTAINED BY STUDENTS**

**(d) Cumulative frequency curve (Ogive)**

The graphical display of cumulative frequency distribution is known as ogive. To draw ogive we plot the cumulative frequencies along y-axis and the corresponding class boundaries along x-axis. The ogive has two types.

- 1) Less than ogive
- 2) More than ogive.

1) Less than ogive

For less than ogive, the upper limit of each class is plotted along x-axis and the corresponding cumulated frequencies are plotted along y-axis. The points so obtained are joined by free hand curve. The curve of less than ogive is always increasing function.

2) More than ogive

For more than ogive the lower limit of each class is plotted along x-axis and the corresponding cumulated frequencies are plotted along y-axis. The points

so obtained is joined by free hand curve to get more than ogive. The graph of more than ogive is decreasing function.

Example 1.

Draw less than and more than ogive from following data.

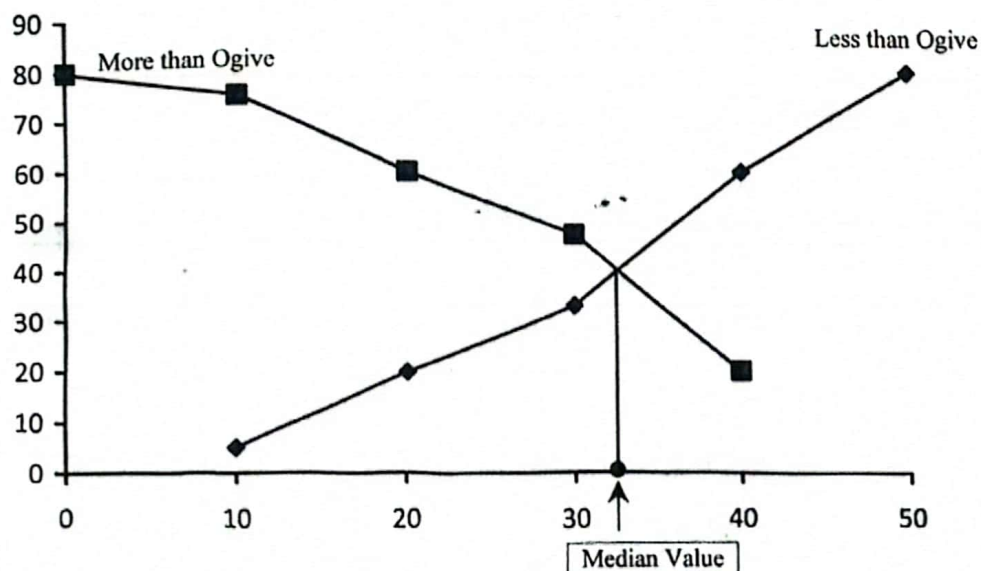
No. of grains	0-10	10-20	20-30	30-40	40-50
No. of plants	5	15	13	27	20

Solution:

Let us make the cumulative frequency distribution table

Class	frequency	less than	Cf	more than	Cf
0-10	5	less than 10	5	more than 0	80
10-20	15	less than 20	20	more than 10	75
20-30	13	less than 30	33	more than 20	60
30-40	27	less than 30	60	more than 30	47
40-50	20	less than 50	80	more than 40	20

OGIVE CURVE REPRESENTING NUMBER OF GRAINS



From the point where less than and more than ogive curves intersect each other, draw a perpendicular on the x-axis. The point where this perpendicular touches the x-axis gives the value of median.

Exercise - 4(I)

1. Why we need to represent the data in graph (or in diagrams) after tabulation? Discuss the different between graph and diagram.
2. Write in brief about the different graphical representation of data taking examples.
3. Write short notes about
 - (a) Histogram
 - (b) Frequency polygon
 - (c) Frequency curve
 - (d) Cumulative frequency curve (Ogive)
4. What are the merits of graphic representation of data?
5. In a study of life of vehicles, the following data were obtained:

Life (in years)	No. of vehicles
10-20	2
20-30	10
30-40	25
40-50	15
50-60	5
60-70	3

Draw a frequency polygon for the above data.

6. Draw histogram, frequency polygon and Ogive with the help of following table. (first change into exclusive classes)

a)

Class interval	Frequency
1-10 "	3
11-20	14
21-30	21
31-40	25
41-50	40
51-60	40
61-70	47
71-80	50

b)

Class interval	Frequency
24.5-29.5	3
30.5-35.5	6
36.5-41.5	14
42.5-47.5	20
48.5-53.5	25
54.5-59.5	37
60.5-65.5	39
66.5-72.5	42

Following are the marks obtained by 50 students in the examination; 30, 41, 45, 45, 40, 35, 28, 25, 35, 49, 62, 60, 56, 55, 44, 65, 52, 50, 47, 42, 54, 43, 28, 35, 42, 45, 57, 68, 62, 65, 35, 30, 35, 42, 47, 25, 28, 48, 50, 72, 60, 56, 68, 75, 49, 55, 78, 57, 65, 55. Construct the grouped frequency table then a histogram with class interval of 10. Also locate the modal value.

8. Height of students measured between the aged 5-30 are presented in the following table:

Height (in cm)	No. of boys	No. of girls
70-75	10	12
75-80	22	18
80-85	55	30
85-90	20	10
90-95	13	3
95-100	5	2

Draw cumulative frequency curve for above data. Also locate the median value.

9. Calculate less than and more than ogive from following frequency distribution. Also mention the median value.

Height (cm)	No. of students
160-162	10
162-164	15
164-166	17
166-168	19
168-170	20
170-172	26
172-174	29
174-176	30

10. Followings are the temperature recorded of last 60 days. Draw the frequency polygon, frequency curve and histogram of the data:

Temperature ($^{\circ}\text{C}$)	5-10	10-15	15-20	20-25	25-30	30-35
No. of days	6	12	15	17	8	2

11. Draw the 'less than' and 'more than' ogives from the data below:

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	8	12	20	6	4	10

4.3 Diagrammatic representation of data

Diagrams are one of the easiest and most commonly used devices of presenting numerical data of any field. Diagram helps biostatisticians to visualize the meaning of numerical complex at a glance. Diagrams are easiest method to understand the quantitative data for non mathematical persons too. In diagrammatic analysis two types of diagrams are commonly used.

- Bar diagram
- Pie-Chart (Angular Diagram).

(a) Bar diagram

Bar diagram are also known as one dimensional diagram in which height of the bars represents the magnitude of the data. Bar diagram consists of set of rectangles of equal width and equal spacing between rectangles the following points should be kept in mind to draw bar diagram.

- All the bars used on a study must have equal width.
- Uniform spacing should be given between bars to make the diagram look more attractive and elegant.
- The height of the rectangles (bars) must represent the frequency or the magnitude of the observation.
- All the bars should be constructed on the same base line.
- The corresponding figure (number) to each bar can be written at the top of each bar.
- In practical, vertical bars are in use but it can be constructed in horizontal direction.

Types of bar diagram

- Simple bar diagram.
- Sub divided or component bar diagram.
- Percentage bar diagram.
- Multiple bar diagram.

10. Followings are the temperature recorded of last 60 days. Draw the frequency polygon, frequency curve and histogram of the data:

Temperature (°C)	5-10	10-15	15-20	20-25	25-30	30-35
No. of days	6	12	15	17	8	2

11. Draw the 'less than' and 'more than' ogives from the data below:

Marks	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	8	12	20	6	4	10

4.3 Diagrammatic representation of data

Diagrams are one of the easiest and most commonly used devices of presenting numerical data of any field. Diagram helps biostatisticians to visualize the meaning of numerical complex at a glance. Diagrams are easiest method to understand the quantitative data for non mathematical persons too. In diagrammatic analysis two types of diagrams are commonly used.

- (a) Bar diagram
- (b) Pie-Chart (Angular Diagram).

(a) Bar diagram

Bar diagram are also known as one dimensional diagram in which height of the bars represents the magnitude of the data. Bar diagram consists of set of rectangles of equal width and equal spacing between rectangles the following points should be kept in mind to draw bar diagram.

1. All the bars used on a study must have equal width.
2. Uniform spacing should be given between bars to make the diagram look more attractive and elegant.
3. The height of the rectangles (bars) must represent the frequency or the magnitude of the observation.
4. All the bars should be constructed on the same base line.
5. The corresponding figure (number) to each bar can be written at the top of each bar.
6. In practical, vertical bars are in use but it can be constructed in horizontal direction.

Types of bar diagram

- a) Simple bar diagram.
- b) Sub divided or component bar diagram.
- c) Percentage bar diagram.
- d) Multiple bar diagram.

a) Simple bar diagram:

It is one of the simplest and most popular diagrams. It consists of a set of equidistant rectangles of equal width. The height of each rectangle represents the given value of variable. It is useful when single variable and its values are given.

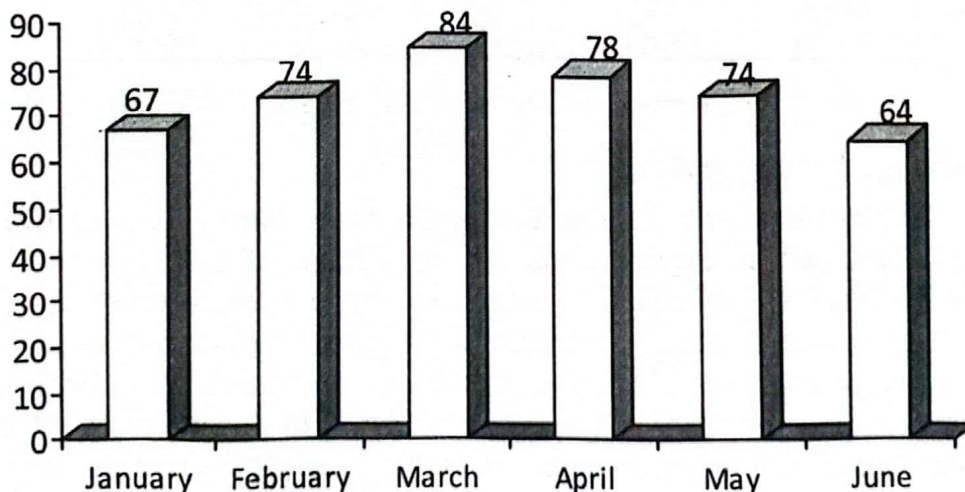
Example 1.

Oxygen consumption in cc/kg/h in 6 months in a species of fish was obtained as below. Draw a simple bar diagram.

Month	January	February	March	April	May	Jun
O ₂ consumed	67	74	84	78	74	64

The simple bar diagram of the given data is as follows:

BAR DIAGRAM REPRESENTING OXYGEN CONSUMPTION OF A SPECIES OF FISH IN 6 MONTH

**b) Subdivided or component bar diagram:**

In this diagram, the frequencies are divided into different components. This diagram consists of set of rectangles of equal width, each of the rectangle is divided into various segments representing the component of total.

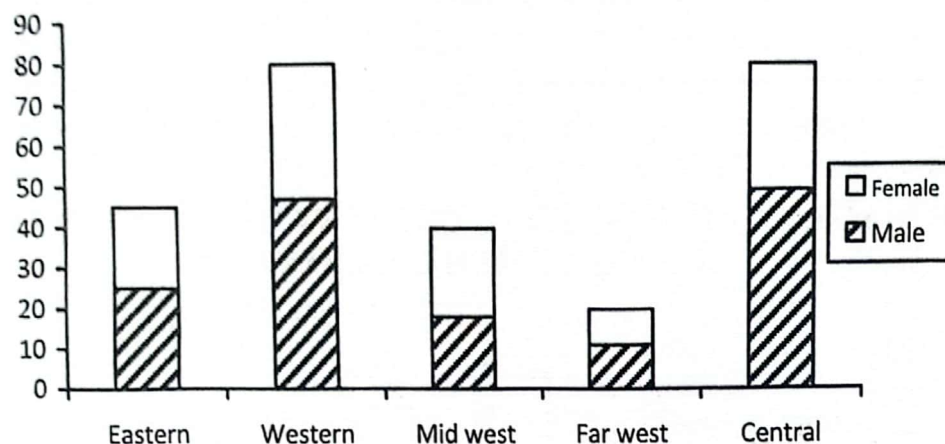
Example 2.

Draw the component bar diagram of the death of male and female in different development region in 2003 given below.

Region	No. of death ('000')		Total
	Male	Female	
Eastern	25	20	45
Western	47	33	80
Mid west	18	22	40
Far west	11	9	20
Central	49	31	80

Solution:

COMPONENT BAR DIAGRAM REPRESENTING DEATH OF MALE AND FEMALE IN 2003



c) Percentage bar diagram

In percentage bar diagram the total for each bar is taken as 100 and the value of each part is expressed as the percentage of respective totals. This diagram is useful when we have to compare the two or more set of observations. In this diagram, the height of rectangle for each set of observations is equal. To compare two or more set of observations the ratio of the breadth of rectangle is taken according as the ratios of the total of the sets. In this diagram, we calculate the percentage of each of component and we find the cumulative percentage too.

Example 3.

Draw the percentage bar diagram from the following table of populations of different towns A and B

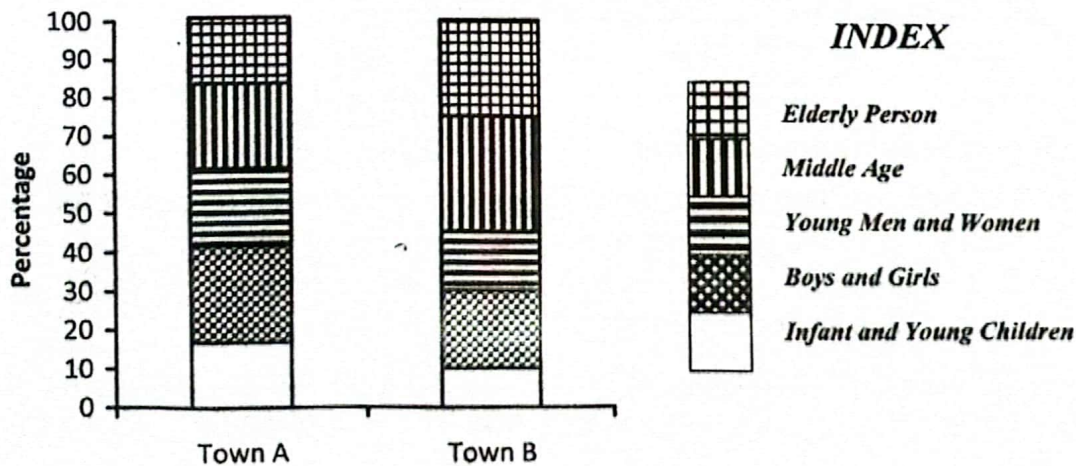
Population group	Town A	Town B
Infants and young children	10,000	4000
Boys and girls	15,000	8000
Young men and women	12,000	6000
Middle aged	13,000	12,000
Elderly persons	10,000	10,000
Total	60,000	40,000

Solution:

Calculation for percentage bar diagram

Category	Town A			Town B		
	Population	%	Cumulated %	Population	%	Cumulated %
Infants and young children	10,000	16.67	16.67	4,000	10	10
Boys and girls	15,000	25	41.67	8,000	20	30
Young men and women	12,000	20	61.67	6,000	15	45
Middle aged	13,000	21.67	83.34	12,000	30	75
Elderly persons	10,000	16.67	100.0	10,000	25	100.0
Total	60,000	100		40,000	100	

PERCENTAGE BAR DIAGRAM REPRESENTING DIFFERENT TYPES OF POPULATION IN TOWN A AND B

**d) Multiple bar diagram**

In this diagram, a set of adjacent rectangles are drawn for each of the variable other procedure are similar to that of the simple bar diagram. Proper and suitable spacing is given between different set of bars. This bar diagram is useful when the comparison between two or more variables is to be made.

Example 4.

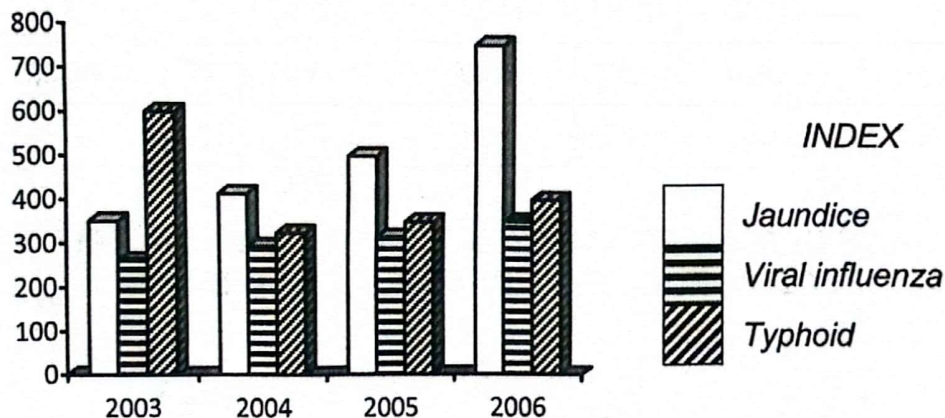
In a certain hospital, the number of patients in three different diseases in 4 years is recorded as follows.

No. of patients			
Year	Jaundice	Viral influenza	Typhoid
2003	350	260	600
2004	415	295	325
2005	500	315	350
2006	750	350	400

Draw multiple bar diagram.

Solution:

MULTIPLE BAR DIAGRAM REPRESENTING CASES OF THREE DISEASES IN 4 DIFFERENT YEARS



2. Pie diagram or Angular diagram

It is a type of two dimensional diagram in which area represent the total value. In pie diagram, we construct a circle whose area represent the total value of variable. The circle is then divided in to different sectors by the radial lines such that area of each sector represents the value of components.

In pie chart, area of sector denotes the frequency of the variable. It helps to study the comparative difference at a glance. The size of angle is obtained by multiplying the class percentage by 3.6 i.e. by $\frac{360}{100}$ or we use the formula

$$\text{Size of angle} = \frac{\text{Class frequency}}{\text{Total observations}} \times 360$$

Example 5.

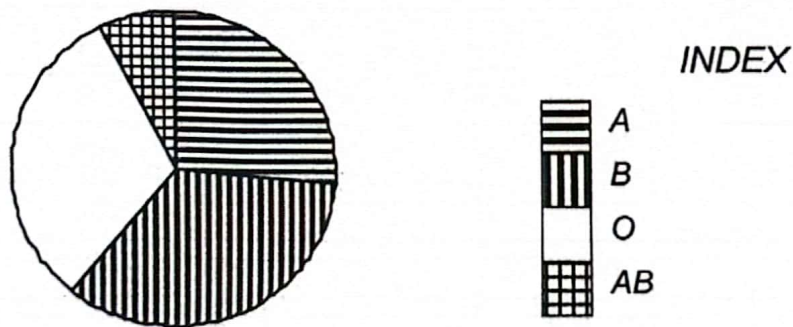
In a study of blood groups in 2810 people it was found that 744 people are of group A, 971 are of group B, 888 are of group O and 207 are of group AB. Represent these data in pie chart.

Solution:

Let us make the following table:

Group	No. of persons	Angle (in degree)
A	744	$\frac{744}{2810} \times 360^\circ = 95.4^\circ$
B	971	$\frac{971}{2810} \times 360^\circ = 124.2^\circ$
O	888	$\frac{888}{2810} \times 360^\circ = 113.8^\circ$
AB	207	$\frac{207}{2810} \times 360^\circ = 26.6^\circ$
	2810	360°

PIE DIAGRAM REPRESENTING DISTRIBUTION OF BLOOD GROUPS



We have to draw two circles when there are two sets of data and their total and component values are to be compared. The areas of the two circles must be in proportion to the total magnitudes. As the area of a circle is πr^2 , the radii of the two circles must be in the ratio of the square roots of their radii. Two circles of suitable radii are drawn and they are divided into sectors to represent the different components.

Example 6.

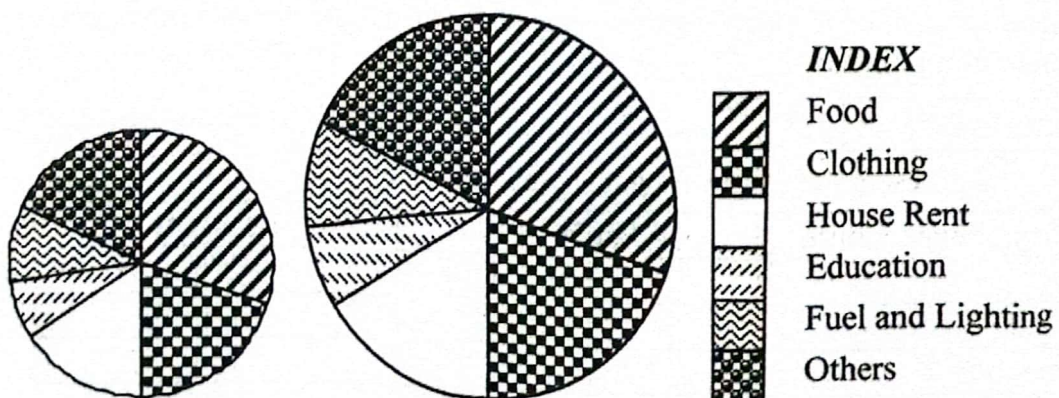
Represent the following data by pie-diagram.

Items of expenditure	Family A	Family B
Food	140	270
Clothing	100	180
House Rent	70	144
Education	40	63
Fuel and Lighting	50	81
Others	80	162
Total	480	900

Solution:

Calculation of Angle of Pie diagram

Items of expenditure	Family A		Family B	
	Rs.	Degrees	Rs.	Degrees
Food	140	$\frac{360}{480} \times 140 = 105$	270	$\frac{360}{900} \times 270 = 108$
Clothing	100	$\frac{360}{480} \times 100 = 75$	180	$\frac{360}{900} \times 180 = 72$
House Rent	70	$\frac{360}{480} \times 70 = 52.5$	144	$\frac{360}{900} \times 144 = 57.6$
Education	40	$\frac{360}{480} \times 40 = 30$	63	$\frac{360}{900} \times 63 = 25.2$
Fuel and Lighting	50	$\frac{360}{480} \times 50 = 37.5$	81	$\frac{360}{900} \times 81 = 32.4$
Others	80	$\frac{360}{480} \times 80 = 60$	162	$\frac{360}{900} \times 162 = 64.8$
Total	480	360	900	360
Square Root	21.9		30	
Radii of Circles	2.2		3	

EXPENDITURES OF TWO FAMILIES

Pie diagrams may be drawn in the same way when three or more sets of data are to be represented.

4.4 Steam Leaf Display

Steam leaf display has mainly two meaningful words as 'steam' and 'leaf'. The first digit (or digits) in a number may be taken as steam and that of the second digit (or digits) are listed in order in a column in the display table. The leaf values are recorded for each steam in a row to the right of the stem. Steam leaf display is an important statistical tool of organizing a set of data. It has a vital role to understand how the values are distributed and cluster over the range of observations in the data set. The basic idea of display is based on analogy of plants. Further we discussed in examples as below.

Example 1.

Display the steam and leaf of the following marks obtained by 40 student's out of 100

16	18	19	19	21	26	28	15	20	32
33	35	35	39	45	47	47	49	52	52
52	58	58	59	59	63	62	67	67	71
73	75	75	75	71	77	79	81	82	94

Solution:

Display table of steam and leaf

Steam	Leaf	Frequency
1	5, 6, 8, 9, 9	5
2	0, 1, 6, 8	4
3	2, 3, 5, 5, 9	5
4	5, 7, 7, 9	4
5	2, 2, 2, 8, 8, 9, 9	7
6	2, 3, 7, 7	4
7	1, 1, 3, 5, 5, 5, 7, 9	8
8	1, 2	2
9	4	1
		$\Sigma f = N = 40$

Example 2.

Construct the steam leaf display table of the lives of 45 bulbs of a particular type in an industrial application.

Bulbs lives in years

3.1	2.6	2.3	2.4	3.0	3.1	3.7
4.1	5.2	2.8	4.9	5.6	4.1	5.4
6.1	4.0	4.3	4.5	3.7	6.1	2.3
4.5	4.9	5.6	4.3	3.9	3.2	5.0
4.8	3.7	1.8	5.1	4.8	3.0	5.1
3.3	6.3	5.0	4.4	4.7	5.3	5.6
3.4	2.6	3.6	4.5	6.2	2.5	5.7

Solution:

The first digit before the decimal is chosen as 'stem' which represent the main magnitude. And the digit after the decimal is chosen as the leaf on its corresponding stem. The decimal point is not usually written in the display table. The number of 'leaf' on each 'stem' is written on the column of frequency as show below.

Steam and leaf display table

Steam	Leaf	Frequency
1	8	1
2	6, 3, 4, 8, 3, 6, 5	7
3	1, 0, 1, 7, 7, 9, 2, 7, 0, 3, 4, 6	12
4	1, 9, 1, 0, 3, 5, 5, 9, 3, 8, 8, 4, 7, 5	14
5	2, 6, 4, 6, 0, 1, 1, 0, 3, 6, 7	11
6	1, 1, 3, 2	4
		$\Sigma f = N = 49$

Notes:

1. The numbers in the column of leaf are listed as the visual indication of relative numbers. We can notice whether the distribution is approximately symmetrical and we may get a preliminary indication of any particular theoretical distribution may fit the data. Again, we may want to short the leaves on each stem in order of magnitude to give more detail and facilitate finding parameters which depend on the order.
2. In another method of displaying the number of stems is to double the number of stems. Especially if the number of data is large in relation to the initial number of stems then for steam 'a' the leaves can be taken from 0 to 4 and for stem 'b' can be taken from 5 to 9. From the above example 2 the result without sorting is shown as below.

Stem Leaf Display Table (Plotted with Double Leaf)

Stem	Leaf	Frequency
1b	8	1
2a	3, 4, 3	3
2b	6, 8, 6, 5	4
3a	1, 0, 1, 2, 0, 3, 4	7
3b	7, 7, 9, 7, 6	5
4a	1, 1, 0, 3, 3, 4	6
4b	9, 5, 5, 9, 8, 8, 7, 5	8
5a	2, 4, 0, 1, 1, 0, 3	7
5b	6, 6, 6, 7	4
6a	1, 1, 3, 2	4
		$\Sigma f = N = 49$

In this case, we might double the number of stems and sort the leaves on each stem. Again, in the next case it might be more appropriate to show two significant figures on each leaf, with appropriate separation between leaves.

Significance of Diagrams and Graphs

Diagrams and graphs are extremely useful because of the following reasons:

- They give a bird's eye view of the entire data and, therefore, the information presented can be easily understood. It is a fact that as the number and magnitude of figures increases they become more confusing and their analysis tends to be more strenuous. Pictorial presentation helps in proper understanding of the data as it gives an interesting form to it. The old saying 'A picture is worth 10,000 words' is very true. The mind through the eye can more readily appreciate the significance of figures in the form of pictures than it can follow the figures themselves.
- They are attractive to the eye. Figures are dry but diagrams delight the eye. For this reason diagrams create greater interest than cold figures. Thus, while going through journals and newspapers, the readers generally skip over the figures but most of them do look at the diagrams and graphs. Since diagrams have attraction value, they are very popular in exhibitions, fairs, conferences, board meetings and public functions.

- (c) They have a great memorizing effect. The impressions created by diagrams last much longer than those created by the figures presented in a tabular form.
- (d) They facilitate comparison of data relating to different periods of time of different regions. Diagrams help one in making quick and accurate comparison of data. They bring out hidden fact and relationship and can stimulate as well as aid analytical thinking and investigation.

4.5 Limitations of Diagram and Graphs

Although diagrams and graphs are powerful and effective media for presenting statistical data, they are not under all circumstances and for all purposes complete substitute for tabular and other forms of presentation. The well trained specialist in this field is one who recognize not only the advantages but also the limitations of these techniques. He knows when to use and when not to use these methods and from its repertoire is able to select the most appropriate form for every purpose. Julin has beautifully said, "Graphic statistics has a role to play of its own; it is not the servant of numerical statistics, but it cannot pretend, on the other hand, to precede or displace the latter".

The main limitations of diagrams and graphs are

- (a) They can present only approximate values.
- (b) They can approximately represent only limited amount of information.
- (c) They are intended mostly to explain quantitative facts to the general public. From the point of view of the statistician, they are not of much help in analyzing data.
- (d) They can be easily mis-interpreted and can be used for grinding one's axe during advertisement, propaganda and electioneering. As such as diagrams should never be accepted without a close inspection of the bonafides because things are very often not what they appear to be.
- (e) The two-dimensional diagrams and the three-dimensional diagrams cannot be accurately appraised visually, and, therefore, as far as possible their use should be avoided.

4.6 Difference between Diagram and Graph

The difference between diagram and graph are as below:

- (a) In diagrams the data are presented by bars, rectangles, circles etc but in graphs they are presented only by lines.
- (b) The diagram may be one, two or three dimensional but the graph is always two dimensional.

- (c) The construction of graph is easier than the diagram.
- (d) Diagrams are constructed on the plane paper but the graphs must be constructed on graph paper.
- (e) Diagrams are only used for comparisons but graphs are also used to find the relationship between the variables.

Exercise - 4(II) /

1. Write about the diagrammatic representation of data and its types in brief.
2. Write short notes about:
 - (a) Bar diagram
 - (b) Pie- chart
 - (c) Steam leaf display
3. Write some points (or rules) that should be followed on constructing a bar-diagram and write its types in brief.
4. What are the merits of a diagrammatic representation of data?
5. What are the limitations of diagrammatic and graphic representation of data?
6. What are the differences between graphic and diagrammatic representation of data?
7. Present the following data by a bar diagram:

Year	2034	2035	2036	2037	2038
Sales (100)	7	8	10	12.5	16

8. Draw subdivided bar diagram

College	Arts	Science	Commerce	Agriculture	Total
A	1200	800	600	400	3,000
B	750	500	300	450	2,000

9. Draw pie chart from the following

District	Illam	Kavre	Nuwakot	Morang	Jhapa
Milk production (in 1000 liter)	800	750	550	400	500

10. Represent the following data by a multiple bar diagram:

Years	1 st div	2 nd div	3 rd div	Failed
2038	30	80	160	70
2039	40	100	210	90
2040	50	130	230	80

11. Draw the percentage bar diagram to represent the following data

Item	Expenditure	
	Family A	Family B
Food	400	480
Clothing	200	400
House Rent	160	200
fuel	80	120
others	160	400
Total	1,000	1,600

12. The densities of population of two towns A and B during 3 census periods are tabulated as below. Draw a sub-divided bar diagram to represent the data.

Year	Town A	Town B
1991	418	636
2001	470	700
2011	544	782

13. Draw the multiple bar-diagram of the following data:
Production (in nearest '0000' m tone)

District	Maize	Wheat	Potato
Tanahu	21	11	14
Kaski	27	25	14
Gorkha	19	15	5
Lamjung	14	12	4
Manang	6	8	3

14. The expenditure of two families are listed below in a table. Represent the data by a pie-chart (angular diagram):

Items	Family A	Family B
Food	20000	30000
Rent	10000	20000
Clothing	1000	1500
Education	5000	2000
Energy	2500	5000
Miscellaneous	2500	1000

15. Display the stem and leaf of the following:

(a) Marks obtained by 50 students out of 100

26	22	25	15	14	37	11	12	13	11
30	32	31	35	42	48	38	31	32	44
45	61	65	51	55	52	51	41	45	46
72	70	75	77	61	65	52	55	76	72
92	90	92	96	85	86	82	85	89	92

(b) Temperature (degree °C) of 30 days in hilly region of Nepal

01	02	03	02	05	03	15	17	12	15
22	23	27	21	23	11	12	17	15	19
07	06	10	11	25	26	30	29	28	19

16. Construct the steam leaf display table of the height of 40 students (of classes 1 to 10) in feet.

2.3	3.2	2.5	2.2	5.5	5.6	2.6	2.3	6.2	1.5
1.7	1.8	2.7	4.2	4.5	4.2	6.0	5.1	4.5	4.6
3.1	3.1	3.2	3.7	3.2	3.5	3.4	4.2	4.6	4.3
6.2	6.4	6.2	5.5	5.2	4.1	4.6	3.2	2.7	5.8

Answers

15. (a) Display table of steam and leaf

Steam	Leaf	Frequency (f)
1	5, 4, 1, 2, 3, 1	6
2	6, 2, 5	3
3	7, 0, 2, 1, 5, 8, 1, 2	8
4	2, 8, 4, 5, 1, 5, 6	7
5	1, 5, 2, 1, 2, 5	6
6	1, 5, 1, 5	4
7	2, 0, 5, 7, 6, 2	6
8	5, 6, 2, 5, 9	5
9	2, 0, 2, 6, 2	5
		$\Sigma f = 50$

(b) Display table of steam and leaf

Steam	Leaf	Frequency (f)
0	1, 2, 3, 2, 5, 3, 7, 6	8
1	5, 7, 2, 5, 1, 2, 7, 5, 9, 0, 1, 9	12
2	2, 3, 7, 1, 3, 5, 6, 9, 8	9
3	0	1
		$\Sigma f = 30$

11. Steam and leaf display table

Steam	Leaf	Frequency (f)
1	5, 7, 8	3
2	3, 5, 2, 6, 3, 7, 7	7
3	2, 1, 1, 2, 7, 2, 5, 4, 2	9
4	2, 5, 2, 5, 6, 2, 6, 3, 1, 6	10
5	5, 6, 1, 5, 2, 8	6
6	2, 0, 2, 4, 2	5
		$\Sigma f = 40$



Summarizing Data Set

5.1 Introduction

Central tendency or averages are the measures which condense a huge mass of numerical data into single numerical values which is the representative of entire distribution. According to **Croxton and Crowden**, "An average value is a single value with the range of data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is sometimes called measure of central value."

From this definition we can conclude that averages are the typical values around which other items of the distribution concentrate. They are the values which lie between two extreme observations, (i.e. smallest and largest) of the distribution of the values in the central part of the distribution.

The averages are sometimes called measure of location or measure of central tendency.

5.2 Requisites of an Ideal Measure of Central Tendency

The following are requirements to be satisfied by an ideal average or central tendency.

- (a) It should be rigidly defined. In other words, the definition should be clear and unambiguous so that it leads to one and only one interpretation by different user.
- (b) It should be simple to understand and easy to calculate. Even for a non mathematical person it should readily understandable and should be computed with sufficient ease and rapidity. It should not involve heavy arithmetical calculation.
- (c) It should be based on all the observation. The ideal measure of central tendency should use all types of data which ensures that there is no loss of any information resulting from not using the available data.
- (d) It should suitable for further mathematical treatment. Which means the average should possess some important and

interesting mathematical properties so that its use in further statistical theory is enhanced.

- (e) It should be least affected by fluctuation of sampling. It means if we take independent random samples of same size from a given population and compute the average for each of these samples then for an ideal average, the values so obtained from different samples should not vary much from one another.
- (f) It should not be affected much by extreme observation. It means that a few very small or very large observations should not much affect the value of an ideal measure of central tendency.

5.3 Measure of Central Tendency

The following are the measure of central tendency which are commonly used in practice.

1. Mean
2. Median
3. Mode

1. Mean

Mean is a measure of central value of given set of observations. It has three types:

- (a) Arithmetic mean
- (b) Geometric mean
- (c) Harmonic mean

(a) Arithmetic mean:

The arithmetic mean or simply mean of a given set of observation is their sum divided by the number of observations. The arithmetic mean is generally denoted by $\bar{X}$. Thus if $X_1, X_2 \dots X_n$ is the given set of observations, then their arithmetic mean is given by

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum X}{n}$$

Where, $\sum X$ = the sum of observations.

n = the number of observations.

Computation of arithmetic mean

The following are the formula to calculate arithmetic means in different series.

Individual series

Let $X_1, X_2, X_3 \dots X_n$ be n individual observations. Then their mean is given by

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\Sigma X}{n}$$

This method is known as direct method.

Example 1.

A random sample of 10 boys had the following intelligence quotient (I.Q.)

70, 120, 110, 101, 88, 83, 95, 98, 107, 100

Find the mean I.Q.

Solution:

The mean I.Q. is given by

$$\begin{aligned} &= \frac{\Sigma X}{n} = \frac{1}{10} [70 + 120 + 110 + 101 + 88 + 83 + 95 + 98 + 107 + 100] \\ &= \frac{972}{10} = 97.2 \end{aligned}$$

Short cut method

This method is generally used when the size of variate value is very big. In this method we find arithmetic mean by taking deviations of the items from any arbitrary number. So the arithmetic mean is given by

$$\bar{X} = A + \frac{\Sigma d}{n}$$

where A = assumed mean

$$d = X - A$$

Example 2.

Find arithmetic mean by short cut method.

Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Tem.(°C)	17	18	19	15	14	20	24

Solution:

Day	Temperature (X)	d = X - 18 (A)
Sunday	17	-1
Monday	18	0
Tuesday	19	1
Wednesday	15	-3
Thursday	14	-4
Friday	20	2
Saturday	24	6
		$\Sigma d = 1$

$$\text{Now, mean } (\bar{X}) = A + \frac{\Sigma d}{n} = 18 + \frac{1}{7} = 18.14$$

Discrete series

Let n variate values $X_1, X_2, \dots, X_n$ have the following frequencies $f_1, f_2, \dots, f_n$ respectively i.e.

x	X_1	X_2	X_3	...	X_n
f	f_1	f_2	f_3	...	f_n

Direct method

Then the arithmetic mean is given by

$$\bar{X} = \frac{X_1 \cdot f_1 + X_2 \cdot f_2 + \dots + X_n \cdot f_n}{f_1 + f_2 + \dots + f_n} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{N}$$

where $N = \Sigma f$ is the total frequency

This method of calculating arithmetic mean is known as direct method.

Example 3.

Calculate the average bonus paid per member from the following data by direct method.

Bonus	50	60	70	80	90	100	110
No. of persons	1	3	5	7	6	2	1

Solution:

Let us make the frequency distribution table from the given data.

CALCULATION OF MEAN

Bonus (X)	No. of persons (f)	$f \times X$
50	1	50
60	3	180
70	5	350
80	7	560
90	6	540
100	2	200
110	1	110
	$\Sigma f = 25$	$\Sigma fX = 1990$

From the table, $\Sigma f = 25$ $\Sigma fX = 1990$

$$\text{Now, mean } (\bar{X}) = \frac{\sum fX}{\sum f} = \frac{1990}{25} = 79.60$$

Indirect method (short cut method)

In this method the deviation of items will be taken from assumed mean A. $\bar{X}$ will be calculated by following formula.

$$\bar{X} = A + \frac{\sum fd}{N} \text{ where } d = X - A \text{ and } N = \sum f.$$

This method is useful if values of X are in decimal or large

Example 4.

Calculate the arithmetic mean by indirect method for the following observations

Weight of patients (in kg)	20	22	23	24	25	27
No. of patients	5	7	11	9	3	2

Let us make the frequency distribution table supposing the assumed mean as $A = 23$.

CALCULATION OF MEAN

Weight (X)	No. of patients (f)	$d = X - 23$	fd
20	5	$20 - 23 = -3$	-15
22	7	$22 - 23 = -1$	-7
23	11	$23 - 23 = 0$	0
24	9	$24 - 23 = 1$	9
25	3	$25 - 23 = 2$	6
27	2	$27 - 23 = 4$	8
	$\Sigma f = 37$		$\Sigma fd = 1$

Now,

$$\bar{X} = A + \frac{\sum fd}{N} = 23 + \frac{1}{37} = 23.02$$

Continuous series

To calculate the arithmetic mean in continuous series we must calculate the midvalue of each class and denote it by X. Other process are similar to the case of discrete series. We have three methods of calculating arithmetic mean in continuous series.

Direct method

Let $X_1, X_2, X_3, \dots, X_n$ be the mid value of n classes having frequencies $f_1, f_2, f_3, \dots, f_n$ respectively, then arithmetic mean is given by

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \dots + f_nX_n}{f_1 + f_2 + \dots + f_n}$$

$$\text{i.e. } \bar{X} = \frac{\sum fX}{N}$$

Where X = mid value of each class

$N = \sum f$ = total frequency

Example 5.

Calculate the mean from the following frequency distribution:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	6	5	8	15	7	6	3

Solution:

CALCULATION OF MEAN

Marks	Mid value (X)	No. of students (f)	fX
0-10	5	6	30
10-20	15	5	75
20-30	25	8	200
30-40	35	15	525
40-50	45	7	315
50-60	55	6	330
60-70	65	3	195
		$\sum f = 50$	$\sum fX = 1670$

$$\therefore \text{Mean } (\bar{X}) = \frac{\sum fX}{N} = \frac{1670}{50} = 33.4$$

Indirect method (short-cut method)

In this method, the arithmetic mean is calculated by the following formula:

$$\bar{X} = A + \frac{\sum fd}{N}$$

Where $d = X - A$; A = assumed mean

$N = \Sigma f = \text{total frequency}$

This method is mainly useful when the value of X are large as well as in decimal.

Example 6.

Find the arithmetic mean from the following frequency distribution

Marks	10-15	15-20	20-25	25-30	30-35	35-40	40-45
No. of students	3	2	7	11	9	8	6

Solution:

CALCULATION OF MEAN

Marks	Mid-value (X)	No. of students (f)	$d = X - A$ $= X - 27.5$	fd
10-15	12.5	3	-15	-45
15-20	17.5	2	-10	-20
20-25	22.5	7	-5	-35
25-30	27.5	11	0	0
30-35	32.5	9	5	45
35-40	37.5	8	10	80
40-45	42.5	6	15	90
		$\Sigma f = 46$		$\Sigma fd = 115$

Now,

$$\begin{aligned}\bar{X} &= A + \frac{\Sigma fd}{N} \\ &= 27.5 + \frac{115}{46} = 27.5 + 2.5 \\ &= 30\end{aligned}$$

Step deviation method

In this method, the arithmetic mean is calculated by

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times h$$

Where, $d' = \frac{X - A}{h}$ and $h = \text{size of the interval}$.

This method is used when the value of X is very large as well as the value is given in decimal.

Example 7.

Find the arithmetic mean from the following frequency distribution.

Pressure	130-135	135-140	140-145	145-150	150-155	155-160
No. of patients	5	3	7	13	11	2

Solution:

CALCULATION OF MEAN

Pressure	mid value (X)	No. of patients (f)	$d' = \frac{X - 147.5}{5}$	fd'
130-135	132.5	5	-3	-15
135-140	137.5	3	-2	-6
140-145	142.5	7	-1	-7
145-150	147.5	13	0	0
150-155	152.5	11	1	11
155-160	157.5	2	2	4
		$\Sigma f = 41$		$\Sigma fd' = -13$

$$\begin{aligned}\text{Now, mean } (\bar{X}) &= A + \frac{\Sigma fd'}{\Sigma f} \times h = 147.5 + \left(\frac{-13}{41} \right) \times 5 \\ &= 147.5 - 0.32 \times 5 = 147.5 - 1.60 = 145.9\end{aligned}$$

Properties of arithmetic mean**Theorem: 1**

The algebraic sum of the deviations of the given set of observations from their arithmetic mean is zero i.e. $\Sigma f(X - \bar{X}) = 0$.

Proof

Let X is the variable and $\bar{X}$ is the arithmetic mean of the frequency distribution.

Now, the sum of the deviations taken from arithmetic mean is

$$= \Sigma f(X - \bar{X})$$

$$= \Sigma fX - \Sigma f \bar{X} \quad \left[\because \frac{\Sigma fX}{N} = \bar{X} \right]$$

$$= N \bar{X} - N \bar{X} = 0$$

$$\text{Hence, } \Sigma f(X - \bar{X}) = 0$$

It is noted that $\Sigma(X - \bar{X}) = 0$ by the same argument.

Theorem: 2

The sum of squares of the deviations of the items is minimum when taken from the arithmetic mean. i.e., $\sum f(X - A)^2$ is minimum when $A = \bar{X}$.

Proof:

Let A is the arbitrary number from which deviations are to be taken and let S be the sum of squares of the deviations taken from A .

$$\text{i.e., } S = \sum f(X - A)^2$$

Differentiating both sides with respect to A we get,

$$\frac{dS}{dA} = -2 \sum f(X - A).$$

Again differentiating both sides with respect to A we get

$$\frac{d}{dA} \left(\frac{dS}{dA} \right) = \frac{d^2S}{dA^2} = 2\sum f = 2N$$

For maxima and minima $\frac{dS}{dA} = 0$

$$\text{or, } -2\sum f(X - A) = 0$$

$$\text{or, } \sum fX = \sum fA$$

$$\text{or, } A = \frac{\sum fX}{\sum f} = \bar{X}$$

$$\text{At } A = \bar{X}.$$

$$\frac{d^2S}{dA^2} = 2N \text{ which is positive.}$$

Hence, S is minimum when $A = \bar{X}$

Notes:

- The above result can also be stated as $\sum f(X - \bar{X})^2 < \sum f(X - A)^2$
- The above result is also true where $S = \sum (X - A)^2$

Combined mean

If $\bar{X}_1$ and $\bar{X}_2$ are the arithmetic means of two set of observations with their respective total frequencies N_1 and N_2 . Then the combined mean of the two sets of observations is denoted by $\bar{X}_{12}$ and given by

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

Example 8.

In a class of 20 students, the average age of 12 boys is 15 years and the average age of 8 girls is 17 years. Find the average age of the class.

Solution:

Number of boys $N_1 = 12$

Average age of boys $\bar{X}_1 = 15$

Numbers of girls $N_2 = 8$

Average age of girls $\bar{X}_2 = 17$

Combined mean age $\bar{X}_{12} = ?$

By formula

$$\begin{aligned}\bar{X}_{12} &= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} = \frac{12 \times 15 + 8 \times 17}{12 + 8} \\ &= \frac{180 + 136}{20} = \frac{316}{20} = 15.8 \text{ years}\end{aligned}$$

Weighted arithmetic mean

The formula discussed so far for computing the arithmetic mean are based on the assumption that all the items in the distribution are of equal importance. However, in practice, all the items under consideration may not have an equal importance. If some items of the distribution are more important than others then this point must be taken in mind in order to compute representative mean of the distribution. In such cases, proper weights are to be given to various items such that the weight attached to each items being proportional to the importance of the item in the distribution.

Let $W_1, W_2, W_3, \dots, W_n$ are weights attached to $X_1, X_2, X_3, \dots, X_n$ respectively. Then the weighted arithmetic mean is denoted by $\bar{X}_w$ and is given by

$$\bar{X}_w = \frac{W_1 X_1 + W_2 X_2 + \dots + W_n X_n}{W_1 + W_2 + \dots + W_n} = \frac{\sum WX}{\sum W}$$

Example 9.

A student got 75 in Nepali, 60 in English, 63 in mathematics, 59 in physics, 55 in chemistry. If the weights given to these subjects are 1, 1, 3, 2 and 2 respectively, find the weighted arithmetic mean.

Solution:

Let the variable X denotes the marks of corresponding subjects and W denotes the weight of corresponding subjects.

CALCULATION OF WEIGHTED ARITHMETIC MEAN

Subject	Marks (X)	Weights (w)	W. X
Nepali	75	1	75
English	60	1	60
Maths	63	3	189
Physics	59	2	118
Chemistry	55	2	110
		$\Sigma W = 9$	$\Sigma WX = 552$

$$\text{Now, } \bar{X}_w = \frac{\Sigma WX}{\Sigma W} = \frac{552}{9} = 61.33$$

Merits and demerits of arithmetic mean**Merits**

1. It is simple to understand and easy to calculate.
2. It is rigidly defined.
3. It is based on all the observations.
4. It is suitable for further mathematical treatment.
5. It is least affected by fluctuation of sampling.

Demerits

1. It is affected by extreme observations, two or three very large value of variable unduly affect the value of the arithmetic mean.
2. It cannot be determined by inspection nor can it be located graphically.
3. It cannot be computed accurately in case of open end classes.
4. It is not useful if we are dealing with qualitative characteristics such as intelligence, honesty, beauty etc.

(b) Geometric mean

The geometric mean of a set of n non-zero and non-negative observation is the n^{th} root of their product. Thus, if $X_1, X_2, X_3, \dots, X_n$ are the n observations then their G.M is given by

$$\text{G.M.} = \sqrt[n]{X_1 \cdot X_2 \dots X_n} = (X_1 \cdot X_2 \cdot X_3 \dots X_n)^{\frac{1}{n}}$$

Calculation of geometric mean**Individual series**

Let $X_1, X_2, \dots, X_n$ be n non-zero and non-negative variate values. Then their geometric mean G.M. is given by

$$G = \sqrt[n]{X_1 \cdot X_2 \dots X_n} \quad \text{or, } G = \text{antilog} \left[\frac{\sum \log X}{n} \right]$$

Where n is the number of observations.

Example 10.

Find the geometric mean of 24, 25, 26, 27, 28, 29.

Solution:

The given data can be written as

CALCULATION OF GEOMETRIC MEAN

X	log X
24	1.3802
25	1.3979
26	1.4149
27	1.4313
28	1.4471
29	1.4623
	$\Sigma \log X = 8.5340$

$$\begin{aligned} \text{Now, } G.M. &= \text{antilog} \left(\frac{\sum \log X}{n} \right) = \text{antilog} \left(\frac{8.5340}{6} \right) \\ &= \text{antilog} (1.4223) = 26.444 \end{aligned}$$

Discrete series

If $f_1, f_2, f_3, \dots, f_n$ be the corresponding frequencies of ' n ' non zero and non-negative variate values $X_1, X_2, \dots, X_n$ then their geometric mean is given by

$$G = \text{antilog} \left[\frac{\sum f \log X}{N} \right]$$

Where N is the total frequency.

Example 11.

Find the geometric mean of the following frequency distribution:

Marks (X)	3	4	5	6	7	8
No. of students (f)	2	7	3	4	9	5

Solution:

The given distribution can be written as

CALCULATION OF GEOMETRIC MEAN

X	f	log X	f. log X
3	2	0.4771	0.9542
4	7	0.6020	4.2144
5	3	0.6989	2.0969
6	4	0.7781	3.1126
7	9	0.8450	7.6058
8	5	0.9030	4.5154
			$\Sigma f \log X = 22.5515$

Now,

$$\begin{aligned} \text{G.M.} &= \text{antilog} \left(\frac{\Sigma f \log X}{N} \right) = \text{antilog} \left(\frac{22.5515}{30} \right) \\ &= \text{antilog} (0.7517) = 3.7299 \end{aligned}$$

Continuous series

For the continuous series we use the same formula for discrete series, where X denotes the mid-value of each class. Thus for continuous series

$$\text{G.M.} = \text{Antilog} \left(\frac{\Sigma f \log X}{N} \right)$$

Where X is the mid-value of classes.

Merits and demerits of G.M.**Merits**

1. It is rigidly defined.
2. It is based on all the observation.
3. It is suitable for further mathematical treatment.
4. It is least affected by extreme observation.
5. It is not affected much by fluctuation of sampling.

Demerits

1. It is not simple to understand and not easy to calculate.
2. If any one observation is non positive then we can not find the value of geometric mean.

(c) Harmonic mean

The harmonic mean of n non-zero observations is defined as the reciprocal of arithmetic mean of the reciprocal of the observations. If X_1, X_2

..., X_n is a given set of n non-zero observations, then their harmonic mean is given by

$$\text{H.M.} = \frac{n}{\Sigma(\frac{1}{X})}$$

This gives the harmonic mean of individual series. For discrete and continuous series we use

$$\text{H.M.} = \frac{N}{\Sigma f \frac{1}{X}}$$

Where $N = \Sigma f$ = total frequency.

It is noted that in case of continuous series X denotes the mid value of class. The merits and demerits of harmonic mean is very much similar to that of the geometric mean.

Example 12.

Find Harmonic mean of weights of 31 persons in a sample inquiry.

Weight (X)	130	135	140	145	146	148	149	150	157
No. of persons (f)	3	4	6	6	3	5	2	1	1

Solution:

Making the frequency distribution table from given values

CALCULATION OF HARMONIC MEAN

Weight (X)	No. of persons (f)	$\frac{1}{X}$	$f \cdot \frac{1}{X}$
130	3	0.00769	0.02307
135	4	0.00741	0.02964
140	6	0.00714	0.04284
145	6	0.00690	0.04140
146	3	0.00685	0.02055
148	5	0.00676	0.03380
149	2	0.00671	0.01342
150	1	0.00667	0.00667
157	1	0.00637	0.00627
	$\Sigma f = 31$		$\Sigma f \left(\frac{1}{X} \right) = 0.21776$

$$\text{Now, H.M.} = \frac{N}{\sum f\left(\frac{1}{X}\right)} = \frac{31}{0.21776} = 142.36$$

Hence the mean weight is 142.36.

2. Median

The median is that value of variable which divides the whole set of observations in two equal halves. One half containing all the values greater than median and the other half containing all the values less than median. Thus, the number of observations above median is equal to the number of observations below the median. So the median is known as the positional average.

Calculation of median

Individual series

In this series, we first arrange the data into ascending or descending order of magnitude and use the following formula:

$$\text{Median} = \left(\frac{n+1}{2} \right)^{\text{th}} \text{ item}$$

If numbers of terms are odd then the middle term of the series is taken for median. If the number of terms are even, the average of two middle terms is taken as median.

Example 13.

Calculate median from following series:

- (a) 35, 37, 34, 29, 39
- (b) 7, 9, 3, 11, 5, 14

Solution:

- (a) Arranging the given data in increasing order, we get,

29, 34, 35, 37, 39

Here, total number of object (n) = 5

$$\text{So, median} = \left(\frac{n+1}{2} \right)^{\text{th}} \text{ item} = \left(\frac{6}{2} \right)^{\text{th}} = 3^{\text{rd}} \text{ item}$$

$\therefore$ Median = 35

- (b) Arranging the given data in increasing order we get,
3, 5, 7, 9, 11, 14

Here, total number of objects (n) = 6

$$\begin{aligned}
 \text{Median} &= \left(\frac{n+1}{2} \right)^{\text{th}} \text{ item} = \left(\frac{7}{2} \right)^{\text{th}} = 3.5^{\text{th}} \text{ item} \\
 &= \text{average of } 3^{\text{rd}} \text{ and } 4^{\text{th}} \text{ item} \\
 &= \frac{7+9}{2} = 8
 \end{aligned}$$

$$\therefore \text{Median} = 8$$

Discrete series

In this method, we first arrange the given data either in ascending or in descending order of magnitude and construct the less than cumulative frequency table. Then we use the formula $\left(\frac{N+1}{2} \right)^{\text{th}}$ item to calculate the median value. To find the median value see cumulative frequency table and note the value corresponding to the cumulative frequency which is equal or just greater than $\left(\frac{N+1}{2} \right)$.

Example 14.

Find the median value from the data given below:

Marks	10	13	19	11	7
No. of students	3	5	2	6	8

Solution:

Making the frequency distribution table by arranging the data in increasing order.

CALCULATION OF MEDIAN

Marks (X)	Frequency (f)	C.f.
7	8	8
10	3	11
11	6	17
13	5	22
19	2	24
	N = 24	

$$\text{Median} = \left(\frac{N+1}{2} \right)^{\text{th}} \text{ item} = \left(\frac{25}{2} \right)^{\text{th}} = 12.5^{\text{th}} \text{ item} = 13^{\text{th}} \text{ item}$$

Here, the cumulative frequency (c.f.) just greater than 13 is 17 and the value of X corresponding to 17 is 11.

Hence median = 11

Continuous series

If the data are given in continuous series, then we first arrange the classes in increasing order and construct less than cumulative frequency distribution table. Then we calculate the value of $\frac{N}{2}$ and see the table to note the value corresponding to cumulative frequency which is equal or just greater than $\frac{N}{2}$. This will give the median class. To find the median value we use the following formula:

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times h$$

Where, L = lower limit of median class

N = total frequency

cf = cumulative frequency previous to the median class

f = frequency of median class

h = class interval of median class

Example 15.

Find the median from the data given below:

Class	10-20	20-30	30-40	40-50	50-60
Frequency	2	5	8	3	2

Solution:

Making cumulative frequency distribution table we get,

CALCULATION OF MEDIAN

Class (X)	Frequency (f)	C.f.
10-20	2	2
20-30	5	7
30-40	8	15
40-50	3	18
50-60	2	20
	$N = 20$	

$$\text{Here, } \frac{N}{2} = \frac{20}{2} = 10$$

The cumulative frequency (c.f.) just greater than 10 is 15 and the corresponding class is 30–40.

$\therefore$ median lies in the interval 30 – 40.

$$\begin{aligned}\text{Now, median} &= L + \frac{\frac{N}{2} - cf}{f} \times h = 30 + \frac{10 - 7}{8} \times 10 \\ &= 30 + \frac{3}{8} \times 10 = 30 + 3.75\end{aligned}$$

$$\therefore \text{Median} = 33.75$$

Determination of median by graphical method

Median can be determined graphically as mentioned in page 358. It can also be obtained by applying following method:

- i) Draw less than ogive curve.
- ii) Take the variable on the x-axis and frequency on the y-axis.
- iii) Determine the median value by the formula.

$$\text{median} = \text{size of } \frac{N}{2} \text{ th item}$$

- iv) Locate this value on the y-axis and from it draw a perpendicular on the cumulative frequency curve.
- v) From the point where it meets the ogive draw another perpendicular on the x-axis and the point where it meets the x-axis is the median.

It should be noted that the other partition values like quartiles, deciles and percentile can also be determined graphically by this method.

Example 16.

Find median graphically from example 15.

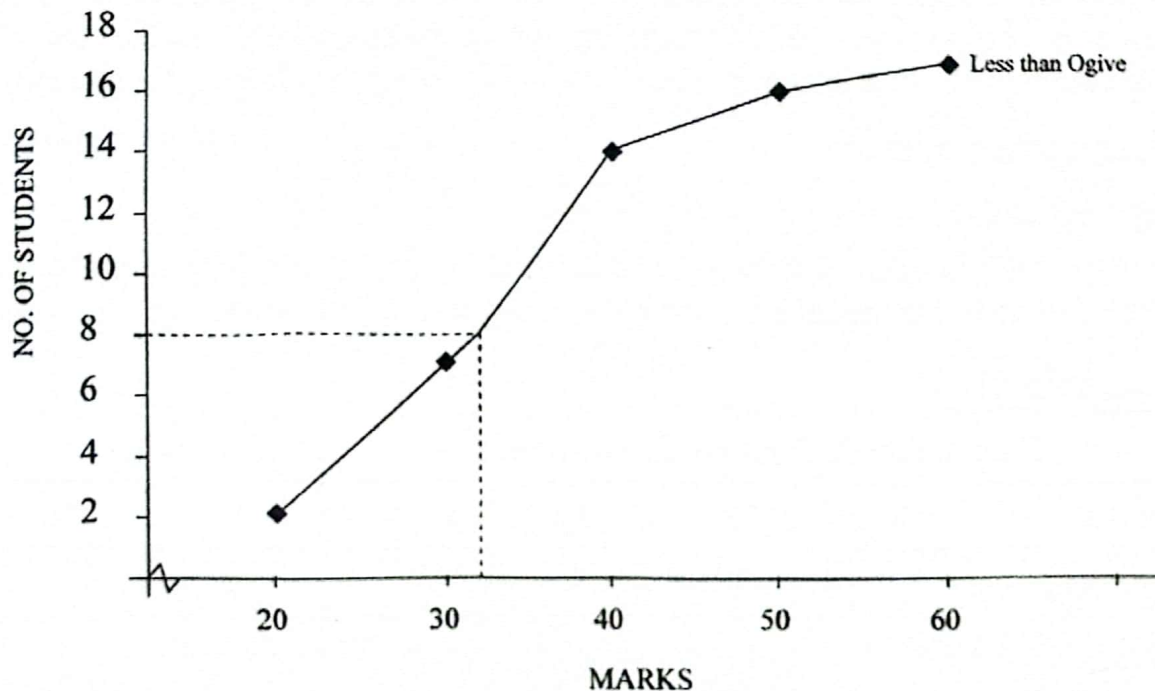
Solution:

First we make less than cumulative frequency table:

Marks	No. of students
Less than 20	2
Less than 30	7
Less than 40	14
Less than 50	16
Less than 60	17

$$\begin{aligned}\text{Median} &= \text{size of } \frac{N}{2} \text{ th item} = \frac{17}{2} \text{ th item} \\ &= 8.5 \text{ th item}\end{aligned}$$

Locate this value on the y-axis and from it draw a perpendicular on the curve. From this point draw a perpendicular line on the x-axis and the point where it meets the x-axis gives the median value.



From the graph it is clear that the median value is 32.5. It should be noted that the value obtained graphically will be the same as obtained algebraically except for errors in plotting and reading the scale.

Merits and demerits of median

Merits

- 1) It is rigidly defined.
- 2) It is simple to understand and easy to calculate.
- 3) It is not affected by extreme observations.
- 4) It can be computed in open end classes.
- 5) It is useful when dealing with qualitative characteristics such as honesty, intelligence etc.
- 6) It can be computed graphically. (The number corresponding to intersection of less than ogive and more than ogive curve).

Demerits

- 1) It is not based on all the observations.
- 2) Arrangement of data in increasing or decreasing order is necessary.

- 3) For ungrouped data (individual series) if the number of observations is even then median can not be computed exactly.
- 4) Median is not suitable for further mathematical treatment.
- 5) It is more affected by fluctuation of sampling.

3. Mode

Mode is the value which occurs maximum number of times in a set of observations and around which other items of series cluster densely. Thus mode is the variate value which occurs most frequently in the set of observations. So, mode is that value of variable which has largest frequency.

Computation of mode

Discrete series

In individual series, there is no direct formula to calculate mode. In discrete series, the mode is the variate value with highest frequency.

Example 17.

Find the mode from the following series:

Marks	11	12	13	14	15	16
Frequency	7	3	9	8	2	1

Solution:

Here the value of variable which has largest frequency is 13. Thus mode = 13.

Continuous Series

In case of the continuous series, the class corresponding to the maximum frequency is called modal class. The value of mode is calculated by using following formula:

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where, L = lower limit of modal class

f_1 = frequency of modal class

f_0 = frequency of class preceding the modal class

f_2 = frequency of class succeeding the modal class

h = interval of the modal class

The above formula can be written as

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

Where, $\Delta_1 = f_1 - f_0$

$\Delta_2 = f_1 - f_2$

Example 18.

Find the mode of following frequency distribution:

Class	10-20	20-30	30-40	40-50	50-60
Frequency	3	7	11	9	2

Solution:

Here the class having highest frequency is 30 – 40

$$\text{So, } f_1 = 11, f_0 = 7, f_2 = 9, h = 10, L = 30$$

We know that

$$\begin{aligned} \text{Mode} &= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h \\ &= 30 + \frac{11 - 7}{22 - 7 - 9} \times 10 \\ &= 30 + \frac{4}{6} \times 10 \\ &= 30 + 6.66 = 36.66 \end{aligned}$$

Note:

Sometimes the value of $2f_1 - f_0 - f_2 = 0$. In this case the value of mode is not definite. So, in this situation, we use the following formula.

$$\text{Mode} = L + \frac{|f_1 - f_0|}{|f_1 - f_2| + |f_1 - f_0|} \times h$$

The above techniques of locating mode is not practicable if

- The maximum frequency is repeated or approximately equal concentration is found in two or more neighbouring classes.
- If the maximum frequency occurs either in very beginning or at the end of the distribution.
- If there are irregularities in the distribution.

In such cases mode is calculated by the **method of grouping** which is explained below.

In this method, we write 6 columns of frequencies. The frequencies in the first column are the original frequencies, 2nd column is obtained by combining the frequencies of column (1) combining two by two in first column. 3rd column is obtained by combining the frequencies of column (1) two by two after leaving the first frequency. Fourth column is obtained by combining three by three elements of 1st column. The fifth column is

obtained by combining three by three frequencies of 1st column after leaving 1st frequency. Similarly, column (6) is obtained by combining three by three frequencies of first column after leaving first two frequencies.

The maximum frequency in each column is represented by 'Italic Type'.

Then we construct the analysis table to locate mode. This table contains three columns. Column number, maximum frequency and value giving maximum frequency.

In analysis table, the value of variable that repeated maximum number of time in last column is mode.

Example 19.

Determine the mode from the following data:

X	20	21	22	23	24	25	26	27	28	29
f	6	9	4	2	10	8	7	5	1	3

Solution:

Here the distribution is irregular since the frequencies increased up to 9 and decreased up to 2 again go up to 10 and start decreasing. Thus, in this case the value of mode may not be 24. To locate mode we use the method of grouping as shown below.

Grouping Table

X	Frequencies					
	(1)	(2)	(3)	(4)	(5)	(6)
20	6					
21	9	} 15				
22	4	} 6	} 13	} 19	} 15	
23	2	} 18	} 12			} 16
24	10					
25	8	} 15				
26	7	} 12				
27	5	} 4				
28	1					
29	3					

Analysis Table

Column Number	Maximum Frequency	Value giving maximum frequency
1	10	24
2	18	24, 25
3	15	25, 26
4	20	23, 24, 25
5	25	24, 25, 26
6	20	25, 26, 27

Here, the value 25 is repeated maximum number of times so mode
= 25

Determination of mode by graphical method

In a frequency distribution the value of mode can also be determined graphically. The steps in calculation are

- Draw a histogram of the given data.
- Draw two lines diagonally in the inside of the modal class bar, starting from each upper corner of the bar to the upper corner of the adjacent bar.
- Draw a perpendicular line from the intersection of the two diagonal lines to the x-axis which gives the modal value.

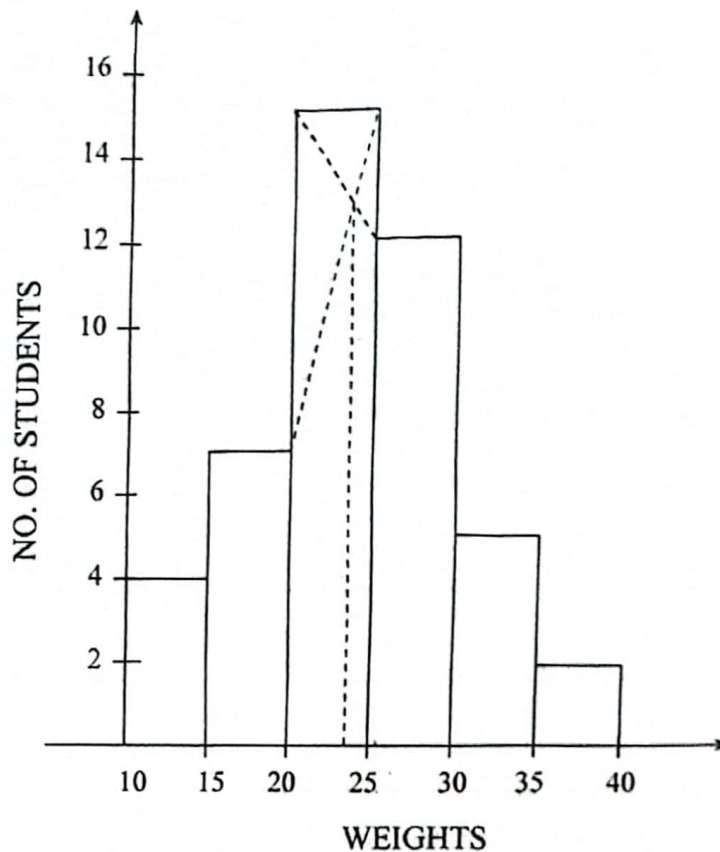
Example 20.

Determine the modal value of the following data by graphical method:

Weight (in kg)	10-15	15-20	20-25	25-30	30-35	35-40
No. of students	4	7	15	12	5	2

Solution:

First we draw histogram of the given data



It is clear from the graph that the modal value is 23.5. It should be noted that mode can be obtained graphically only where there is one class containing the highest frequency. If two or more classes have the same highest frequency, mode can not be determined graphically.

Merits and demerits of mode

Merits:

- 1) It is simple to understand and easy to calculate.
- 2) Mode is not affected by extreme observations.
- 3) It can be obtained in case of open end classes.
- 4) In some cases, it can be obtained simply inspection and by graph.
- 5) Mode can be calculated even in case of unequal class interval.

Demerits:

- 1) Mode is not rigidly defined. It is ill defined if maximum frequency is repeated or if maximum frequency lies in very beginning or very end of the distribution or if the distribution is irregular. In these cases we use method of grouping. If this method gives two values of mode, the distribution is called bimodal.
- 2) It is not based on all the observation.
- 3) It is not suitable for further mathematical treatment.
- 4) Mode is much affected by fluctuation of sampling.

5.4 Relation among Three Central Tendencies

In some distribution the value of mean, median and mode are equal to each other. This type of distribution is called symmetrical distribution. If mean, median and mode are not mutually equal then the distribution is known as asymmetrical or skewed distribution. In this case, median usually lies between mean and mode. In such case, we have the following relationship.

$$\text{Mean} - \text{mode} = 3(\text{mean} - \text{median})$$

$$\therefore \text{Mode} = 3 \text{ median} - 2 \text{ mean}$$

This relation is called empirical relationship. It was introduced by **Karl Pearson**. This formula is very useful to locate mode when the distribution has two or more largest frequencies. If any two of mean, median and mode are known then the third can be obtained easily by this relation.

Notes:

The skewed (asymmetrical) distribution are of two types

- (i) If $\text{mean} > \text{median} > \text{mode}$ then the distribution is called positively skewed.
- (ii) If $\text{mean} < \text{median} < \text{mode}$ then the distribution is called negatively skewed.

Example 21.

What will be the mode of a distribution, if $\text{mean} = 32$ and $\text{median} = 30$?

Solution:

$$\text{We have, } \bar{x} = 32, M_d = 30$$

$$\begin{aligned} \text{Using } M_0 &= 3M_d - 2\bar{x} \\ &= 3 \times 30 - 2 \times 32 \\ &= 90 - 64 \\ &= 26 \end{aligned}$$

$$\text{Hence, Mode} = 26$$

5.5 Partition Values

The values which divide the series into a number of equal parts are called the partition values. Thus we can consider median as a type of partition values. Here we discuss on Quartiles, Deciles and Percentiles.

1. Quartiles

The values which divide the given data into four equal parts is known as quartiles. There are three quartiles denoted by Q_1 , Q_2 and Q_3 such that $Q_1 \leq Q_2 \leq Q_3$.

Computation of quartiles

1) For individual series

$$Q_i = \text{value of } \left(\frac{i(n+1)}{4} \right)^{\text{th}} \text{ item}$$

where $i = 1, 2, 3$

For example,

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

2) For discrete series

$$Q_i = \text{value of } \left(\frac{N+1}{4} \right)^{\text{th}} \text{ item}$$

$i = 1, 2, 3$

For example,

$$Q_3 = \text{value of } \frac{3(N+1)}{4} \text{ item}$$

3) For continuous series (Exclusive Continuous)

$$Q_i = \text{lies in the class containing } \left(\frac{iN}{4} \right)^{\text{th}} \text{ item}$$

For quartile value, we use the formula

$$Q_i = L + \frac{\frac{iN}{4} - cf}{f} \times h$$

Particularly,

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times h$$

Where, $N = \sum f$

L = lower limit of corresponding class

cf = cumulative frequency of previous class

f = frequency of corresponding class

h = length of the interval

Example 22.

Find Q_1 and Q_3 from the following data:

(a) 30, 20, 25, 32, 22, 35, 26 (b) 16, 20, 25, 15, 21, 12, 19, 24

Solution:

(a) Arranging the given data in ascending order,

20, 22, 25, 26, 30, 32, 35

$\therefore$ Number of observation (n) = 7

$$\begin{aligned}\text{Now, } Q_1 &= \text{Value of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item} \\ &= \text{Value of } \left(\frac{7+1}{4}\right)^{\text{th}} \\ &= \text{Value of } 2^{\text{nd}} \text{ item} \\ &= 22\end{aligned}$$

$$\begin{aligned}\text{Also, } Q_3 &= \text{Value of } \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{ item} \\ &= \text{Value of } \left(\frac{3(7+1)}{4}\right)^{\text{th}} \\ &= \text{Value of } 6^{\text{th}} \text{ item} \\ &= 32\end{aligned}$$

Hence, $Q_1 = 22$, $Q_3 = 32$

(b) Arranging the given data in ascending order,

12, 15, 16, 19, 20, 21, 24, 25

$\therefore$ Number of observation (n) = 8

$$\begin{aligned}\text{Now, } Q_1 &= \text{Value of } \left(\frac{n+1}{4}\right)^{\text{th}} \text{ item} \\ &= \text{Value of } \left(\frac{8+1}{4}\right)^{\text{th}} \\ &= \text{Value of } 2.25^{\text{th}} \text{ item} \\ &= 2^{\text{nd}} \text{ item} + 0.25 (3^{\text{rd}} \text{ item} - 2^{\text{nd}} \text{ item}) \\ &= 15 + 0.25 (16 - 15) \\ &= 15 + 0.25 \times 1 \\ &= 15.25\end{aligned}$$

$$\text{Also, } Q_3 = \text{Value of } \left(\frac{3(n+1)}{4}\right)^{\text{th}} \text{ item}$$

$$\begin{aligned}
 &= \text{Value of } \left(\frac{3 \times 9}{4}\right)^{\text{th}} \text{ item} \\
 &= \text{Value of } 6.75^{\text{th}} \text{ item} \\
 &= 6^{\text{th}} \text{ item} + 0.75 (7^{\text{th}} - 6^{\text{th}}) \text{ item} \\
 &= 21 + 0.75 (24 - 21) \\
 &= 21 + 0.75 \times 3 \\
 &= 23.25
 \end{aligned}$$

$$\text{Hence, } Q_1 = 15.25, Q_3 = 23.25$$

Example 23.

Find quartiles from following data:

X	13	14	15	16	17	18	19
f	2	3	7	9	3	5	1

Solution:

Making the frequency distribution table we get

X	f	C.f.
13	2	2
14	3	5
15	7	12
16	9	21
17	3	24
18	5	29
19	1	30
N = 30		

$$\begin{aligned}
 Q_1 &= \text{value of } \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item} = \text{value of } \left(\frac{31}{4}\right)^{\text{th}} \text{ item} \\
 &= \text{value } 7.75^{\text{th}} \text{ item} = 15
 \end{aligned}$$

$$\begin{aligned}
 Q_2 &= \text{value of } 2\left(\frac{N+1}{4}\right)^{\text{th}} \text{ item} = \text{value of } \left(2 \times \frac{31}{4}\right)^{\text{th}} \text{ item} \\
 &= \text{value } 15.5^{\text{th}} \text{ item} = 16
 \end{aligned}$$

$$\begin{aligned}
 Q_3 &= \text{value of } 3\left(\frac{N+1}{4}\right)^{\text{th}} \text{ item} \\
 &= \text{value of } \left(3 \times \frac{31}{4}\right)^{\text{th}} \text{ item}
 \end{aligned}$$

$$= \text{value } 23.25^{\text{th}} \text{ item}$$

$$= 17$$

Example 24.

Find quartiles from following data:

Class	0-10	10-20	20-30	30-40	40-50
frequency	2	7	9	11	3

Solution:

Making the frequency distribution table

Class (X)	Frequency (f)	c.f.
0-10	2	2
10-20	7	9
20-30	9	18
30-40	11	29
40-50	3	32
	N = 32	

$$\text{Here, } \frac{N}{4} = \frac{32}{4} = 8$$

The c.f. just greater than 8 is 9 and the corresponding class is 10-20.
i.e., Q_1 lies in 10-20

$$\begin{aligned} \therefore Q_1 &= L + \frac{\frac{N}{4} - cf}{f} \times h \\ &= 10 + \frac{8-2}{7} \times 10 \\ &= 10 + \frac{6}{7} \times 10 = 18.57 \end{aligned}$$

$$\text{Again, } \frac{2N}{4} = \frac{2(32)}{4} = 16$$

The c.f. just greater than 16 is 18 and the corresponding class is 20-30.
i.e., Q_2 lies in 20-30

$$\begin{aligned} \text{Thus, } \therefore Q_2 &= L + \frac{\frac{2N}{4} - cf}{f} \times h \\ &= 20 + \frac{16-9}{7} \times 10 \\ &= 20 + \frac{7}{7} \times 10 = 27.77 \end{aligned}$$

$$\text{Finally, } \frac{3N}{4} = \frac{3(32)}{4} = 24$$

The c.f. just greater than 24 is 29 and the corresponding class is 30-40, i.e., Q_3 lies in 30-40

Thus,

$$\begin{aligned}\therefore Q_3 &= L + \frac{\frac{3N}{4} - cf}{f} \times h \\ &= 30 + \frac{24 - 18}{11} \times 10 \\ &= 30 + \frac{6}{11} \times 10 = 35.45\end{aligned}$$

2. Deciles

Deciles are the values which divide the series into ten equal parts. There are nine deciles namely $D_1, D_2 \dots D_9$ such that

$$D_1 \leq D_2 \leq D_3 \leq \dots \leq D_9$$

Computation of deciles

For deciles we make less than frequency distribution table.

Individual series

$$D_j = \text{value of } \left[\frac{j(n+1)}{10} \right]^{\text{th}} \text{ item}$$

Where $j = 1, 2, 3, \dots, 9$.

Discrete series

$$D_j = \text{value of } \frac{j(N+1)}{10}^{\text{th}} \text{ item}$$

Where $j = 1, 2, \dots, 9$.

Continuous series

In continuous series, we first find the class where $D_j, j = 1, 2, \dots, n$ lies. For this we use

$$D_j \text{ lies in containing } \left(\frac{j \times N}{10} \right)^{\text{th}} \text{ item}$$

After locating the class, we use the following formula to calculate deciles.

$$D_j = L + \frac{\frac{j \times N}{10} - cf}{f} \times h$$

Where, $j = 1, 2, \dots, 9$.

L = lower limit of deciles class

$N = \sum f$

cf = cumulative frequency of the class previous to deciles class

f = frequency of deciles class

h = length of interval of deciles class.

3. Percentiles

Percentiles are the values which divide the series in to 100 equal parts. There are 99 percentiles namely $P_1, P_2, \dots, P_{99}$ such that

$$P_1 \leq P_2 \leq P_3 \leq \dots \leq P_{99}.$$

Computation of percentiles

For percentiles we must make less than cumulative frequency distribution table. Then we use the following formula:

Individual series

$$P_k = \text{value of } \frac{k \times (n+1)}{100}^{\text{th}} \text{ item}$$

$$k = 1, 2, \dots, 99$$

Discrete series

$$P_k = \text{value of } \left[\frac{k(N+1)}{100} \right]^{\text{th}} \text{ item}$$

Where, $k = 1, 2, \dots, 99$.

Continuous series

In this series we first find the class where P_k lies.

For this we use the following formula:

$$P_k \text{ lies in containing } \left[\frac{k \times N}{100} \right]^{\text{th}} \text{ item.}$$

After locating the class where P_k lies, we use the following formula:

$$P_k = L + \frac{\frac{k \times N}{100} - cf}{f} \times h$$

L = lower limit of percentile class

$N = \sum f$

cf = cumulative frequency of the class previous to percentile class

f = frequency of to percentile class

h = length of interval of percentile class.

Note: The median, Q_2 , D_5 and P_{50} are equal.

Example 25.

Find the 5th decile and 30th percentile of the following data:

Height in cm	148	143	160	152	157	150	155
No. of children	7	23	20	13	27	9	1

Solution:

Making the frequency distribution table by arranging the data in ascending order. We get,

X	Frequency (f)	cf
143	23	23
148	7	30
150	9	39
152	13	52
155	1	53
157	27	80
160	20	100
	N = 100	

$$\begin{aligned}
 \text{Now, 5}^{\text{th}} \text{ deciles} &= \text{value of } \left(\frac{5(n+1)}{10} \right)^{\text{th}} \text{ item} \\
 &= \text{value of } \left(\frac{5(100+1)}{10} \right)^{\text{th}} \text{ item} \\
 &= \text{value of } (50.5)^{\text{th}} \text{ item}
 \end{aligned}$$

The cf equal or greater than 50.5 is 52 and the corresponding value of 52 is 152.

$$\therefore D_5 = 152$$

Again, 30th percentile

$$\begin{aligned}
 P_{30} &= \text{value of } \left(\frac{30 \times (n+1)}{100} \right)^{\text{th}} \text{ item} \\
 &= \text{value of } \left(\frac{30 \times 101}{100} \right)^{\text{th}} \text{ item} \\
 &= \text{value of } (30.3)^{\text{th}} \text{ item}
 \end{aligned}$$

The cf equal or greater than 30.3 is 39 and the corresponding value of 39 is 150.

$$\therefore P_{30} = 150$$

Example 26.

Find the 7th decile and 55th percentile from the following distribution:

Weight in kg	60-62	62-64	64-66	66-68	68-70	70-72	72-74
No. of patients	10	20	45	50	60	40	15

Solution:

Making less than cumulative frequency distribution table we get,

Class	Frequency (f)	C.f.
60-62	10	10
62-64	20	30
64-66	45	75
66-68	50	125
68-70	60	185
70-72	40	225
72-74	15	240
	N = 240	

$$\text{Here, for } D_7, \frac{7 \times N}{10} = \frac{7 \times 240}{10} = 168$$

The c.f. just greater than 168 is 185 and the corresponding class 68 - 70, i.e. D_7 lies in 68-70 class.

$$\begin{aligned} \text{Now, } D_7 &= L + \frac{\frac{7 \times N}{10} - cf}{f} \times h = 68 + \frac{168 - 125}{60} \times 2 \\ &= 68 + 1.433 = 69.33 \end{aligned}$$

$$\text{Also for } P_{55}, \frac{55 \times N}{100} = \frac{55 \times 240}{100} = 132.$$

The c.f. just greater than 132 is 185 and the corresponding class is 68 - 70 i.e. P_{55} lies in 68-70.

$$\begin{aligned} \text{Now, } P_{55} &= L + \frac{\frac{55 \times N}{100} - cf}{f} \times h = 68 + \frac{132 - 125}{60} \times 2 \\ &= 68 + 0.23 = 68.23. \end{aligned}$$

Example 27.

Find the missing frequency of the following distribution if the arithmetic mean is 4.5

X	2	3	4	5	6	7	8
f	2	4	6	—	3	2	1

Solution:

Let the missing frequency is f_1 , making the frequency distribution table.

X	Frequency (f)	fx
2	2	4
3	4	12
4	6	24
5	f_1	$5f_1$
6	3	18
7	2	15
8	1	8
	$\Sigma f = 18 + f_1$	$\Sigma fx = 80 + 5f_1$

$$\text{Here, } \bar{X} = 4.5$$

$$\Sigma f = 18 + f_1$$

$$\Sigma fX = 80 + 5f_1$$

$$\text{We know, } \bar{X} = \frac{\Sigma fx}{\Sigma f}$$

$$\text{or, } 4.5 = \frac{18 + 5f_1}{18 + f_1}$$

$$\text{or, } 80 + 4.5f_1 = 81 + 5f_1$$

$$\text{or, } 5f_1 - 4.5f_1 = 81 - 80$$

$$\text{or, } 0.5f_1 = 1$$

$$\therefore f_1 = \frac{1}{0.5} = 2$$

Thus the missing frequency is 2.

Example 28.

If average height of all the 51 plants is 40 cm. Find the missing frequencies corresponding to the height 30 and 50. [CTEVT 073 R/B]

Height (cm)	10	20	30	40	50	60
No. of plants	2	3	—	21	—	5

Solution:

Suppose missing frequencies are f_1 and f_2 corresponding to 30 and 50. The frequency distribution table is

Heights (X)	Frequency (f)	fX
10	2	20
20	3	60
30	f_1	$30f_1$
40	21	840
50	f_2	$50f_2$
60	5	300
	$\Sigma f = 31 + f_1 + f_2$	$\Sigma fX = 1120 + 30f_1 + 50f_2$

Here, we have given that

$$N = \Sigma f = 51$$

From the table we have,

$$N = \Sigma f = 31 + f_1 + f_2$$

$$\text{or, } 51 = 31 + f_1 + f_2$$

$$\text{or, } 51 - 31 = f_1 + f_2$$

$$\therefore f_1 + f_2 = 20$$

$$\text{i.e. } f_2 = 20 - f_1 \quad (1)$$

Again, we have, $\bar{X} = 40$

$$\text{i.e., } \frac{\Sigma fX}{N} = 40$$

$$\text{or, } \frac{1120 + 30f_1 + 50f_2}{51} = 40$$

$$\text{or, } 1120 + 30f_1 + 50f_2 = 2040$$

$$\text{or, } 30f_1 + 50f_2 = 2040 - 1120 = 920$$

$$\text{or, } 3f_1 + 5f_2 = 92 \quad (2)$$

Using (1) in (2) we get,

$$3f_1 + 5(20 - f_2) = 82$$

$$\text{or, } 3f_1 + 100 - 5f_2 = 82$$

$$\begin{aligned} \text{or, } -2f_1 &= -100 + 82 \\ &= -18 \end{aligned}$$

$$\text{i.e., } f_1 = \frac{-18}{-2} = 9$$

$$\therefore f_1 = 9$$

$$\text{Again, from (1), } f_2 = 20 - f_1 = 20 - 9 = 11$$

$\therefore$ The missing frequencies are $f_1 = 9$ and $f_2 = 11$.

Example 29.

Calculate the missing frequencies from the following table if the median is 50 and $N = 100$.

Class	0-20	20-40	40-60	60-80	80-100
frequency	14	—	27	—	15

Solution:

Let the missing frequencies are f_1 and f_2 .

Now, making the frequency distribution we get

Class	Frequency (f)	C.f.
0-20	14	14
20-40	f_1	$14 + f_1$
40-60	27	$41 + f_1$
60-80	f_2	$41 + f_1 + f_2$
80-100	15	$56 + f_1 + f_2$
	$100 = 56 + f_1 + f_2$	

From the table we have,

$$N = 56 + f_1 + f_2$$

$$100 = 56 + f_1 + f_2$$

$$\therefore f_1 + f_2 = 100 - 56 = 44 \quad (1)$$

Since median is given as 50 so, it lies in the class 40-60.

So, the median class is 40-60.

$$\text{Now, Median} = L + \frac{\frac{N}{2} - cf}{f} \times h$$

$$\text{or, } 50 = 40 + \frac{50 - (14 + f_1)}{27} \times 20$$

$$\text{or, } 50 - 40 = \frac{20}{27} (50 - 14 - f_1)$$

$$\text{or, } \frac{10 \times 27}{20} = 36 - f_1$$

$$\text{or, } \frac{27}{2} = 36 - f_1$$

$$\text{or, } 27 = 72 - 2f_1$$

$$\text{or, } 2f_1 = 72 - 27$$

$$f_1 = \frac{45}{2} = 22.5 = 23$$

Also, we have from (1)

$$f_1 + f_2 = 44$$

$$\therefore f_2 = 44 - f_1 = 44 - 23 = 21$$

Thus the missing frequencies are $f_1 = 23$, $f_2 = 21$.

EXERCISE - 5(I)

1. Find the arithmetic mean from the data:
 - (a) 15, 13, 27, 20, 25
 - (b) 50, 57, 63, 72, 88
2. Hemoglobin percentage (Hb%) of 9 patients suffering from AIDS was recorded as 5.2mg, 5.3mg, 5.6mg, 5.7mg, 5.4mg, 5.2mg, 5.3mg, 5.3mg and 5.2mg. Find out the mean Hb%.
3. Find the mean incubation (the time between somebody being infected with a disease and the appearance of the first symptoms) of 9 polio cases are given below:

23, 22, 20, 24, 16, 17, 18, 19, 21.
4. Find the median of the series
 - (a) 5, 3, 2, 4, 10, 7, 12
 - (b) 4, 2, 6, 1, 9
 - (c) 40, 55, 37, 60, 12, 96, 25
 - (d) 18, 14, 31, 22, 16, 9, 12, 37

5. a) Oxygen consumption of 9 fishes in cc/100ml/hour was recorded as: 35, 44, 38, 36, 39, 40, 42, 41, 45. Find the median from given data.
 b) In the blood samples of 7 patients, the Hb% was recorded as 7, 13, 22, 16, 10, 18, 14. Find the median value.

6. Find the mode value of

(a) 3, 4, 2, 5, 4, 8, 3, 4

(b) 16, 14, 10, 14, 11, 12, 14, 18, 20

7. If $\bar{x}$ and M_d are the mean and median for a certain distribution. Write the expression for mode.

8. In a certain distribution has mode 3.2 and median 3.3 then find mean.

9. If mean and mode of a given data are 7.5 and 6 respectively. Find the median value.

10. If mean and median of given data are 3.5 and 4.5. Find M_o .

11. Calculate the first quartile for the data: 3, 7, 4, 10, 8, 5, 2.

12. Calculate the Q_3 for the data: 12, 8, 13, 16, 20, 17, 22.

13. (a) Find the mean days of confinement after delivery in the following series.

Days of confinement	6	7	8	9	10
No. of patients	5	4	4	3	2

(b) Find the average of temperature of the following information.

Temp ($^{\circ}\text{C}$)	21	23	25	27	29
No. of days	7	11	16	14	2

(c) Find the arithmetic mean of the following distribution.

Height in cms (x)	10	20	30	40	50	60
No. of plants (f)	2	3	9	21	11	5

14. (a) Calculate the missing frequency of the following distribution given that A.M. is equal to 19.42.

Marks	5	10	15	20	25	30
No. of students	2	3	?	8	5	4

(b) Given the A.M. is equal to 7.3. Find the missing frequency in the following data:

x	5	6	7	8	9
f	4	6	12	?	8

15. Calculate geometric mean from the following distribution.

(a)

X	2	3	4	5	6	7	8
f	2	4	6	2	3	2	1

(b)

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	5	7	15	25	8

16. Calculate the harmonic mean from the following distributions:

(a)

Weight (kg)	30	35	40	45	46	48	49	50	57
No. of persons	3	4	5	6	3	5	2	1	1

(b)

Class interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	12	18	27	20	17	6

17. Find the median of the data:

(a) 7, 18, 55, 33, 67, 41, 29, 73

(b) 23, 71, 55, 36, 43, 56, 29, 40

18. (a) The distribution of age of males at the time of marriage was as follows:

Age (years)	18-22	22-26	26-30
No. of males	23	65	46

Calculate the median age of marriage.

- (b) Calculate the median from the data given below:

x	10-20	20-30	30-40	40-50	50-60
f	2	5	7	2	1

- (c) Obtain the median and modal marks of the following data:

Marks	20	30	50	70	60	95	80
No. of students	4	5	8	10	12	5	6

19. (a) Calculate the modal marks for the given data:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	7	12	24	10	7

- (b) Find the mode from the given data:

Age	0-10	10-15	15-20	20-30	30-40
No. of people	5	8	12	7	3

20. (a) Calculate the average daily wages for the workers of two factories:

	Factory A	Factory B
No. of wage earners	350	300
Average daily wage	Rs. 2.00	Rs. 2.50

- (b) There are 60 diabetes patients in a Hospital of which 25 are women. The average weight of 25 women is 40kg and that of 35 men is 53kg. Find the mean weight in kg of the all patients.

21. Compute the weighted arithmetic mean of the index number from the data below:

Group	Index no.	Weight
Food	125	7
Clothing	133	5
Fuel and Light	141	4
Miscellaneous	182	3
House rent	173	1

22. Calculate median, first quartile and eighty-fifth percentile from the following data of incomes.

Income in '00 Rs.	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of families	75	250	350	192	68	35	24	6

23. Amend the following table and locate the median from the following data:

Size	10-16	16-17.5	17.5-20	20-30	30-35	35-40	40-50
frequency	10	15	17	25	28	30	40

24. The following data related to age of staffs in a hospital and the number of days they were reported sick in a month

Age	30	32	35	40	48	50	52
Sick days	1	0	2	5	2	4	6

Calculate mean, median and other two quartiles.

25. The median marks of 230 students is 46, find the missing frequencies of following:

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
frequency	12	30	-	65	-	25	19

26. Find the missing frequencies from the following distribution with median 413.11 and total number of frequency is 1000.

Class	300-325	325-350	350-375	375-400	400-425	425-450
frequencies	55	61	80	-	326	-

ANSWERS

1. (a) 20 (b) 66 2. 5.36 mg 3. 20
4. (a) 5 (b) 4 (c) 40 (d) 17
5. (a) 40 (b) 14
6. (a) 4 (b) 14
7. $M_o = 3M_d - 2\bar{x}$ 8. 3.35 9. 7
10. 6.5 11. 3 12. 20
13. (a) 7.61 (b) 24.72 (c) 40
14. (a) 4. (b) 10
15. (a) 4.19 (b) 25.61
16. (a) 41.65 (b) 17.437
17. (a) 37 (b) 41.5
18. (a) 24.71 (b) 32.14 (c) 70; 60
19. (a) 24.61 (b) 17.22
20. (a) 2.23 (b) 47.58 21. 141.15
22. 12.5; 8.5; 19.56 23. 32.77
24. $\bar{x} = 45.4$; $M_d = 50$; $Q_1 = 40$; $Q_3 = 52$
25. 34; 45 26. 133; 345

5.6 Variability of Dispersion

By computing only central values, we can not find the exact nature of observations because two different sets of observation may have same central values. Thus to know the exact nature of given distribution we must know how that data are scattered from their central values. Here we discuss the different methods of measuring the scatterness of data from their central values called dispersion.

The literal meaning of dispersion is scatter, deviation, fluctuation, spread or variability of data. Thus the measure of dispersion means finding

the degree to which the individual values of the variable scatter away from average or central value.

The measure of variation is one of the important tools of statistics for biological and medical sciences because biology and medical science have phenomenon that are more variable than that of other sciences. For example, height, weight, Hb%, RBCs number, rate of oxygen consumption etc of two individual of same species, of same age and sex will differ definitely. The cure rate of same drug varies in different patients, even of same age and sex.

Dispersion helps us to compare two or more series with regard to their variability and to determine the reliability of averages.

5.7 Various Measures of Variations

1. Range
2. Quartile deviation or semi inter quartile range
3. Standard deviation

Among these methods the Lorenz Curve is the graphic method of measuring variability of given distribution. In the following section we will discuss these measures in detail. Due to the beyond the scope of the course, mean deviation and Lorenz curve will not be discussed.

1) Range

The range of raw data is the difference between the maximum and minimum observations in the data. i.e.

$$\text{Range} = X_{\max} - X_{\min} = L - S$$

Where L = the largest item & S = the smallest item and $X_{\max}$ is largest observation and $X_{\min}$ is the smallest observation of the given data.

Coefficient of range

The above formula of range is an absolute measure of dispersion and contains the units of measurement. It is useful only when we compare the variability of two or more distribution with same unit of measurement. But if we have to compare the variability of distributions given in different units of measurements we can not use the above formula. In such case we need a relative measure which is independent of units of measurement. This relative measure is called coefficient of range and it is given by

$$\text{Coefficient of range} = \frac{L - S}{L + S}$$

Where L is the largest variate value and S is the smallest variate value.

Example 1.

Find the range and its coefficient from the following data:

- (a) 20, 10, 30, 100, 5, 25.

(b)	X	10	11	12	13	15
	f	3	4	17	15	13

(c)	X	0-10	10-20	20-30	30-40	40-50
	f	3	7	12	9	8

Solution:

- (a) Here, Largest item (L) = 100

Smallest item (S) = 5

$$\therefore \text{Range (R)} = L - S = 100 - 5 = 95$$

$$\text{Coeff. of range} = \frac{L - S}{L + S} = \frac{100 - 5}{100 + 5} = \frac{95}{105} = 0.94$$

- (b) Largest item (L) = 15

Smallest item (S) = 10

$$\therefore \text{Range (R)} = L - S = 15 - 10 = 5$$

$$\text{Also, coeff. of range} = \frac{L - S}{L + S} = \frac{15 - 10}{15 + 10} = \frac{05}{25} = 0.2$$

- (c) Largest item (L) = 50

Smallest item (S) = 0

$$\therefore \text{Range (R)} = L - S = 50 - 0 = 50$$

$$\text{Also, coeff. of range} = \frac{L - S}{L + S} = \frac{50 - 0}{50 + 0} = \frac{50}{50} = 1$$

Merits and demerits of range

Merits:

- 1) It is simple to understand and easy to calculate.
- 2) It takes little time to compute.
- 3) It is rigidly defined.

Demerits:

- 1) It is not based on all the observations.
- 2) Range is very much affected by fluctuation of sampling i.e. its values varies widely from sample to sample.
- 3) It can not be calculated in case of open end classes.
- 4) If any item is added or deleted from the set of observations except the smallest and largest values then the values of range is unaltered.
- 5) It is not suitable for further mathematical treatment.

2) Quartile Deviation (or Inter Quartile Range)

The half of the difference between third quartile and first quartile is called quartile deviation or semi-inter quartile range. Thus quartile deviation is given by

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

This formula for Q.D. is the absolute measure of dispersion. For comparative studies of variability of the distributions we need the relative measure which is known as coefficient of quartile deviation and it is given by

$$\text{Coefficient of Q. D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 2.

Find the Q.D. and its coefficient from the following data:

10, 31, 16, 24, 35, 14, 18, 28, 20, 32

Solution:

Arranging the given data in ascending order,

10, 14, 16, 18, 20, 24, 28, 31, 32, 35

Here, $n = 10$

Now,

$$\begin{aligned} Q_1 &= \text{Value of } \left(\frac{n+1}{4} \right)^{\text{th}} \text{ item} \\ &= \text{Value of } \left(\frac{10+1}{4} \right)^{\text{th}} \text{ item} \\ &= \text{Value of } 2.75^{\text{th}} \text{ item} \\ &= 2^{\text{nd}} \text{ item} + 0.75 (3^{\text{rd}} - 2^{\text{nd}}) \text{ item} \\ &= 14 + 0.75 (16 - 14) \\ &= 14 + 0.75 \times 2 = 14 + 1.5 \\ &= 15.5 \end{aligned}$$

$$\begin{aligned} \text{Also, } Q_3 &= \text{Value of } \left(\frac{3(n+1)}{4} \right)^{\text{th}} \text{ item} \\ &= \text{Value of } \left(\frac{3 \times 11}{4} \right)^{\text{th}} \text{ item} \\ &= \text{Value of } 8.25^{\text{th}} \text{ item} \\ &= 8^{\text{th}} \text{ item} + 0.25 (9^{\text{th}} - 8^{\text{th}}) \text{ item} \end{aligned}$$

$$\begin{aligned}
 &= 31 + 0.25(32 - 31) \\
 &= 31 + 0.25 \times 1 = 31 + 0.25 \times 1 = 31.25 \\
 \text{Now, Q.D.} &= \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(31.25 - 15.5) \\
 &= \frac{1}{2}(15.75) = 7.875 \\
 \therefore \text{Coeff. of Q. D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{31.25 - 15.5}{31.25 + 15.5} = \frac{15.75}{46.75} = 0.37
 \end{aligned}$$

Example 3.

The length of hookworms in **cm** and their frequencies was studied and given in the form of following table. Find the quartile deviation and its coefficient.

Length(cm)	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65
Frequency	4	3	8	9	14	3	3	2	2	2

Solution:

Converting the given series into exclusive continuous series and making the less than cumulative frequency distribution table.

COMPUTATION OF Q.D.

Class	Frequency	c.f.
15.5-20.5	4	4
20.5-25.5	3	7
25.5-30.5	8	15
30.5-35.5	9	24
35.5-40.5	14	38
40.5-45.5	3	41
45.5-50.5	3	44
50.5-55.5	2	46
55.5-60.5	2	48
60.5-65.5	2	50
	N = 50	

$$\text{Here, } \frac{N}{4} = \frac{50}{4} = 12.5$$

The c.f. just greater than 12.5 is 15 and the corresponding class is 25.5 - 30.5.

$$\begin{aligned}\therefore Q_1 &= 25.5 + \frac{12.5 - 7}{8} \times 5 \\ &= 25.5 + \frac{5.5 \times 5}{8} = 28.9375\end{aligned}$$

$$\text{Here, } \frac{3N}{4} = \frac{3 \times 50}{4} = 37.5$$

The c.f. just greater than 37.5 is 38 and the corresponding class is 35.5 - 40.5.

$$\begin{aligned}\therefore Q_3 &= 35.5 + \frac{37.5 - 24}{14} \times 5 = 35.5 + \frac{13.5 \times 5}{14} \\ &= 40.3214\end{aligned}$$

$$\text{Now, Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{40.32 - 28.93}{2} = 11.39 = 5.69$$

$$\text{Also, coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40.32 - 28.93}{40.32 + 28.93} = \frac{11.39}{69.25} = 0.16$$

Merits and Demerits of Q.D.

Merits

1. It is based on 50% of the observation so it is better measure of dispersion than range.
2. It is not affected by extreme observations.
3. It can be computed in case of open end classes.

Demerits

1. It is not based on all the observation. It ignores first and last 25% of the data.
2. Q.D. is affected considerably by fluctuation of sampling
3. It is not suitable for further mathematical treatment.

3) Standard Deviation

Standard deviation is defined as the positive square root of the arithmetic mean of the squares of deviations of the given observations from their arithmetic mean. We denote the standard deviation by Greek alphabet σ (small sigma). Thus, if $X_1, X_2, \dots, X_n$ is a set of n observations then the standard deviation is given by

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}} = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$$

Computation of Standard Deviation

Individual Series

Direct method:

$$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}} = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$$

Where n = total number of observations

Shortcut method:

$$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

Where $d = X - A$

A = Assumed mean

Discrete series and continuous series

Direct method:

$$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma f(X - \bar{X})^2}{N}} = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$

Where, $N = \Sigma f$

X = mid value for continuous series

Shortcut method:

In above formula, the deviations are taken from arithmetic mean. If we take the deviations from assumed mean A (say) then the formula will be as follows:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

where $d = X - A$, A is assumed mean.

Step deviation method:

If the deviations taken from assumed mean are large we use this method. Where we use the following formula.

$$\sigma = h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$$

Where, $d' = \frac{X - A}{h}$, A is the assumed mean and h is the common factor.

Merits and Demerits of Standard Deviation**Merits**

1. It is rigidly defined.
2. It is based on all the observations
3. The squaring of the deviations $(X - \bar{X})$ removes the drawback of ignoring the signs of mean deviations in computing mean deviation.
4. It is suitable for further mathematical treatment.
5. It is least affected by fluctuation of sampling.

Demerits

1. It is difficult to calculate for non mathematical person
2. It gives greater weight to extreme values.

Example 4.

Hemoglobin percent g/100 ml of liver fed wallago attu (a type of fish) was recorded as 23, 22, 20, 24, 16, 18, 19, 21 and 17. Calculate the standard deviation.

Solution:

Here, first we calculate the arithmetic mean

$$\bar{X} = \frac{\Sigma X}{n} = \frac{23 + 22 + 20 + 24 + 16 + 18 + 19 + 21 + 17}{9} = 20$$

CALCULATION OF S.D.

X	X-20	(X-20) ²
23	3	9
22	2	4
20	0	0
24	4	16
16	-4	16
18	-2	4
19	-1	1
21	1	1
17	-3	9
		$\Sigma(X - \bar{X})^2 = 60$

$$\begin{aligned} \text{Now, } \sigma &= \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}} = \sqrt{\frac{60}{9}} \\ &= \sqrt{6.66} = 2.58 \end{aligned}$$

Example 5.

Calculate the standard deviation of the marks obtained by 40 students (using short-cut method)

X	24	26	30	31	34	36	37	38
f	3	7	6	10	5	4	3	2

Solution:

CALCULATION OF S.D.

X	f	d = X - A(30)	d ²	fd	fd ²
24	3	-6	36	-18	108
26	7	-4	16	-28	112
30	6	0	0	0	0
31	10	1	1	10	10
34	5	4	16	20	80
36	4	6	36	24	144
37	3	7	49	21	147
38	2	8	64	16	128
	40			45	729

$$\begin{aligned}
 \text{Now, } \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \\
 &= \sqrt{\frac{729}{40} - \left(\frac{45}{40}\right)^2} = \sqrt{\frac{729}{40} - \left(\frac{9}{8}\right)^2} \\
 &= \sqrt{\frac{729}{40} - \frac{81}{64}} \\
 &= \sqrt{18.23 - 1.27} \\
 &= \sqrt{16.96} \\
 &= 4.12
 \end{aligned}$$

Example 6.

Calculate standard deviation from the following data (using short-cut method):

Class (X)	0-10	10-20	20-30	30-40	40-50
Frequency (f)	3	7	11	9	5

Solution:

COMPUTATION OF S.D.

Class (X)	Frequency (f)	Mid-value (m)	$d = m - A(25)$	d^2	fd	fd^2
0-10	3	5	-20	400	-60	1200
10-20	7	15	-10	100	-70	700
20-30	11	25	0	0	0	0
30-40	9	35	10	100	90	900
40-50	5	45	20	400	100	2000
					60	4800

$$\begin{aligned}
 \text{Now, } \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{4800}{35} - \left(\frac{60}{35}\right)^2} \\
 &= \sqrt{137.14 - (1.71)^2} = \sqrt{137.14 - 2.94} \\
 &= \sqrt{134.2} = 11.58
 \end{aligned}$$

Example 7.

The intelligence quotient (IQ) of 50 boys is given below. Calculate the standard deviation. (Step-deviation method)

Class	0-20	20-40	40-60	60-80	80-100	100-120	120-140	140-160
Frequency	3	4	3	4	13	12	8	3

Solution:

Making the frequency distribution table of above information, with assumed mean (A) = 90 and h = 20

COMPUTATION OF S.D.

Class (X)	Mid-value (m)	f	$d' = \frac{m - 90}{20}$	d'^2	$f \cdot d'$	$f \cdot d'^2$
0-20	10	3	-4	16	-12	48
20-40	30	4	-3	9	-12	36
40-60	50	3	-2	4	-6	12
60-80	70	4	-1	1	-4	4
80-100	90	13	0	0	0	0
100-120	110	12	1	1	12	12
120-140	130	8	2	4	16	32
140-160	150	3	3	9	9	27
		N = 50			$\sum fd' = 3$	$\sum fd'^2 = 171$

$$\begin{aligned}
 \text{Now, } \sigma &= h \times \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \\
 &= 20 \times \sqrt{\frac{171}{50} - \left(\frac{3}{50}\right)^2} \\
 &= 20 \times \sqrt{3.42 - 0.0036} \\
 &= 20 \times \sqrt{3.4164} \\
 &= 20 \times 1.8484 = 36.96
 \end{aligned}$$

Example 8.

Marks distribution of 100 students are given below:

Marks	0-5	5-10	10-15	15-20	20-25
No. of students	13	—	27	—	16

If average marks is 13.5, find standard deviation.

Solution:

Let f_1 and f_2 be two missing frequencies.

COMPUTATION OF MEAN

Marks (X)	No. of students (f)	Mid-value (m)	$d' = \frac{m-12.5}{5}$	fd'
0-5	13	2.5	-2	-26
5-10	f_1	7.5	-1	$-f_1$
10-15	27	12.5	0	0
15-20	f_2	17.5	1	f_2
20-25	16	22.5	2	32
Total	$N = \sum f = 56 + f_1 + f_2$			$\sum fd' = 6 - f_1 + f_2$

We have,

$$\text{Total number of students (N)} = 100$$

$$\text{or, } 56 + f_1 + f_2 = 100$$

$$\text{or, } f_1 + f_2 = 44 \quad (1)$$

$$\text{Again, mean } (\bar{X}) = A + \frac{\sum fd'}{N} \times h$$

$$\text{or, } 13.5 = 12.5 + \frac{6 - f_1 + f_2}{100} \times 5$$

$$\begin{aligned} \text{or,} \quad & 1 = \frac{6 + f_1 + f_2}{20} \\ \text{or,} \quad & 20 = 6 + f_1 + f_2 \\ \text{or,} \quad & -f_1 + f_2 = 14 \end{aligned} \quad (2)$$

Solving (1) and (2), we get

$$\begin{array}{rcl} f_1 + f_2 & = & 44 \\ -f_1 + f_2 & = & 14 \\ \hline 0 + 2f_2 & = & 58 \\ \therefore f_2 & = & 29 \end{array}$$

Putting the value of f_2 in equation (1), we get

$$\begin{aligned} f_1 + 29 & = 44 \\ \therefore f_1 & = 15 \end{aligned}$$

Now,

COMPUTATION OF STANDARD DEVIATION

Marks (X)	No. of students (f)	Mid-value (m)	$d' = \frac{m-12.5}{5}$	fd'	fd'^2
0-5	13	2.5	-2	-26	52
5-10	15	7.5	-1	-15	15
10-15	27	12.5	0	0	0
15-20	29	17.5	1	29	29
20-25	16	22.5	2	32	64
	$N = \Sigma f = 50$			$\Sigma fd' = 20$	$\Sigma fd'^2 = 160$

$$\begin{aligned} SD (\sigma) &= \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h = \sqrt{\frac{160}{100} - \left(\frac{20}{100}\right)^2} \times 5 \\ &= \sqrt{1.6 - 0.04} \times 5 = \sqrt{1.56} \times 5 = 6.245 \end{aligned}$$

Variance

The variance of a distribution is defined as the square of the standard deviation. Thus

$$\text{Variance} = \sigma^2$$

Hence to calculate variance, we calculate the standard deviation and square this value.

Relative measure of standard deviation

The standard deviation defined above is an absolute measure of dispersion. The corresponding relative measure is known as the coefficient of variation. The coefficient of variation used in such problems where we want to compare the variability of two or more than two series. To calculate the coefficient of variation, we use the following formula:

$$\text{Coefficient of variation (C.V.)} = \frac{\text{S.D.}}{\text{mean}} \times 100$$

$$\therefore \text{C.V.} = \frac{\sigma}{\bar{X}} \times 100$$

The less C.V. implies the more consistency or uniformity or reliability or less variability and vice versa.

Example 9.

Calculate the mean, standard deviation and coefficient of variation from the data given below.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	7	10	20	15	13	5

Solution:

CALCULATION OF MEAN AND SD

Marks (X)	Mid-value (m)	(f)	$d' = \frac{m - 25}{10}$	d'^2	fd'	fd'^2
0-10	5	7	-2	4	-14	28
10-20	15	10	-1	1	-10	10
20-30	25	20	0	0	0	0
30-40	35	15	1	1	15	15
40-50	45	13	2	4	26	52
50-60	55	5	3	9	15	45
		$\Sigma f = 70$			$\Sigma fd' = 32$	$\Sigma fd'^2 = 150$

We have, $N = 70$, $A = 25$, $h = 10$

$$\text{Now, } (\bar{X}) = A + \frac{\Sigma fd'}{N} \times h = 25 + \frac{32}{70} \times 10 = 25 + 4.57 = 29.57$$

$$\begin{aligned} \text{SD } (\sigma) &= h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \\ &= 10 \times \sqrt{\frac{150}{70} - \left(\frac{32}{70}\right)^2} \\ &= 10 \times \sqrt{2.14 - 0.21} = 10 \times \sqrt{1.93} \\ &= 10 \times 1.39 = 13.9 \end{aligned}$$

$$\therefore \text{Coefficient of variation (CV)} = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{13.9}{29.57} \times 100 = 47\%$$

Example 10.

The following are the weights of 10 rats in ten months. Both were fed same normal diet. If the consistency performance is the criterion, calculate which one maintains consistency.

Rat X	50	45	55	40	47	50	45	40	48	50
Rat Y	55	46	51	45	52	45	50	45	40	47

Solution:

Making the frequency distribution table we get

CALCULATION OF C.V.

X	$X - \bar{X}$	$(X - \bar{X})^2$	Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
50	+3	9	55	+7.4	54.76
45	-2	4	46	-1.6	2.56
55	+8	64	51	+3.4	11.56
40	-7	49	45	-2.6	6.76
47	0	0	52	+4.4	19.36
50	+3	9	45	-2.6	6.76
45	-2	4	50	+2.4	5.76
40	-7	49	45	+2.6	6.76
48	+1	1	40	-7.6	57.76
50	+3	9	47	0.6	0.36
$\Sigma X = 470$		$\Sigma (X - \bar{X})^2 = 198$	$\Sigma f = 476$		$\Sigma (Y - \bar{Y})^2 = 172.40$

Here, mean weight of rat X is $\bar{X} = \frac{\Sigma X}{n} = \frac{470}{10} = 47$

Mean weight of rat Y i.e. $\bar{Y} = \frac{\Sigma Y}{n} = \frac{476}{10} = 47.6$

For the rat X,

$$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n}} = \sqrt{\frac{198}{10}} = \sqrt{19.8} = 4.45$$

$$\therefore \text{C.V.} = \frac{\text{S.D.}}{\bar{X}} \times 100 = \frac{4.45}{47} \times 100 = 9.47$$

For the rat Y,

$$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{n}} = \sqrt{\frac{172.40}{10}} = \sqrt{17.24} = 4.15$$

$$\therefore \text{C.V.} = \frac{\text{S.D.}}{\bar{Y}} \times 100 = \frac{4.15}{47.6} \times 100 = 8.72$$

Here the coefficient of variation of rat Y is less than that of rat X. Thus the weight of rat Y is more consistent.

Combined Standard Deviation

Let two frequency distributions have means $\bar{X}_1$ and $\bar{X}_2$ and respective standard deviations σ_1 and σ_2 . Then their combined standard deviation is denoted by σ_{12} and given by

$$\sigma_{12} = \sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2}}$$

Where, $N_1 = \Sigma f$ for first frequency distribution

$N_2 = \Sigma f$ for second frequency distribution

$$\bar{X}_{12} = \text{Combined mean} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$$d_1 = \bar{X}_{12} - \bar{X}_1$$

$$d_2 = \bar{X}_{12} - \bar{X}_2$$

Example 11.

From the following two series, find the combined standard deviation.

	Mean	S. D.	No. of items
Series A	75	20	40
Series B	65	25	60

Solution:

We have given that

$$N_1 = 40, \sigma_1 = 20, \bar{X}_1 = 75$$

$$N_2 = 60, \sigma_2 = 25, \bar{X}_2 = 65$$

$$\begin{aligned}\text{Now, } \bar{X}_{12} &= \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} \\ &= \frac{40 \times 75 + 60 \times 65}{40 + 60} = 69\end{aligned}$$

$$\text{Also, } d_1 = \bar{X}_{12} - \bar{X}_1 = 69 - 75 = -6$$

$$d_2 = \bar{X}_{12} - \bar{X}_2 = 69 - 65 = 4$$

$$\begin{aligned}\text{Thus, } \sigma_{12} &= \sqrt{\frac{40(400 + 36) + 60(625 + 16)}{100}} \\ &= \sqrt{559} = 23.64\end{aligned}$$

EXERCISE - 5(II)

1. If $n = 10$, $\Sigma x = 60$, $\Sigma x^2 = 1000$ find standard deviation.
2. For a set of ungrouped values the following information are found $n = 15$, $\Sigma x = 480$, $\Sigma x^2 = 15735$. Find the standard deviation.
3. For a frequency distribution, S.D. = 2, $\bar{x} = 15$. Find the value of CV.
4. In a certain distribution, if $CV = 26.73\%$ and $\bar{x} = 39.5$, find its S.D.
5. Find the coefficient of the range
 - (a) 4, 9, 2, 7, 13, 8, 18.
 - (b) 32, 34, 31, 25, 21, 35, 35.
6. (a) If x be the variable values of n object and $\bar{x}$ their arithmetic mean then write the standard deviation and variance of the values.
 - (b) In a certain distribution, the arithmetic mean and S.D. are respectively $\bar{x}$ and σ , find coefficient of S.D. and CV.
 - (c) The mean and variance of certain distribution are 10 and 16 respectively. Find the coefficient of variation of the distribution.
 - (d) The mean and standard deviation of a certain distribution are respectively 45 and 22.5, find the C.V. of the distribution.
7. Find the Q.D. and its coefficient from the following:

- (a) 23, 25, 26, 27, 28, 29, 30

- (b)

Height (cm)	153	155	157	159	161	163
No. of students	25	21	28	20	18	24

8. Calculate the S.D. from the following data

- (a) 8, 9, 15, 23, 5, 11, 19, 10, 12
 (b) 30, 40, 35, 22, 25, 48, 45.
 (c) 40, 60, 70, 80, 90, 100.
 (d) 40, 50, 60, 70, 80, 90, 100, 110, 120, 130
 (e) 2, 4, 7, 13, 14, 16, 18, 20.

9. Find the standard deviation of

x	10	11	12	13	14
f	3	12	18	12	2

10. Calculate quartile deviation and its coefficient from the following data: 16, 20, 25, 15, 21, 12, 19, 24.

11. Calculate the quartiles deviation for the data given below:

Age	0-10	10-20	20-30	30-40	40-50	50-60
No. of patient	2	9	10	8	6	3

12. Define standard deviation and variance. Why S.D. is the best measure of dispersion? Calculate the standard deviation from mean of the following distribution. Also calculate the coefficient of variation of the respiratory rate of patients per minute are 23, 22, 20, 24, 16, 17, 18, 19, 21.

13. Find the standard deviation from mean of the following distribution.

Age (yr) Above	0	10	20	30	40
No. of persons	100	88	55	25	10

14. (a) Find C.V. from following:

X	10-20	20-30	30-40	40-50	50-60
f	2	5	9	3	1

(b) Compute the C.V. from following:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	3	7	11	8	6	6

(c) Calculate the standard deviation and C.V. from the data given below:

Weight (kg) above	0	10	20	30	40
No. of person	100	88	55	25	10

(d) Calculate standard deviation and C.V. for the data given below:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	7	12	24	10	7

(e) Find the standard deviation and C.V. from the following data:

Variable	0-10	10-20	20-30	30-40	40-50
Frequency	2	9	10	8	3

(f) Compute the coefficient of variation (C.V.) from the given data:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	3	7	11	8	6	5

(g) Calculate the coefficient of variation for the following data:

x	5-10	10-15	15-20	20-25	25-30
f	2	5	7	4	1

15. Marks distribution of 200 students are given below:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	26	—	54	—	32

If the average marks are 27, find standard deviation.

ANSWERS

- 8
- 10.56
- 0.8
 - 0.25
- $$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}; \text{ Variance } (\sigma^2) = \frac{\sum(x - \bar{x})^2}{n}$$
- $\frac{\sigma}{\bar{x}} \times 100$
 - 40%
 - 50%
- 2; 0.074
 - 3; 0.0189
 - 5.34
 - 9.17
 - 28.82
 - 6.22
 - 0.96
 10. 4; 0.10
 - 12, 91
 13. 11.45
 - 29.69
 - 46.96
 - 11.39; 46.21
 - 10.46; 41.11
 - 14.48; 47.48
 - 5.19; 31.10
- 19.72
 11. 18.33; 39.38
 - 11.06; 98.31
- 12.48



Concepts of Probability

6.1 Introduction

The word probability can be taken as the synonym of the words, almost certain, chance, most likely, probably etc. All the words mentioned above are related to uncertainty which are usually used in our daily life. The numerical measure of the degree of uncertainty is known as the "Theory of Probability". This theory was first originated by a gambler named Chevalier de Mere in 17th century on the process of playing the games of chances. For examples, tossing a coin, throwing a die, drawing a card etc. But now a days it has wide applications in the field of Engineering, Medical Science, Economics, Business, Statistics and many other fields also. And hence the probability has become one of the important branches of modern mathematics.

Probability and Engineering

Probability study is very important to engineers for it examines how likely events could happen and so the risks could be determined and resolved professionally. Another significant application of probability theory in everyday life is reliability. Many consumer products, such as automobiles and consumer electronics, use reliability theory in product design to reduce the probability of failure.

6.2 Basic Terms in Probability

Before defining probability, we shall give basic terms and concepts used there in.

Experiment

An operation, which can produce some well-defined outcomes, is known as an experiment. There are two types of experiment.

Deterministic experiments: When we perform experiments in science and engineering, repeatedly under very nearly identical conditions, we get almost the same result. Such experiments are called deterministic

experiments. In other words, an experiment whose results are already determined is known as deterministic experiments.

Random experiments: An experiment is called a random experiment if when conducted repeatedly under essentially homogeneous conditions, the result is not unique but may be any one of the various possible outcomes. In other words, an experiment whose results are not already determined is known as random experiments. When tossing a coin, either head or tail come up, which cannot be pre-determined.

Sample space

The set of all possible outcomes of the random experiment is called the sample space. The possible outcomes are called sample points. The sample space is generally denoted by the letter S. We list the sample space of some random experiments.

<u>Random Experiment</u>	<u>Sample space</u>
• Tossing of an unbiased coin	$S = \{H, T\}$
• Throwing of a fair die	$S = \{1, 2, 3, 4, 5, 6\}$
• Tossing of an unbiased coin twice	$S = \{HH, HT, TH, TT\}$
• A family of two children	$S = \{BB, BG, GB, GG\}$

In an experiment, if a trial having m possible outcomes is performed n times then there are m^n sample points.

Example 1.

If a coin is tossed three times how many possible outcomes are there?

Solution:

Here number of possible outcomes in a single trial (m) = 2

Total number of trial (n) = 3

So, total number of sample points = $m^n = 2^3 = 8$

Thus when a coin is tossed three times there are 8 possible outcomes.

Example 2.

If a die is thrown twice find the number of possible outcomes.

Solution:

Here number of possible outcomes in a single throw (m) = 6

Total number of trial (n) = 2

So, total number of sample points = $m^n = 6^2 = 36$.

Trial

Performing of a random experiment is called a trial. For example, throwing a die, tossing an unbiased coin, drawing a ball from an urn and so on are some trials.

Event

An event is defined as a subset of the sample space of a random experiment. There are several types of events. Some of them are given below.

Simple and composite event: An event is called an simple event, if it contains only one sample point. It is also known as an elementary event. If an event corresponds to more than one possible outcome of the experiment, it is known as composite or compound event. For example, in tossing of a single die, the event of getting '5' is a simple event but the event "getting an even number" is a composite event because it contains $\{2, 4, 6\}$.

Impossible and sure event: An event is called an impossible event, if it can never occur. In the experiment of rolling a die, the event of getting 7 is an impossible event. An event, which is sure to occur, is called a sure event. In the above example of rolling a die, the event of getting a number less than 7 is a sure event.

Equally likely cases (events)

The outcomes of a random experiment are called equally likely, if all of these have equal preferences or chances to occur. In the experiment of tossing a unbiased coin, the outcomes 'head' (H) and 'tail' (T) are equally likely.

Favourable cases (events)

The number of outcomes of a random experiment which entail (or result in) the happening of an event are termed as the cases favourable to the event. For example, in drawing a card from a pack of cards, the cases favourable to 'getting a diamond' are 13 and to 'getting an ace of spade' is only one 1.

Exhaustive cases (events)

The total number of possible outcomes of a random experiment is called exhaustive cases for the event. In toss of a single coin, we can get head (H) or tail (T). Hence, exhaustive number of cases is $(H, T) = 2$.

Mutually exclusive events

Two or more events are said to be mutually exclusive if the happening of anyone of them excludes the happening of all others in the same trial. For example, in toss of a coin, the events H and T are mutually exclusive because if H comes, we can't get T and if T comes we can't get H.

Thus, events are said to be mutually exclusive if no two or more of them can happen simultaneously.

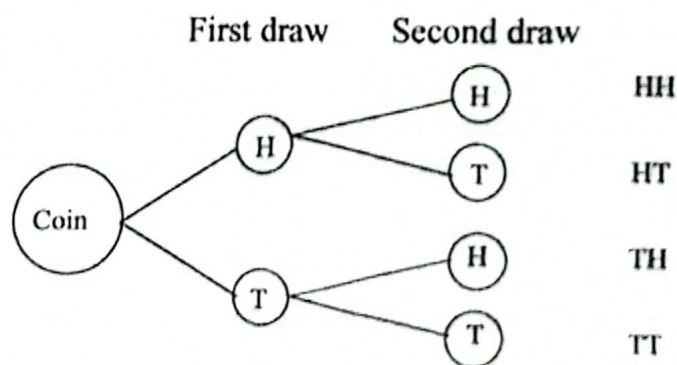
Independent events

Two events are said to be independent to each other if the occurrence or non-occurrence of one event does not affect the occurrence of the other event. For example, in tossing of a die repeatedly, the event of getting '5' in the first throw is independent of getting '5' in the second, third or subsequent throws.

6.3 Tree Diagram

A tree diagram is a device used to enumerate all the logical possibilities of a sequence of things where each thing can occur in a finite number of ways. A tree diagram is constructed from left to right and the number of branches at any point corresponds to the number of ways the next thing can occur.

The tree diagram of the sample space of the toss of two coins is shown in the following diagram.



6.4 Classical or priori (or mathematical) Theory of Probability

In a random experiment, if there are n exhaustive, mutually exclusive and equally likely cases out of which m are favourable to the happening of an event A , then the probability of occurrence of A , usually denoted by $P(A)$ and is given by

$$P(A) = \frac{\text{Total number of cases favourable to } A}{\text{Total number of exhaustive, mutually exclusive and equally likely cases}} = \frac{m}{n}$$

This definition was given by James Bernoulli who is the first man to obtain a quantitative measure of uncertainty.

Notes:

- (i) The probability of non-occurrence of event A is denoted by $P(\bar{A})$ and is given by

$$P(\bar{A}) = 1 - P(A)$$

$$\Rightarrow P(A) + P(\bar{A}) = 1$$

- (ii) The probability of any event is a number lying between 0 and 1.

i.e. $0 \leq P(A) \leq 1$ for any event A.

- (iii) If A is impossible event then $P(A) = 0$ and
if A is sure event, then $P(A) = 1$

Limitations:

The classical probability has its short-comings and fails in the following situations:

- If the exhaustive number of outcomes of the random experiment is infinite.
- If the various outcomes of the random experiment are not equally likely.
- If the actual value of n is unknown.

Example 3.

Three coins are tossed simultaneously. Write the sample space and the probability of getting two heads.

Solution:

When the three coins are tossed simultaneously, the sample space

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

Here, total no. of exhaustive cases = 8

Let A denotes the event containing two heads. Then

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

Favourable cases to the event of getting two heads = 3

By the definition of probability,

$$P(\text{two heads}) = \frac{3}{8}.$$

Emperical or statistical definition:

When the number of trials tends to become infinite, we can use the relative frequency theory to find the probability of an event. Suppose that an event A occurs m times in n repetitions of a random experiment. Then the

ratio $\frac{m}{n}$ gives the relative frequency of the event A and it will not vary appreciably from one trial to another.

According to **Croxton** and **Crowden** the probability of the event A is the limit of the relative frequency of success in an infinite number of trials.

Symbolically,

$$P(A) = \lim_{n \rightarrow \infty} \frac{m}{n}$$

Limitations:

- (a) A large number of experiments may not be remaining essentially homogeneous.
- (b) The relative frequency $\frac{m}{n}$, may not attain a unique value, no matter however large n may be.

6.5 Basic Laws of Probability Theory

The two theorems are known as the basic laws of probability.

(a) Addition theorem (theorem of total probability)

If A and B are two events with their respective probabilities $P(A)$ and $P(B)$ then the probability of occurrence of at least one of these two events denoted by $P(A \cup B)$ and is given by

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

where $P(A \cap B)$ is the probability of the simultaneous occurrence of the events A and B.

Proof:

Let n be the total number of exhaustive and equally likely cases of the experiment. Let m_1 , m_2 and m_3 be the number of cases favourable to the happening of the events A, B and $A \cap B$ respectively.

Then,

$$P(A) = \frac{m_1}{n}, P(B) = \frac{m_2}{n} \text{ and } P(A \cap B) = \frac{m_3}{n}.$$

Since the outcomes m_3 are common to both events A and B, then the favourable cases to the event $(A \cup B) = m_1 + m_2 - m_3$. Here m_3 is subtracted from $m_1 + m_2$ to avoid counting of common outcomes twice.

$$\text{Therefore, } P(A \cup B) = \frac{m_1 + m_2 - m_3}{n}$$

$$\begin{aligned}
 &= \frac{m_1}{n} + \frac{m_2}{n} - \frac{m_3}{n} \\
 &= P(A) + P(B) - P(A \cap B).
 \end{aligned}$$

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

This completes the proof of theorem of total probability.

Notes:

- (i) If events A and B are mutually exclusive, then $P(A \cap B) = 0$, and in this case addition theorem implies $P(A \cup B) = P(A) + P(B) - 0 = P(A) + P(B)$.
- (ii) This theorem can also be extended to more than two events. If A, B and C are the three events with their respective probabilities $P(A)$, $P(B)$ and $P(C)$ respectively, then $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$.
- (iii) If events A, B and C are mutually exclusive, then $P(A \cup B \cup C) = P(A) + P(B) + P(C)$.

Example 4.

If A and B are events such that $P(A) = 0.3$, $P(B) = 0.6$ and $P(A \cap B) = 0.2$, then find (a) $P(\bar{A})$ (b) $P(\bar{B})$ (c) $P(A \cup B)$.

Solution:

We have $P(A) = 0.3$, $P(B) = 0.6$ and $P(A \cap B) = 0.2$

- (a) $P(\bar{A}) = 1 - P(A) = 1 - 0.3 = 0.7$
- (b) $P(\bar{B}) = 1 - P(B) = 1 - 0.6 = 0.4$
- (c) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 $= 0.3 + 0.6 - 0.2$
 $= 0.9 - 0.2$
 $= 0.7$

Example 5.

A card is drawn from a well-shuffled pack of 52 cards. What is the probability that it is either a spade or an ace.

Solution:

Let A be the event of drawing a spade and B the event of drawing an ace. The events A and B are not mutually exclusive as one ace is a spade also.

$$\therefore P(A) = \frac{13}{52}$$

$$P(B) = \frac{4}{52}$$

$A \cap B$ = the event of drawing the ace of spades

$$P(A \cap B) = \frac{1}{52}$$

$$\text{We have, } P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$

$$= \frac{4}{13}$$

Example 6.

The probability that a person will travel by a plane is $\frac{2}{3}$ and that he will travel by bus is $\frac{1}{5}$. Find the probability of his traveling by plane or bus.

Solution:

Let A = event that the person travels by plane

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

Let B = event that the person travels by bus.

$$\therefore P(B) = \frac{1}{5}$$

The events A and B are mutually exclusive, because he cannot travel by plane and bus at the same time.

Therefore, the probability of his traveling by plane or bus

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &= \frac{2}{3} + \frac{1}{5} = \frac{13}{15} \end{aligned}$$

Example 7.

A die is thrown twice. What is the probability that at least one of the two throws comes up with the number 4?

Solution:

Let S be sample space. Then, $n(S) = 6^2 = 36$.

Let A = the event that the first throw shows 4

and B = the event that the second throw shows 4

$\therefore A \cap B$ = the event showing 4 in each throw.

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

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

and $A \cap B$ = $\{(4, 4)\}$

$\therefore n(A) = 6, n(B) = 6$ and $n(A \cap B) = 1$

So, $P(A) = \frac{6}{36} = \frac{1}{6}$; $P(B) = \frac{6}{36} = \frac{1}{6}$ and $P(A \cap B) = \frac{1}{36}$

$$\begin{aligned}\therefore P(\text{at least one 4}) &= P(A \text{ or } B) = P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= \left(\frac{1}{6} + \frac{1}{6} - \frac{1}{36} \right) = \frac{11}{36}\end{aligned}$$

Example 8.

Let A and B be the events such that, $P(A) = 0.3$, $P(A \cup B) = 0.4$ and $P(B) = x$. Find the value of x such that

- A and B are mutually exclusive,
- A and B are independent.

Solution:

- Let A and B be mutually exclusive events.

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

or, $P(B) = P(A \cup B) - P(A)$

$$\therefore x = (0.4 - 0.3) = 0.1$$

- Let A and B be independent events.

Then, $P(A \cap B) = P(A) \cdot P(B) = 0.3x$ (1)

$$\begin{aligned}\text{But, } P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= 0.3 + x - 0.4 \\ &= (x - 0.1)\end{aligned}$$

Thus, $P(A \cap B) = (x - 0.1)$ (2)

From (1) and (2), we get

$$0.3x = x - 0.1$$

$$\text{or, } 0.1 = x - 0.3x$$

$$\text{or, } 0.1 = x(1 - 0.3)$$

$$\text{or, } 0.1 = 0.7x$$

$$\begin{aligned} \text{psp} \quad \text{or, } x &= \frac{0.1}{0.7} = \frac{1}{7} \\ \therefore x &= \frac{1}{7} \end{aligned}$$

(b) Multiplication theorem on probability (Theorem of compound probability)

If two events A and B are independent, then probability that they will both occur is equal to the product of their individual probabilities. i.e. $P(A \text{ and } B) = P(A) \times P(B)$, i.e., $P(A \cap B) = P(A) \times P(B)$.

Proof:

Let n_1 and n_2 be the total number of possible cases for the events A and B respectively and m_1 and m_2 are their respective favourable cases.

$$\text{Then, } P(A) = \frac{m_1}{n_1}$$

$$P(B) = \frac{m_2}{n_2}$$

Clearly, by the fundamental principle of events, we can combine each of n_1 possible cases of the event A with each of the n_2 possible cases of the event B for their simultaneous occurrence.

Then,

the total number of possible cases = $n_1 \cdot n_2$

the total number of favourable cases = $m_1 \cdot m_2$

$\therefore$ By definition of probability

$$P(A \text{ and } B) = \frac{m_1 \cdot m_2}{n_1 \cdot n_2}$$

$$= \frac{m_1}{n_1} \times \frac{m_2}{n_2}$$

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

Hence, $P(A \cap B) = P(A) \times P(B)$.

This completes the proof of the theorem of compound probability.

Notes:

- (i) $P(A \cap B) = P(A) \times P(B)$ is the necessary and sufficient condition for the events A and B to be independent.

- (ii) The probability that none of the two independent events A and B will happen = $[1 - P(A)][1 - P(B)]$.
- (iii) The probability that at least one event will happen is
 $= 1 - [1 - P(A)][1 - P(B)]$.
- (iv) If A, B and C are three independent events, then the probability of their simultaneous occurrence $P(A \cap B \cap C)$ is $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

Example 9.

Two dice are thrown. Find the probability of getting an odd number on the first die and a multiple of 3 on the other.

Solution:

Let A = event of getting odd number on first die,
 $= \{1, 3, 5\}$

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

Let B = event of getting a multiple of 3 on second die
 $= \{3, 6\}$

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

The events A and B are independent.

$$\begin{aligned} \therefore \text{Required Probability} &= P(A \cap B) = P(A) \times P(B) \\ &= \frac{1}{2} \times \frac{1}{3} = \frac{1}{6} \end{aligned}$$

Example 10.

Let A and B are two non-mutually exclusive events. If $P(A) = \frac{1}{4}$, $P(B) = \frac{2}{5}$ and $P(A \cup B) = \frac{1}{2}$, find the value of $P(A \cap B)$.

Solution:

We have $P(A) = \frac{1}{4}$, $P(B) = \frac{2}{5}$ and $P(A \cup B) = \frac{1}{2}$.

By addition theorem on probability,

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

$$\begin{aligned}\text{or, } \frac{1}{2} &= \frac{1}{4} + \frac{2}{5} - P(A \cap B) \\ \text{or, } P(A \cap B) &= \frac{1}{4} + \frac{2}{5} - \frac{1}{2} = \frac{5+8-10}{20} \\ &= \frac{3}{20}.\end{aligned}$$

Example 11.

A can solve 90% of the problems given in a book and B can solve 70%. What is the probability that at least one of them will solve a problem, selected at random from the book?

Solution:

Let A = the event that A solves the problem.

B = the event that B solves the problem.

Clearly, A and B are independent events.

$$\text{Now, } P(A) = \frac{90}{100} = \frac{9}{10}; P(B) = \frac{70}{100} = \frac{7}{10}$$

$$\therefore P(A \cap B) = P(A) \cdot P(B) = \left(\frac{9}{10} \times \frac{7}{10} \right) = \frac{63}{100}$$

[$\because$ A and B are independent]

$$\begin{aligned}\text{So, } P(A \text{ or } B) &= P(A \cup B) = P(A) + P(B) - P(A \cap B) \\ &= \left(\frac{9}{10} + \frac{7}{10} - \frac{63}{100} \right) = \frac{97}{100} = 0.97\end{aligned}$$

Example 12.

The probability that a person will win a car race is $\frac{2}{3}$, and the probability that he will not win a horse race is $\frac{5}{9}$. If the probability of getting success in at least one of the events is $\frac{4}{5}$, what is the probability that he will be successful in both the events?

Solution:

Let A denotes the event of winning a game and B denote the event of winning a horse race.

$$\text{Then, } P(A) = \frac{2}{3}$$

$$P(\bar{B}) = \frac{5}{9}$$

$$\therefore P(B) = 1 - P(\bar{B}) = 1 - \frac{5}{9} = \frac{4}{9}$$

$$\text{Also, } P(A \cup B) = \frac{4}{5}$$

These are non-mutually exclusive events as both events can happen together.

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

$$\text{or, } \frac{4}{5} = \frac{2}{3} + \frac{4}{9} - P(A \cap B)$$

$$\text{or, } P(A \cap B) = \frac{2}{3} + \frac{4}{9} - \frac{4}{5} = \frac{42}{135} = \frac{14}{45}$$

$\therefore$ Hence the probability that a person will be successful in both the events $P(A \cap B) = \frac{14}{45}$.

Example 13.

A husband and a wife appear in an interview of two vacancies in the same post. The probability of husband's selection is $\frac{1}{7}$ and that of wife's selection is $\frac{1}{5}$. What is the probability that

- (a) both of them will be selected
- (b) only one of them will be selected
- (c) none of them will be selected
- (d) at least one of them will be selected

Solution:

Let E = the event that the husband is selected.

F = the even that the wife is selected

$$\text{Now, } P(E) = \frac{1}{7} \text{ and } P(F) = \frac{1}{5}$$

$$\therefore P(\bar{E}) = 1 - \frac{1}{7} = \frac{6}{7}, P(\bar{F}) = 1 - \frac{1}{5} = \frac{4}{5}$$

$$(a) \quad P(\text{both of them will be selected}) = P(E \cap F)$$

$$= P(E) \cdot P(F) = \frac{1}{7} \times \frac{1}{5} = \frac{1}{35}$$

$$(b) \quad P(\text{only one of them will be selected})$$

$$= P[(E \text{ and not } F) \text{ or } (F \text{ and not } E)]$$

$$= P[(E \cap \bar{F}) \cup (F \cap \bar{E})]$$

$$= P(E \cap \bar{F}) + P(F \cap \bar{E})$$

$$[\because (E \cap \bar{F}) \cap (F \cap \bar{E}) = \phi]$$

$$= P(E) \cdot P(\bar{F}) + P(F) \cdot P(\bar{E})$$

$$= \frac{1}{7} \times \frac{4}{5} + \frac{1}{5} \times \frac{6}{7} = \frac{4}{35} + \frac{6}{35} = \frac{10}{35} = \frac{2}{7}$$

$$(c) \quad P(\text{none of them will be selected})$$

$$= P[(\text{not } E) \text{ and } (\text{not } F)]$$

$$= P(\bar{E} \cap \bar{F}) = P(\bar{E}) \times P(\bar{F})$$

$$= \frac{6}{7} \times \frac{4}{5} = \frac{24}{35}$$

$$(d) \quad P(\text{at least one of them will be selected})$$

$$= 1 - P(\text{none will be selected})$$

$$= \left(1 - \frac{24}{35}\right) = \frac{11}{35}$$

Example 14.

The probabilities for solving a problem of statistics by three students A, B and C are $\frac{1}{5}$, $\frac{1}{4}$ and $\frac{1}{3}$ respectively. What is the probability that the problem will be solved?

Solution:

The respective probabilities of the events A, B and C are $P(A) = \frac{1}{5}$, $P(B) = \frac{1}{4}$ and $P(C) = \frac{1}{3}$.

Probability that the problem will be solved by A or B or C is

$$\begin{aligned}
 P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C) \\
 &= P(A) + P(B) + P(C) - P(A) \cdot P(B) - P(B) \cdot P(C) - P(C) \cdot P(A) + P(A) \cdot P(B) \cdot P(C) \\
 &= \frac{1}{5} + \frac{1}{4} + \frac{1}{3} - \frac{1}{5} \times \frac{1}{4} - \frac{1}{4} \times \frac{1}{3} - \frac{1}{3} \times \frac{1}{5} + \frac{1}{5} \times \frac{1}{4} \times \frac{1}{3} \\
 &= \frac{1}{5} + \frac{1}{4} + \frac{1}{3} - \frac{1}{20} - \frac{1}{12} - \frac{1}{15} + \frac{1}{60} \\
 &= \frac{12 + 15 + 20 - 3 - 5 - 4 + 1}{60} = \frac{36}{60} = \frac{3}{5}
 \end{aligned}$$

Alternative Method

Let $\bar{A}$, $\bar{B}$ and $\bar{C}$ are the events of non-solving the problem.

$$\therefore P(\bar{A}) = 1 - P(A) = 1 - \frac{1}{5} = \frac{4}{5}$$

$$P(\bar{B}) = 1 - P(B) = 1 - \frac{1}{4} = \frac{3}{4}$$

$$P(\bar{C}) = 1 - P(C) = 1 - \frac{1}{3} = \frac{2}{3}$$

Clearly A, B, C are independent events and the problem will be considered to have been solved if at least one student solves it.

$\therefore$ The required probability = $P(A \text{ or } B \text{ or } C)$

$$= P(A \cup B \cup C) = 1 - P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C})$$

$$= 1 - \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} = 1 - \frac{2}{5} = \frac{3}{5}$$

Example 15.

The probabilities of A, B, C solving a problem are $\frac{1}{3}$, $\frac{2}{7}$ and $\frac{3}{8}$ respectively. If all the three try to solve the problem simultaneously, find the probability that the problem will not be solved.

Solution:

Let A, B, C be the events of solving the problem by A, B, C respectively.

Then, $\bar{A}, \bar{B}, \bar{C}$ are the respective events of not solving the problem.

$$\text{Now, } P(A) = \frac{1}{3}; P(B) = \frac{2}{7} \text{ and } P(C) = \frac{3}{8}$$

$$\therefore P(\bar{A}) = \left(1 - \frac{1}{3}\right) = \frac{2}{3}; P(\bar{B}) = \left(1 - \frac{2}{7}\right) = \frac{5}{7}$$

$$\text{and } P(\bar{C}) = \left(1 - \frac{3}{8}\right) = \frac{5}{8}$$

$\therefore$ The probability that the problem will not be solved

$$\text{i.e., } P(\text{none solving the problems}) = P(\bar{A} \cap \bar{B} \cap \bar{C})$$

Since $\bar{A}, \bar{B}$ and $\bar{C}$ are independent

$$= P(\bar{A}) \cdot P(\bar{B}) \cdot P(\bar{C})$$

$$= \frac{2}{3} \cdot \frac{5}{7} \cdot \frac{5}{8} = \frac{25}{84}$$

6.6 Counting Rules (Permutation and Combination)

Permutation and combination, in mathematics is certain arrangement of objects or elements. Permutation and combinations are important in many branches of mathematics. For example, in probability theory and statistics, they can be used to count the number of possible arrangements of a system. It has important applications to the design and operation of computers as well as to the physical and social science. Indeed in any area where the possible arrangement of a finite number of elements play a role, the theory of permutation and combination is useful.

Example 16.

Suppose that committees of 5 people are selected at random from a group of 7 boys and 5 girls. What is the probability that 3 boys and 2 girls are selected?

Solution:

$$\text{Total number of people} = 7 + 5 = 12$$

$$\text{Let } n = \text{total number possible cases}$$

$$= \text{Number of selection of 5 out of 12 people}$$

$$= {}^{12}C_5 = \frac{12!}{7!5!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 792$$

Since we need 3 boys and 2 girls, so 3 boys can be selected from 7 boys and 2 girls from 5 girls in ${}^7C_3 \times {}^5C_2$ ways.

$$\begin{aligned}\therefore \text{Number of favourable cases (m)} &= {}^7C_3 \times {}^5C_2 \\ &= \frac{7!}{4!3!} \times \frac{5!}{3!2!} = 35 \times 10 \\ &= 350\end{aligned}$$

$$\therefore P(3 \text{ boys and 2 girls}) = \frac{m}{n} = \frac{350}{792} = \frac{175}{396}$$

Example 17.

A bag contains 12 items of which 4 are defective. From this 2 items are drawn at random (c) both defective items.

Solution:

$$\begin{aligned}\text{Let } n &= \text{Total no. of possible cases} \\ &= \text{No. of selection of 2 out of 12 items} \\ &= {}^{12}C_2 = \frac{12!}{2!10!} = \frac{12 \cdot 11}{2 \cdot 1} = 66\end{aligned}$$

Now,

(a) No. of defective item = 4.

So number of non-defective item = 8

So, the number of selection that 2 items can be selected out of 8 is

$$= {}^8C_2 = \frac{8!}{2!6!} = \frac{8 \times 7}{2} = 28$$

$$\therefore P(\text{no item defective}) = \frac{m}{n} = \frac{28}{66} = \frac{14}{33}$$

(b) One defective item

Since we need 1 item defective and 1 item non-defective, so 1 defective item be selected from 4 defectives and other non-defective item from rest 8 non-defectives in ${}^4C_1 \times {}^8C_1$ ways.

$$\begin{aligned}\therefore \text{Number of favourable cases (m)} &= {}^4C_1 \times {}^8C_1 = 4 \times 8 \\ &= 32\end{aligned}$$

$$\therefore P(\text{one item defective}) = \frac{m}{n} = \frac{32}{66} = \frac{16}{33}$$

(c) Both defective items

2 defective items can be selected from 4 defective items, in 4C_2 ways.

So, no. of favourable cases

$$m = {}^4C_2 = \frac{4!}{2!2!} = \frac{4 \cdot 3}{2 \cdot 1} = 6$$

$$\therefore P(\text{both items defective}) = \frac{6}{66} = \frac{1}{11}.$$

Example 18.

A bag contains 8 red and 5 white balls. Three balls are drawn at random. Find the probability that

- (a) all three balls are white;
- (b) all three balls are red;
- (c) one ball is red and two balls are white.

Solution:

Total number of balls = $(8 + 5) = 13$.

Let S be the sample space. Then,

$n(S)$ = number of ways of selecting 3 balls out of 13

$$\begin{aligned} &= C(13, 3) = \frac{13!}{(13-3)!3!} \\ &= \frac{13!}{10!3!} = \frac{13 \times 12 \times 11 \times 10!}{10! 3 \times 2 \times 1} \\ &= \frac{13 \times 12 \times 11}{3 \times 2 \times 1} \\ &= 286 \end{aligned}$$

(a) Let E_1 be the event of getting 3 white balls.

Then, $n(E_1)$ = number of ways of selecting 3 balls out of 5

$$\begin{aligned} &= C(5, 3) = \frac{5!}{(5-3)!3!} \\ &= \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = 10 \end{aligned}$$

$$\therefore P(\text{getting 3 white balls}) = \frac{n(E_1)}{n(S)} = \frac{10}{286} = \frac{5}{143}$$

- (b) Let E_2 be the event of getting 3 red balls.

Then, $n(E_2)$ = number of ways of selecting 3 balls out of 8

$$= C(8, 3) = \frac{8!}{(8-3)!3!}$$

$$= \frac{8 \times 7 \times 6 \times 5!}{5! \times 3 \times 2 \times 1} = 56$$

Ans

$$\therefore P(\text{getting 3 red balls}) = \frac{n(E_2)}{n(S)} = \frac{56}{286} = \frac{28}{143}$$

- (c) Let E_3 be the event of getting 1 red and 2 white balls.

Then, $n(E_3)$ = (number of ways of selecting 1 ball out of 8) $\times$
(number of ways of selecting 2 balls out of 5)

$$= C(8, 1) \times C(5, 2) = (8 \times 10) = 80$$

$$\therefore P(\text{getting 1 red and 2 white balls}) = \frac{n(E_3)}{n(S)} = \frac{80}{286} = \frac{40}{143}$$

EXERCISE - 6

1. If $P(A) = 0.3$, find the value of $P(\bar{A})$.
2. What is the probability of getting tails in a throw of a coin?
3. A card is drawn at random from a well shuffled deck of 52 cards. What is the probability that it is
 - (a) a spade?
 - (b) an ace?
 - (c) a red card?
 - (d) a picture (face) card?
4. A dice is thrown once. Find the probability of getting
 - (a) a number less than or equal to 4.
 - (b) a prime number
 - (c) a multiple of 3
 - (d) an even number
 - (e) number greater than 2.
5. Two coins are tossed simultaneously. What is the sample space?
6. Two unbiased coins are tossed simultaneously. Find the probability of getting:

- (a) one head (b) one tail
 (c) at most one head (d) at least one head
 (e) more than 2 heads (f) less than 3 tails.
7. What is the probability that an alphabet selected at random is (i) a vowel (ii) a consonant?
8. A card is drawn randomly from a well shuffled deck of 52 cards. What is the probability that it is
 (a) an ace or knave?
 (b) a heart or a spade?
 (c) a king or a queen?
 (d) a heart or an ace?
9. Two unbiased dice are thrown simultaneously. Find the probability of getting:
 (a) sum of 10 (b) sum of 7
 (c) not same number on the dice (d) multiple of 3 as the sum.
10. Let A and B be two possible events of an experiment, such that $P(A) = 0.4$, $P(B) = K$ and $P(A \cup B) = 0.7$
 (a) For what value of K are A and B mutually exclusive?
 (b) For what value of K are A and B independent?
11. If E and F be the events in a sample space such that $P(E) = 0.4$, $P(F) = 0.5$ and $P(E \cup F) = 0.6$, find $P(E \cap F)$.
12. In a class of 200 students, 120 take mathematics, 110 take physics and 40 take both mathematics and physics. If a student is selected at random from this class, find the probability that the student takes mathematics or physics.
13. In a class, 30% of the students offered mathematics, 20% offered chemistry and 10% offered both. If a student is selected at random, find the probability that he has offered mathematics or chemistry.
14. A, B, C are three mutually exclusive and exhaustive events associated with a random experiment. Find $P(A)$, it being given that

$$P(B) = \frac{3}{2} P(A) \text{ and } P(C) = \frac{1}{2} P(B).$$
15. If A, B and C are three mutually exclusive events with

$$\frac{1}{3} P(A) = \frac{2}{3} P(B) = \frac{1}{6} P(C),$$
 find the value of $P(A)$, $P(B)$ and $P(C)$.

16. The probability that a contractor will get a plumbing contract is $\frac{2}{3}$, and he will not get an electric contract is $\frac{5}{9}$. If the probability of getting at least one contract is $\frac{4}{5}$. What is the probability that he will get both?
17. Arun and Tarun appeared for an interview for two vacancies. The probability of Arun's selection is $\frac{1}{4}$ and that of Tarun's rejection is $\frac{2}{3}$. Find the probability that at least one of them will be selected.
18. A problem in statistics is given to three students whose chances of solving it are $\frac{1}{3}$, $\frac{1}{4}$ and $\frac{2}{5}$ respectively. What is the probability that the problem will be solved?
19. In a class, there are 60 boys and 40 girls. If two students are chosen at random, what will be the probability that:
- both are boys?
 - both are girls?
 - boy and one girl?
20. Suppose 3 people are selected out of 7 men and 6 women. What is the probability that 2 men and 1 woman are selected?

ANSWERS

- | | | |
|------------------------------------|--------------------|--------------------------------------|
| 1. 0.7 | 2. $\frac{1}{2}$ | |
| 3. (a) $\frac{1}{4}$ | (b) $\frac{1}{13}$ | (c) $\frac{1}{2}$ (d) $\frac{3}{13}$ |
| 4. (a) $\frac{2}{3}$ | (b) $\frac{1}{2}$ | (c) $\frac{1}{3}$ |
| (d) $\frac{1}{2}$ | (e) $\frac{2}{3}$ | |
| 5. Sample space = {HH, HT, TH, TT} | | |
| 6. (a) $\frac{1}{2}$ | (b) $\frac{1}{2}$ | (c) $\frac{3}{4}$ |

- (d) $\frac{3}{4}$ (e) 0 (f) 1
7. (a) $\frac{5}{26}$ (b) $\frac{21}{26}$
8. (a) $\frac{2}{13}$ (b) $\frac{1}{2}$ (c) $\frac{2}{13}$ (d) $\frac{4}{13}$
9. (a) $\frac{1}{12}$ (b) $\frac{1}{6}$ (c) $\frac{5}{6}$ (d) $\frac{1}{3}$
10. (a) 0.3 (b) 0.5 11. 0.3
12. $\frac{19}{20}$ 13. $\frac{2}{5}$
14. $\frac{4}{13}$ 15. $\frac{2}{7}; \frac{1}{7}; \frac{4}{7}$ 16. $\frac{14}{45}$
17. $\frac{1}{2}$ 18. $\frac{7}{10}$
19. (a) $\frac{59}{165}$ (b) $\frac{26}{165}$ (c) $\frac{16}{33}$
20. $\frac{63}{143}$

...

Theoretical Probability Distribution

7.1 Introduction

The distribution which consists of the data arising either from actual observation or from some experiment is called the frequency distribution. But if it is starting from certain assumption and is deduced mathematically then such (that) distribution is called the theoretical distribution of certain universe. Here, we study only the two types of **theoretical probability distributions** namely Discrete Probability Distribution (Binomial distribution, Poisson distribution, Normal distribution) which are introduced by James Bernoulli (1654-1705) and De-Moivre (1667-1754) respectively.

7.2 Random Variables

A variable whose value depends upon the outcomes of a random experiment is known as a random variable. That is, a random variable is a function whose domain is the sample space of a random experiment and its range is a set of real numbers. The random variable is usually denoted by capital letters and its corresponding values are denoted by their corresponding small letters. Such as if X be a random variable then its values are denoted by $x_1, x_2, x_3, \dots$. For example, if a die is thrown at once then the face which is turned up out of 6 faces marked by the numbers respectively 1, 2, 3, 4, 5, 6, is called the random variable. Suppose we throw a die at once to get the face marked by 3 then the values, it may associate are $p = 1$ (for the face 3 if it turned up) and $q = 0$ (for not getting the face marked by 3 i.e. getting the other faces except the face marked by 3).

There are two types of random variables. They are discrete random variable and continuous random variable.

(a) Discrete Random Variable

A random variable 'X' is said to be discrete either if it consists of (assumes) finite number of values or countably infinite values that can be arranged as a sequence. In other words a random variable which takes only the integral values such as 0, 1, 2, 3, ... is known as "discrete random variable". For example,

- 1) The number of defective bulbs in a basket.
- 2) The number of students in a class, etc.

(b) Continuous Random Variable

A random variable 'X' is said to be continuous if it assumes (consists of) all the values within a certain interval. For examples,

- 1) The duration of a telephone call.
- 2) The temperatures recorded within 24 hours, etc.

7.3 Probability Distribution of a Random Variable

Probability distribution is a table in which different values of random variable are listed with their corresponding probabilities. Such as if $x_1, x_2, x_3, \dots, x_n$ be the values of random variable X and if their corresponding values of probabilities be $P_1, P_2, P_3, \dots, P_n$. Then the probability distribution is formed as below in a table.

Random variable (X)	x_1	x_2	x_3	...	x_n
Probability P(X)	P_1	P_2	P_3	...	P_n

$$\text{Here, } \Sigma P(x) = P_1 + P_2 + P_3 + \dots + P_n = 1$$

Otherwise, the probability distribution is said to be not admissible. Thus, mathematically a set $\{[X, P(x)] ; X \text{ is a random variable}\}$ is called a probability distribution. A function $P(X)$ where $\Sigma P(x) = 1$ is called a probability function. If 'X' be discrete random variable then $P(X)$ is called the probability mass function and the distribution is called the discrete probability distribution.

Example 1.

A coin is tossed twice and if getting head is a success then find the probability distribution of the number of success. And state whether the probability distribution is admissible or not.

Solution:

If a coin is tossed at two times the sample space will be $S = \{HH, HT, TH, TT\}$. Here the random variable 'X' can take the values as 0 (for no head), 1 (for one head) and 2 (for both heads). Thus,

$$\text{if } X = 0, P_1 = P(X = 0) = \frac{1}{4},$$

$$\text{if } X = 1, P_2 = P(X = 1) = \frac{2}{4} \text{ and}$$

$$\text{if } X = 2, P_3 = P(X = 2) = \frac{1}{4}$$

Now, the probability distribution of the discrete random variable X can be expressed as

X	0	1	2
P(X)	1/4	2/4	1/4

$$\text{Here, } \Sigma P(X) = P_1 + P_2 + P_3 = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = \frac{4}{4} = 1$$

Since $\Sigma P(X) = 1$ so the probability distribution is admissible.

Notes:

- (i) If X is a discrete random variable then $\Sigma P(X) = 1$.
- (ii) The expected value (mean) of X i.e. $E(X)$ or $\mu = \Sigma [X \cdot P(X)] = np$
- (iii) The variance of X i.e. $V(X)$ or $(\sigma^2) = \Sigma [(X - \mu)^2 \cdot P(X)]$

$$= \Sigma [X^2 \cdot P(X)] - \mu^2 = npq$$
- (iv) Standard deviation of X i.e. $\sigma = \sqrt{V(X)} = \sqrt{npq}$

Example 2.

By using the following probability distribution, find the expected sales (1000 Rs.) and variance [CTEVT 2073 R/B]

Day	I	II	III	IV	V	VI	VII
Sales	90	60	190	50	75	55	80
Probability	0.25	0.18	0.12	0.05	0.1	0.2	0.1

Solution:

Calculation of expected sales and variance

Day	Sales (x)	Probability P(x)	xP(x)	x ²	x ² .P(x)
I	90	0.25	22.5	8100	2025
II	60	0.18	10.8	3600	648
III	190	0.12	22.8	36100	4332
IV	50	0.05	2.5	2500	125
V	75	0.1	7.5	5625	562.5
VI	55	0.2	11	3025	605
VII	80	0.1	8	6400	640
			$\Sigma x.P(x) = 85.1$		$\Sigma x^2.P(x) = 3937.5$

Here, $\Sigma x.P(x) = 85.1$ and $\Sigma x^2.P(x) = 8937.5$

We have, expected sales ($\bar{X}$) = $\Sigma [x.P(x)] = 85.1 \times 1000$
 $= 85100$

$$\begin{aligned}\text{Variance } (\sigma_x)^2 &= \Sigma x^2.P(x) - [\Sigma x.P(x)]^2 \\ &= 8937.5 - (85.1)^2 \\ &= 8937.5 - 7242.01 = 1695.49\end{aligned}$$

$\therefore$ Expected sale ($\bar{X}$) = 85100 (in 1000 Rs.) and

$$\text{Variance } (\sigma_x)^2 = 1695.49$$

Example 3.

Find the value of 'k' and obtain the mean and variance of a random variable X having following probability distribution:

Value of X(x)	-2	-1	0	1	2	3
P(x)	0.1	k	0.2	2k	0.3	k

Solution:

Since 'X' be discrete random variable so we have

$$\Sigma P(x) = 1$$

i.e. $P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = 1$

or, $0.1 + k + 0.2 + 2k + 0.3 + k = 1$

or, $4k + 0.6 = 1$

or, $4k = 1 - 0.6 = 0.4$

$$\text{or,} \quad k = \frac{0.4}{4} = 0.1$$

$$\therefore k = 0.1$$

After putting the value of k the probability distribution of the random variable X is as below.

$X(x)$	$P(x)$	$x.P(x)$	x^2	$x^2.P(x)$
-2	0.1	-0.2	4	0.4
-1	0.1	-0.1	1	0.1
0	0.2	0	0	0
1	0.2	0.2	1	0.2
2	0.3	0.6	4	1.2
3	0.1	0.3	9	0.9
		$\Sigma xP(x) = 0.8$		$\Sigma x^2.P(x) = 2.8$

$$\text{Here, } \Sigma x.P(x) = 0.8 \text{ and } \Sigma x^2.P(x) = 2.8$$

$$\text{value } (\bar{x}) = ?$$

$$\text{Variance } (\sigma_x)^2 = ?$$

$$\text{We have } \bar{X} = \Sigma x.P(x) = 0.8$$

$$\begin{aligned} \text{And } \sigma_x^2 &= \Sigma x^2.P(x) - [\Sigma x.P(x)]^2 \\ &= 2.8 - (0.8)^2 \\ &= 2.8 - 0.64 = 2.16 \end{aligned}$$

$$\therefore k = 0.1, \bar{x} = 0.8, \sigma_x^2 = 2.16$$

7.4 Binomial Distribution

Binomial distribution was discovered by the Swiss-mathematician James Bernoulli (1654-1705) in the year 1700. If an experiment has two possible outcomes 'success' and 'failure' with probabilities p and q respectively i.e. $p + q = 1$, it is said to have Bernoulli distribution. The process by which a binomial observation is made is called the binomial experiment. A binomial experiment satisfies the following properties:

- The number of trials n is finite.
- Each trial has only two possible outcomes usually called success and failure.
- The trials are independent of each other.
- The probability of success p is constant for each trial.

The problems related to tossing of a coin or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probability distribution. Let us take the following examples:

- If a coin is tossed once there are two outcomes namely head (H) and tail (T), i.e.

H T

The probabilities corresponding to these results are:

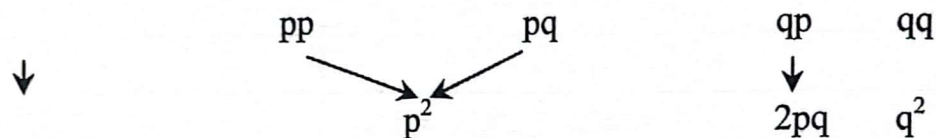
p q

∴ These are the terms of binomial $(p + q)$.

- If two coins are tossed simultaneously there are four possible outcomes

HH HT TH TT

The probabilities corresponding to these results are:



These are the terms of binomial $(p + q)^2$ because

$$(p + q)^2 = p^2 + 2pq + q^2$$

$$= {}^2C_0 p^2 + {}^2C_1 pq + {}^2C_2 q^2$$

- If three coins are tossed, the following are the possible outcomes and the probabilities corresponding to these results are

HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
p^3	p^2q	p^2q	p^2q	pq^2	pq^2	pq^2	q^3

These are the terms of the binomial $(p + q)^3$ because

$$(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$$

$$= {}^3C_0 p^3 + {}^3C_1 p^2q + {}^3C_2 pq^2 + {}^3C_3 q^3$$

In general, in n tosses of a coin the probabilities of the various possible events, i.e. obtaining 0, 1, 2, 3, ..., n heads are given by the successive terms of the binomial expansion $(p + q)^n$, which is

$$(p + q)^n =$$

$${}^nC_0 p^n + {}^nC_1 p^{n-1}q + {}^nC_2 p^{n-2}q^2 + \dots + {}^nC_{n-1} p q^{n-1} + {}^nC_n q^n.$$

Thus the probability of getting r success i.e. $P(r)$ in a set of n independent trials is

$$P(r) = {}^nC_r p^r q^{n-r} \quad (0 \leq r \leq n)$$

The mean and standard deviation:

The mean and standard deviation of the binomial distribution are np and $\sqrt{npq}$ respectively.

Example 4.

If the mean and variance of binomial distribution are 9 and 6 respectively, find the distribution.

Solution:

Let p and q be the probability of success and failure in any one trial and n be the number of trials. Then binomial distribution is $(p + q)^n$.

Now, mean of binomial distribution = $np = 9$ (1)

Variance of binomial distribution = $npq = 6$ (2)

From (1) and (2), we have

$$q = \frac{6}{9} = \frac{2}{3}$$

Also, $p + q = 1 \Rightarrow p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$

From (1),

$$n \times \frac{1}{3} = 9 \therefore n = 27$$

Hence the binomial distribution is $\left(\frac{1}{3} + \frac{2}{3}\right)^{27}$.

Example 5.

What are the chances of getting 2 boys, 2 girls, one boy and one girl when number of pregnancies is 2?

Solution:

Here total number of pregnancies (n) = 2

Probability of getting a boy (p) = $\frac{1}{2}$ $\therefore q = \frac{1}{2}$ [$\because p + q = 1$]

Now, prob. of getting 2 boys

$$P(r = 2) = {}^2C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{2-2} = {}^2C_2 \frac{1}{4} = 1 \times \frac{1}{4} = \frac{1}{4}$$

$$\text{Prob. of getting 2 girls} = P(r = 2) = {}^2C_2 \left(\frac{1}{2}\right)^{2-2} \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

Prob. of getting one boy and one girls

$$P(r=1) = {}^2C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{2-1} = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

Example 6.

If three dice are thrown, what is the probability of getting (a) exactly 3 sixes (b) exactly 2 sixes

Solution:

No. of trials $n = 3$

Prob. of getting a six $(p) = \frac{1}{6}$

$$\therefore q = 1 - \frac{1}{6} = \frac{5}{6}$$

Now,

The prob. of getting exactly 3 sixes

$$\begin{aligned} P(r=3) &= {}^3C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{3-3} = {}^3C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^0 \\ &= 1 \times \frac{1}{216} \times 1 = \frac{1}{216} \end{aligned}$$

Again,

The probability of getting exactly 2 sixes

$$\begin{aligned} P(r=2) &= {}^3C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{3-2} = {}^3C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^1 \\ &= 3 \times \frac{1}{36} \times \frac{5}{6} = \frac{5}{72} \end{aligned}$$

Example 7.

If 5% of the syringe manufactured by a company are defective, what is the probability that out of 3 syringe choosen at random (a) none of them is defective (b) exactly one of them is defective (c) at least one of them is defective.

Solution:

Let probability of defective syringe $(p) = 5\% = \frac{5}{100} = \frac{1}{20}$

$$\therefore q = 1 - \frac{1}{20} = \frac{19}{20}$$

No. of trials = 3

Probability of getting r successes in n trials $P(r) = {}^nC_r p^r q^{n-r}$

- (a) The probability that none of syringe is defective

$$P(r = 0) = {}^3C_0 \left(\frac{1}{20}\right)^0 \left(\frac{19}{20}\right)^3 = 0.8574$$

- (b) The probability that exactly one syringe is defective

$$\begin{aligned} P(r = 1) &= {}^3C_1 p q^2 \\ &= 3 \cdot \left(\frac{1}{20}\right) \left(\frac{19}{20}\right)^2 = \frac{3 \times (19)^2}{(20)^3} = 0.1354 \end{aligned}$$

- (c) The probability that at least one of syringe is defective

$$\begin{aligned} P(r \geq 1) &= 1 - P(r = 0) \\ &= 1 - \left(\frac{19}{20}\right)^3 = 0.1426 \end{aligned}$$

Example 8.

A coin is tossed three times. Find the probability of getting

- (a) exactly two heads
(b) at least one head.

Solution:

Number of trials (n) = 3

Probability of getting head (p) = $\frac{1}{2}$

$$\therefore q = 1 - p = \frac{1}{2}$$

Now,

- (a) Probability of getting exactly two heads $P(r = 2)$

$$= {}^3C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{3-2} = \frac{3!}{(3-2)!2!} \times \frac{1}{4} \times \frac{1}{2} = \frac{3}{8}$$

- (b) Probability of getting at least one head

$$\begin{aligned} P(r \geq 1) &= 1 - P(r = 0) \\ &= 1 - {}^3C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^3 = 1 - 1 \cdot 1 \cdot \frac{1}{8} \\ &= 1 - \frac{1}{8} = \frac{7}{8} \end{aligned}$$

Example 9.

The probability of hitting a target is $\frac{1}{5}$. If six hitting are made, find the probability that (a) none will strike the target, (b) exactly two will strike the target.

Solution:

Let the probability of strike the target (p) = $\frac{1}{5}$

$$\therefore q = 1 - p = 1 - \frac{1}{5} = \frac{4}{5}$$

No. of hitting (n) = 6

Probability of successful hitting = $P(r) = {}^nC_r p^r q^{n-r}$

Now,

$$\begin{aligned} \text{(a) Probability that none will strike the target } P(r=0) &= {}^6C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^6 \\ &= 1.1 \cdot \frac{4096}{15625} = \frac{4096}{15625} \end{aligned}$$

(b) Probability that exactly two strike the target

$$\begin{aligned} P(r=2) &= {}^6C_2 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^{6-2} = \frac{6!}{(6-2)!2!} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^4 \\ &= \frac{6 \cdot 5 \cdot 4!}{4! \cdot 2 \cdot 1} \times \frac{1}{25} \times \frac{256}{3125} = \frac{768}{3125} \end{aligned}$$

EXERCISE - 7(I)

1. Determine whether the following probability distributions are admissible or not. Give reason.

(a)

X	0	1	2	3	4
P(X)	0.2	0.1	0.4	0.2	0.1

(b)

X	-2	-1	0	1	2
P(X)	0.4	0.3	0.1	0.1	0.2

2. Compute the expectation, variance and the standard deviation from the probability distributions below of a random variable X:

(a)

X	-1	0	1	2
P(X)	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

(b)

X	0	1	2	3
P(X)	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{2}{8}$

3. For what value of k, the following probability distribution of a random variable X, becomes admissible. Also find mean, variance and the standard deviation.

X	-2	-1	0	1	2	3
P(X)	0.1	0.3	0.1	2k	k	0.2

4. Obtain the mean, variance and the standard deviation of the following probability distribution of a random variable (sales) X:

Monthly sales (X)	50	100	150	200	250	300
P(X)	0.10	0.30	0.30	0.15	0.10	0.05

5. If the probability of persol's earning profit of Rs. 40,000 is 0.8 and that of suffering a loss of Rs. 10,000 is 0.2 then find his mathematical expectation.
6. In the given discrete probability distribution, find
- the value of 'k'
 - the mean (or expected value)
 - $P(1 < x \leq 4)$
 - the variance
 - standard deviation.
7. In a binomial distribution if $p = 0.6$ and $n = 50$, find the mean.
8. In a binomial distribution, if $n = 400$, $p = 0.1$, find the standard deviation.
9. If $q = \frac{1}{2}$ and $n = 4$, find the mean and the standard deviation in a binomial distribution.
10. The mean of the binomial distribution is 40 and the standard deviation is 6. Find the value of n, p and q.
11. Find the binomial distribution having mean and variance are 12 and 8 respectively.
12. Find the probability of getting two heads in 5 tosses of two coins.
13. Four coins are tossed simultaneously. What is the probability of getting (a) 2 heads and 2 tails (b) at least 2 heads (c) all 4 heads (d) no head.

14. A coin is tossed three times. Find the probability of getting
 - (a) all three heads
 - (b) at least one heads.
15. The incidence of occupation disease in an industry is such that the workman has a 20% chance of suffering it, what is the probability that out of six workmen four or more will contract the disease?
If 20% of the injections manufactured by a company are defective, find the probability that out of 4 injections chosen at random (a) 1 (b) 0 (c) at most 2 injections will be defective.
16. Suppose that in a certain city 52% of all recorded births are males. Suppose we select 5 birth records from population. What is the probability that:
 - (a) exactly three of them are males?
 - (b) less or equal to 2 are males?
17. The probability of hitting a target is $\frac{1}{4}$. If 5 hitting are made, find the probability that
 - (a) none will strike the target
 - (b) exactly one will strike the target.
18. The probability of a HIV patient suffering from TB in his AIDs stage is $\frac{1}{7}$. If 4 patients are selected, find the probability that
 - (a) exactly 3 are suffering from TB
 - (b) at least 3 are suffering from TB.
19. A certain manufacturing process produce syringe of which 15% are defective. Find the probability that in a sample of 10 syringe selected at random there will be
 - (a) no defective
 - (b) at least one defective
 - (c) not more than one defective
20. A sample of 100 fuses is known to have an average 5 defective fuses. Three fuses of the sample are tested. What is the probability that
 - (a) none of them is defective?
 - (b) exactly one of them is defective?

ANSWERS

1. (a) Yes, $\Sigma P(X) = 1$, (b) No, $\Sigma P(x) > 1$

2. (a) $0.5, \frac{19}{12}, 1.258$ (b) $\frac{7}{4}, \frac{15}{16}, 0.97$
3. $K = 0.1$, mean $(\mu) = 0.5$, $\sigma^2 = 2.85$, $\sigma = 1.688$
4. $\bar{X} = E(X) = 150$, $\sigma_x^2 = 4250$, $\sigma_x = 65.19$
5. Rs. 30,000
6. $k = 0.025$ (b) 3.7 (c) 0.55
(d) 3.81 (e) 1.95
7. 30 8. 6 9. 2; 3.162
10. 400; 0.1; 0.9 11. $\left(\frac{2}{3} + \frac{1}{3}\right)^{36}$ 12. $\frac{5}{16}$
13. (a) $\frac{3}{8}$ (b) $\frac{11}{16}$ (c) $\frac{1}{16}$ (d) $\frac{1}{16}$
14. (a) $\frac{1}{8}$ (b) $\frac{7}{8}$ 15. $\frac{53}{3125}$
16. (a) $\frac{256}{625}$ (b) $\frac{256}{625}$ (c) $\frac{608}{625}$
17. (a) 0.32 (b) 0.4625
18. (a) $\frac{243}{1024}$ (b) $\frac{405}{1024}$
19. (a) $\frac{24}{2401}$ (b) $\frac{25}{2401}$
20. (a) 0.1969 (b) 0.8031 (c) 0.5443
21. (a) $\frac{6859}{8000}$ (b) $\frac{1083}{8000}$

7.5 Poisson Distribution

In binomial distribution if total number of trials is very large i.e. $n \rightarrow \infty$ and number of success (P) is very small then $nP = 'k'$ a finite number, then the distribution is called the poisson distribution. It is used to indicate (describe) the rare events as typing mistakes, twin births, road accidents etc. If 'X' be a random variable and follows a poisson distribution then

$$P(X = r) = \frac{e^{-k} k^r}{r!}, r = 0, 1, 2, 3, \dots$$

Note: For a poisson distribution $\mu = \sigma^2 = k$

Example 1.

Usually there are 5 mistakes on average per page while typing the material of mathematics. Find the probabilities of (a) 4 mistake (b) 7 mistakes (c) no. mistake (d) at least 1 mistake (e) not more than 4 mistakes, all in per page.

Solution:

Let X be the number of mistakes per page and also suppose that X follows poisson distribution. Then, here $k = 5$

$$\text{We have, } P(X = r) = \frac{e^{-k}(k)^r}{r!}, r = 0, 1, 2, 3, \dots$$

Thus, (a) $P(4 \text{ mistakes per page})$

$$\begin{aligned} P(X = 4) &= \frac{e^{-5}(5)^4}{4!} = e^{-5} \times \frac{5 \times 5 \times 5 \times 5}{4 \times 3 \times 2 \times 1} \\ &= e^{-5} \times \frac{25 \times 25}{12 \times 2} = \frac{625}{24 \times e^5} \\ &= \frac{625}{24 \times (2.72)^5} = \frac{625}{24 \times 148.90} \\ &= \frac{625}{3573.60} \\ &= 0.175 \end{aligned}$$

(b) $P(7 \text{ mistakes})$

$$\begin{aligned} P(X = 7) &= \frac{e^{-5}(5)^7}{7!} = \bar{X}(Y - \bar{Y}) \\ &= \frac{15625}{148.90 \times 1008} = \frac{15625}{150091.2} \\ &= 0.104 \end{aligned}$$

(c) $P(\text{no mistake})$

$$\begin{aligned} P(X = 0) &= \frac{e^{-5}(5)^0}{0!} = \frac{1}{e^5 \times 1} \\ &= \frac{1}{148.90} = 0.0067 \end{aligned}$$

(d) $P(\text{at least 1 mistake})$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.0067 = 0.993$$

$$\begin{aligned}
 (e) \quad P(X \leq 4) &= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \\
 &= \frac{e^{-5}(5)^0}{0!} + \frac{e^{-5}(5)^1}{1!} + \frac{e^{-5}(5)^2}{2!} + \frac{e^{-5}(5)^3}{3!} + \frac{e^{-5}(5)^4}{4!} \\
 &= e^{-5} \left[1 + 5 + \frac{25}{2} + \frac{125}{6} + \frac{25}{24} \right] \\
 &= e^{-5} [1 + 5 + 12.5 + 20.84 + 26.042] \\
 &= \frac{65.382}{148.90} = 0.439
 \end{aligned}$$

7.6 Normal distribution

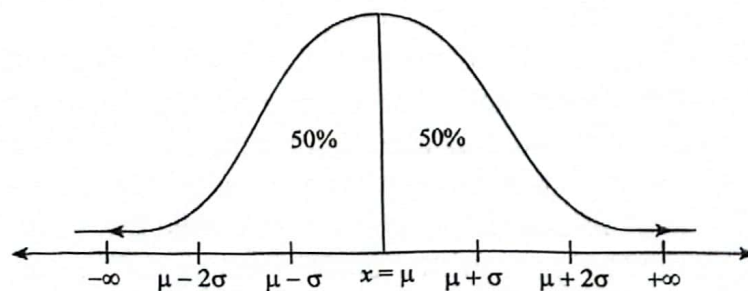
The normal distribution is the limiting condition (case) of the binomial distribution when the number of trials ' n ' becomes large enough and the values of p and q (i.e. probability of success and failure respectively) are almost equal. A continuous random variables X is said to have a normal distribution with parameter μ and σ^2 if its probability density function is given by

$$P(a \leq X \leq b) = \int_a^b P(X) dx$$

$$\text{where } P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \quad (1)$$

Where, $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$ is said to possess a normal distribution. π and e are the constants such that $\pi = \frac{22}{7}$, $2 < e < 3$ and

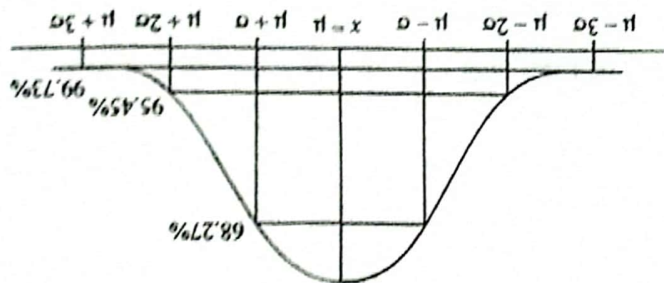
$X \sim N(\mu, \sigma^2)$ is used to denote the normal distribution with mean $(\mu) = np$ and variance $(\sigma^2) = npq$. The curve of the normal distribution $P(X)$ is bell shaped and symmetrical about the line $X = \mu$ which is shown as below:



Some properties of normal distribution

- The curve is symmetrical about the line mean = μ .
- Since the curve is symmetrical so, mean = median = mode.
- X-axis is an asymptote to the normal curve.
- The curve has unique model value.
- The maximum height of the curve divides the total area into two equal parts i.e. the area right to the line $\bar{X} = \mu$ is equal to the area left to the line $\bar{X} = \mu = 0.5$.
- The area under the curve represents the value of the probability where we have the property:

Range of X	Range of Z	Area covered
$X = \mu \pm \sigma$	$Z = \pm 1$	68.27% i.e. $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6827$
$X = \mu \pm 2\sigma$	$Z = \pm 2$	95.45% i.e. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
$X = \mu \pm 3\sigma$	$Z = \pm 3$	99.73% i.e. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$



Standard normal distribution

If $Z = \frac{X - \mu}{\sigma}$ then the above equation (1) reduces as

$$P(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2} \quad (2)$$

which is known as the standard normal distribution where mean (μ) = 0, variance (σ^2) = 1, X be a normal random variable and the random variable Z is said to follow a standard normal variate (distribution). The standard normal curve is symmetrical about the line $x = 0$ (i.e. Y-axis). The total area under the curve (1) or (2) with X-axis is 1 which represents the probability of the normal distribution. The probability corresponding to the area under the normal curve given by (2) can be obtained by using the standard normal area table which is given as below:

Method of obtaining the area

The standard normal variate is given by

$$Z = \frac{X - \mu}{\sigma}$$

where X = value of random variable

μ = mean of normal distribution

σ = standard deviation of the normal distribution

Z = standard normal variate corresponding to the normal variate X

Since the normal curve is symmetrical about the mean so the line

mean = m (i.e. $X = \mu$) or $Z = 0$ divides the entire area into two equal parts.

That is the area between $Z = 0$ and $Z = a$ is same as the area between $Z = -a$ and $Z = 0$. Thus, $P(-a < Z < 0) = P(0 < Z < a)$.

Consider $Z = z_1$ and $Z = z_2$ both are positive and

A_1 = area between $Z = 0$ and $Z = z_1$ and

A_2 = area between $Z = 0$ and $Z = z_2$ then we have following

procedures:

(a) For $P(Z = z_1)$ = probability for $Z = z_1$

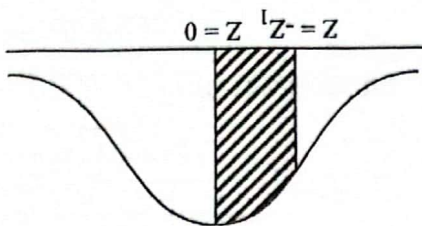
= area between $Z = 0$

and $Z = z_1$

(b) For $P(Z = -z_1)$ = probability of $Z = -z_1$

= area between $Z = -z_1$

and $Z = 0$



(c) For $P(z_1 \leq Z \leq z_2)$

= $P(0 \leq Z \leq z_2) -$

$P(0 \leq Z \leq z_1)$

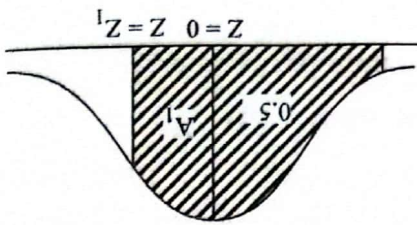
= $A_2 - A_1$

Similar result holds for the negative z_1 and z_2 .

(d) For $P(Z \leq z_1) = P(Z \leq 0) +$

$P(0 \leq Z \leq z_1)$

= $0.5 + A_1$

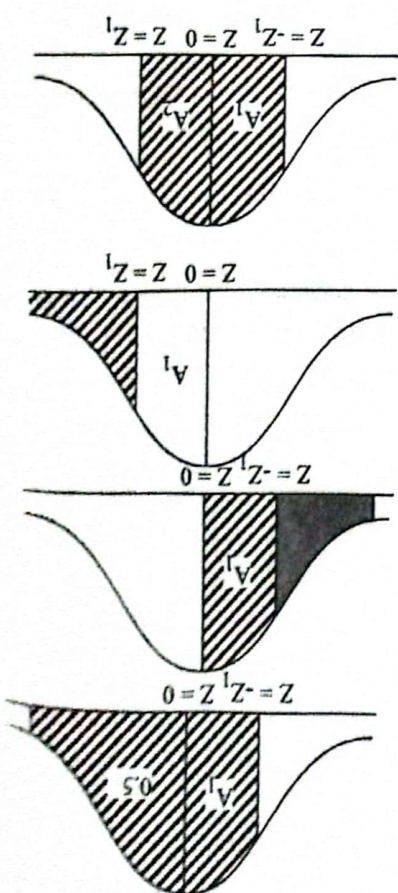


$$(c) \quad \text{For } P(Z \geq -z_1) = P(-z_1 \leq Z \leq 0) + P(Z \geq 0) \\ = A_1 + 0.5$$

$$(f) \quad \text{For } P(Z \leq -z_1) = P(Z \leq 0) - P(-z_1 \leq Z \leq 0) \\ = 0.5 - A_1$$

$$(g) \quad \text{For } P(Z \leq z_1) = P(Z \geq 0) - P(0 \leq Z \leq z_1) \\ = 0.5 - A_1$$

$$(h) \quad \text{For } P(-z_1 \leq Z \leq z_2) \\ = P(-z_1 \leq Z \leq 0) + P(0 \leq Z \leq z_2) \\ = A_1 + A_2$$



Note: The area under the normal curve between the standard variate $Z = 0$ (mean) and $Z = a$ gives the probability of getting a value (or proportional of cases falling) between $Z = 0$ and $Z = a$.

Example 2.

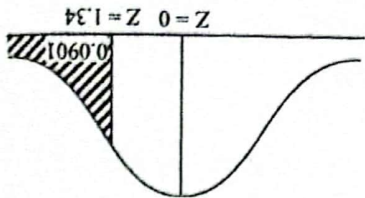
Find the area under the normal curve to the right of $Z = 1.34$

Solution:

The area under the normal curve for $Z = 1.34$ means the area under the curve between $Z = 0$ and $Z = 1.34$. On the area table at first see the column of 1.3 and go along its row to meet the column of 0.04. The interesting value = 0.4099. Also the area under the curve to the right of $Z = 0$ is 0.5. Thus the required area under the normal curve to the right of $Z = 1.34$ is

$$0.5 - 0.4099 = 0.0901$$

The required area is shaded alongside in the figure.



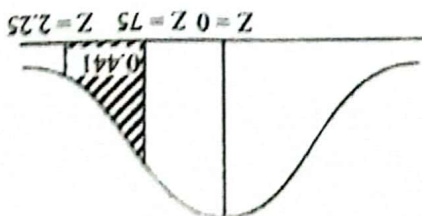
Example 3.

Find the area under the curve between $Z = 0.75$ and $Z = 2.25$.

Solution:

Area under the curve between $Z = 0$ and $Z = 0.75$ is 0.2734 (from table)

Again, the area under the curve between $Z = 0$ and $Z = 2.25$ is 0.4875 (from table)



Now, the required area under the curve between $Z = 0.75$ and $Z = 2.25$ is

$$= 0.4875 - 0.2734 = 0.2141$$

The required area is shown in the figure alongside.

Example 4.

A normal distribution with mean 50 and standard deviation 8, find the probability that the normal variate X assumes a value between 34 and 62.

Solution:

Here, mean (μ) = 50

s.d. (σ) = 8

The standard normal variate (Z)

$$= \frac{X - \mu}{\sigma} = \frac{X - 50}{8}$$

$$\text{For } X = 34, \quad Z = \frac{34 - 50}{8} = \frac{-16}{8} = -2.0$$

$$\text{For } X = 62, \quad Z = \frac{62 - 50}{8} = \frac{12}{8} = 1.5$$

Now, the required value of the probability = Area under the normal curve between

$$P(34 \leq X \leq 62) = P(-2 \leq Z \leq 1.5)$$

$$= P(-2.0 \leq Z \leq 0) + P(0 \leq Z \leq 1.5)$$

$$= \text{Area between } Z = 0, Z = -2.0 + \text{Area between } Z = 0, Z = 1.5$$

$$= 0.4772 + 0.4332 \quad (\text{from the table})$$

$$= 0.9104$$

Example 5.

The weekly wages of 1000 workers are normally distributed around a mean of Rs. 70 and with a standard deviation of Rs. 5. Estimate the number of workers whose weekly wage will be:

- (a) between Rs. 70 and Rs. 72
- (b) more than Rs. 75.

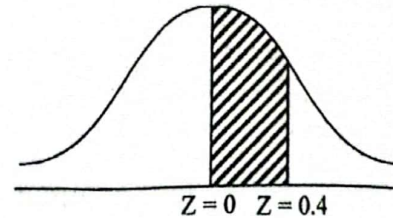
Solution:

Here, mean (μ) = Rs. 70

s.d. (σ) = Rs. 5

Standard normal variate

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 70}{5} \quad (1)$$



(a) When $X = 70$, then $Z = \frac{70 - 70}{5} = \frac{0}{5} = 0$ i.e. $Z = 0$

When $X = 72$, then $Z = \frac{72 - 70}{5} = \frac{2}{5} = 0.4$

i.e. $Z = 0.4$

Now, $P(70 \leq X \leq 72) = P(0 \leq Z \leq 0.4)$
 $= 0.1554$ (from table)

$\therefore$ No. of workers with wages between Rs. 70 to Rs. 72
 $= 1000 \times 0.1554 = 155.4 = 155$

(b) Here, for $X = 75$, $Z = \frac{75 - 70}{5} = \frac{5}{5} = 1.0$ i.e. $Z = 1.0$

Now, area between $Z = 0$ and $Z = 1.0$ is
 $= P(0 \leq Z \leq 1) = 0.3413$

For $P(X > 75) = P(Z > 1)$
 $= P(Z \leq 0) - P(0 \leq Z \leq 1)$
 $= 0.5 - 0.3413 = 0.1587$

$\therefore$ No. of workers having wages greater than Rs. 75
 $= 1000 \times 0.1587 = 158.7$
 $= \text{Rs. } 159$

EXERCISE - 7(II)

1. The average road accidents per week in KTM be 4, find the probability of (a) 3 accidents (b) 6 accidents (c) at least 2 accidents (d) no accident (e) not more than 3 accidents all per week.
2. A local telephone booth receives 2 calls per minute. Find the probability that there will be (a) 3 calls (b) more than 3 calls (c) no call, all in given time period.
3. Find the area under the normal curve for $Z = 0.74$.
4. Find the area under the normal curve between $Z = -0.45$ and $Z = 1.65$.
5. Find the standard normal variate Z if the area under the normal curve between 0 and Z is 0.4222.
6. If a random variable X follow a normal distribution with mean 200 and the standard deviation is 20, find the probability that:
(a) $P(200 \leq X \leq 220)$ (b) $P(X \leq 220)$ (c) $P(X > 240)$.
7. If a standard normal variate Z follows a standard normal distribution, find
(a) $P(0 < Z < 2.12)$ (b) $P(-2 < Z < 0)$
(c) $P(-1.96 < Z \leq 1.96)$
8. If 49% of a normal population lies between $\bar{X} = \mu = 83$ and a value of $X = 99$, find the standard deviation (σ).
9. The weekly wages of 1000 workers are normally distributed around a mean of Rs. 70 and with a standard deviation of Rs. 5. Estimate the number of workers whose weekly wage will be:
(a) between Rs. 69 and 72 (b) less than Rs. 63.
10. The average test mark in a particular class is 79. The standard deviation is 5. If the marks are distributed normally, how many students in the class of 200 did not receive marks between 75 and 82? Given $P(0 \leq Z \leq 0.6) = Z_{0.6} = 0.2257$, $P(0 \leq Z \leq 0.8) = Z_{0.8} = 0.2881$ where Z is a standard normal variate.
11. The marks obtained by students are normally distributed with mean 60 and standard deviation 10. Find the probability that a randomly selected student will have marks (i) less than 70 (ii) more than 90 (iii) between 70 and 60 (iv) below 45 (v) above 50 (vi) between 50 and 70.
12. The life duration of a calculator manufactured by a company is normally distributed with mean 54 months and the standard deviation is 8 months. Find the probability that a calculator chosen randomly will have life duration less than 36 months.

ANSWERS

- | | | | |
|-----|------------|-------------|-----------------------|
| 1. | (a) 0.1954 | (b) 0.1042 | (c) 0.9084 |
| | (d) 0.0183 | (e) 0.4335 | |
| 2. | (a) 0.1804 | (b) 0.1429 | (c) 0.1353 |
| 3. | 0.2704 | 4. 0.6241 | 5. $Z_1 = 1.42$ |
| 6. | (a) 0.3413 | (b) 0.8413 | (c) 0.0228 |
| 7. | (a) 0.4830 | (b) 0.4722 | (c) 0.95 |
| 8. | 6.87 | | |
| 9. | (a) 235 | (b) 81 | 10. 97 |
| 11. | (a) 0.8413 | (b) 0.00135 | (c) 0.1574 (d) 0.0668 |
| | (e) 0.8413 | (f) 0.6826 | |
| 12. | 0.0122 | | |

● ● ●

Bivariate Data Analysis

8.1 Introduction

So far we have confined our discussion to the distribution involving only one variable (univariate distribution) and also saw how the various measures of central tendency, dispersion and shape of frequency distribution or curve can be used for the purpose of comparison and analysis. We may, however, come across certain series where each item of the series may assume the values of two or more variables.

Sometimes two continuous characters are measured in the same person such as weight and cholesterol, weight and height, age and weight, age and height etc. At other times, the same character is measured in two related groups such as tallness in parents and tallness in children, study of intelligent quotient (IQ) in brothers and in corresponding sisters and so on. The relationship between two quantitatively measured variables are called correlation and regression. *Thus the correlation and regression are statistical tools which study the relationship between two or more than two variables.*

8.2 Correlation and Its Types

Correlation is described in different ways. Three of most important ways of classifying correlation are:

1. Positive and negative correlation
2. Linear and non-linear correlation
3. Simple, partial and multiple correlation

1. Positive and negative correlation

The correlation between two variables is said to be positive, if the increase (or decrease) in the value of one variable is accompanied by an increase (or decrease) in the value of the other variable. For example, consider the data

X:	5	9	17	20	32
Y:	3	6	8	15	16

Here the value of Y has increased corresponding to every increasing value of X. The correlation between the variables X and Y is positive. The following data also reveals the positive correlation between two variables X and Y.

X:	27	16	13	9	5
Y:	9	7	4	3	1

The correlation between two variables is said to be negative, if increase (or decrease) in the value of one variable is accompanied by a decrease (or increase) in the value of the other variable. For example, consider the data

X:	115	107	104	98	70
Y:	25	27	43	49	64

Here a decrease in the value of X is accompanied by an increase in the value of Y. The correlation between X and Y is negative.

2. Linear and non-linear correlation

The correlation between two variables is said to be linear, if the ratio of change in one variable to the change in the other variable is almost constant. For example, consider the data,

X:	1	2	3	4	5
Y:	9	11	13	15	17

Here the variables X and Y is linear.

The correlation between two variables is said to be non-linear, if the ratio of change in one variable to the change in the other variable is not constant. Let us consider the following data:

X:	1	2	3	4	5
Y:	7	10	23	30	42

Here the correlation between the variables X and Y is non-linear.

3. Simple, multiple and partial correlation

The correlation between two variables is known as simple correlation, when three or more variables are considered, then the correlation may be multiple or partial. The degree of the relationship among three or more variables at a time is called multiple correlation. In other words, in multiple correlation all the given variables are studied at a time by keeping one variable as dependent and all the remaining variables as independent. Thus in multiple correlation three or more variables studied simultaneously. For example, when we study the relationship between the yield of rice per hectare and both the amount of rainfall and the amount of fertilizers used, it is a problem of multiple correlation.

In partial correlation, there will be three or more variables. The degree of relationship between dependent variable and one independent variable by keeping all other independent variables as constant is called

partial correlation. For example, the achievement of the child in the school is affected by the age, sex, IQ, number of hours studied at home, quality of teaching etc. When we study the relationship between the achievement of the child and age by keeping all other variable as constant is a problem of partial correlations.

8.3 Methods of Measuring Correlation

The following are the methods of measuring correlation between two variables.

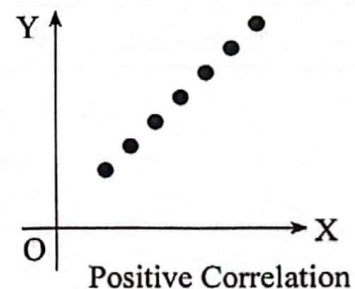
1. Scatter diagram method
2. Karl Pearson's method
3. Spearman's rank correlation

Scatter diagram method:

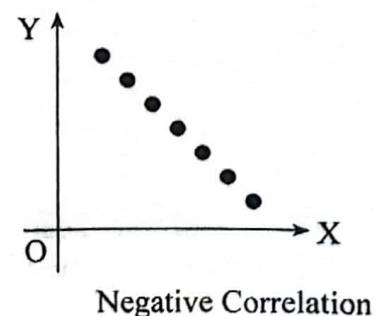
Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be n pairs of values of two variables X and Y with respect to some characteristic. If we plot the corresponding values of X and Y , taking the values of X along x -axis and the values of Y along y -axis, we obtain a collection of dots, called dot diagram or a scatter diagram.

The correlation between two variables by scatter diagram can be analyzed in the following manner:

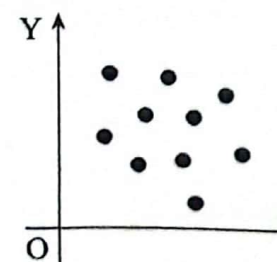
- (a) If the plotted points are in a straight line raising from lower left to upper right, then there is positive correlation between the variables.



- (b) If the plotted points are in a straight line falling from upper left to lower right, then there is negative correlation between the variables.



- (c) If the plotted points show no trend at all, then the variables are independent and are not correlated.



Merits and demerits of scatter diagram method**Merits**

1. This is the simplest method.
2. This method is most suitable, if just an idea is to be formed whether the variables are correlated or not.
3. This method is least affected by the extreme items.
4. Scatter diagrams are attractive and can be used for in exhibition displaying the correlation between variables.

Demerits

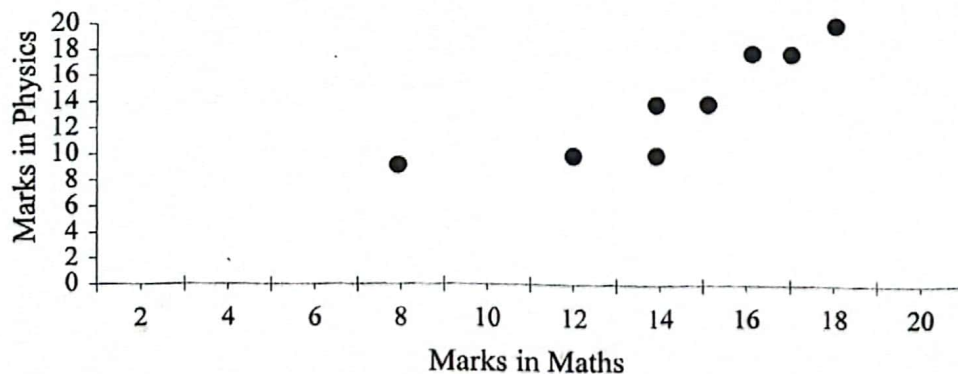
1. This method is subjective in nature.
2. The degree of correlation assessed, cannot be numerically expressed.

Example 1.

Draw a scatter diagram for the following data and comment on the presence of correlation between the variables:

Marks in maths	8	12	16	15	14	17	18
Marks in physics	9	10	18	14	14	18	20

Scatter Diagram



From the scatter diagram, we see that there is positive linear correlation between the marks obtain in mathematics and physics.

8.4 Covariance

Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$ be n pairs of values of

two variables X and Y . The measure $\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n}$ is defined as the

covariance between the variables X and Y . This is denoted by $\text{cov}(X, Y)$. For simplicity, we write

$$\text{Cov}(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} \quad \text{This formula is used when } \bar{X} \text{ and } \bar{Y} \text{ are in whole number.}$$

$\bar{Y}$ are in whole number.

Let us derive another expression for covariance.

We know that

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{1}{n} \{ \sum (x_i - \bar{x})(y_i - \bar{y}) \} \\ &= \frac{1}{n} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \} \\ &= \frac{1}{n} \{ x_1 y_1 - x_1 \bar{y} - \bar{x} y_1 + \bar{x} \bar{y} + x_2 y_2 - x_2 \bar{y} - \bar{x} y_2 + \bar{x} \bar{y} + \dots + x_n y_n - \bar{x} y_n - x_n \bar{y} + \bar{x} \bar{y} \} \\ &= \frac{1}{n} \{ (x_1 y_1 + x_2 y_2 + \dots + x_n y_n) - (x_1 + x_2 + \dots + x_n) \bar{y} - (\bar{y} y_1 + \bar{y} y_2 + \dots + \bar{y} y_n) + n \bar{x} \bar{y} \} \\ &= \frac{1}{n} \left\{ \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{y} y_i + n \bar{x} \bar{y} \right\} \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \sum_{i=1}^n \frac{x_i}{n} \bar{y} - \sum_{i=1}^n \frac{y_i}{n} \bar{x} + \bar{x} \bar{y} \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} - \bar{x} \bar{y} + \bar{x} \bar{y} \quad \left[\because \sum_{i=1}^n \frac{x_i}{n} = \bar{x} \text{ \& } \sum_{i=1}^n \frac{y_i}{n} = \bar{y} \right] \\ \therefore \text{Cov}(X, Y) &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \end{aligned}$$

This formula is used when the $\bar{x}$ and $\bar{y}$ are not whole number (i.e. if $\bar{x}$, $\bar{y}$ are in fraction or decimals).

Example 2.

Find $\text{Cov}(X, Y)$ from the following data:

X	15	20	25	30	40	50
Y	44	43	45	37	34	37

Solution:**CALCULATION OF COVARIANCE**

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
15	44	-15	4	-60
20	43	-10	3	-30
25	45	-5	5	-25
30	37	0	-3	0
40	34	10	-6	-60
50	37	20	-3	-60
$\Sigma X = 180$	$\Sigma Y = 240$			$\Sigma(X - \bar{X})(Y - \bar{Y}) = -235$

Now,

$$\bar{X} = \frac{\Sigma x}{N} = \frac{180}{6} = 30$$

$$\bar{Y} = \frac{\Sigma y}{N} = \frac{240}{6} = 40$$

$$\therefore \text{Cov}(X, Y) = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n} = \frac{-235}{6} = -39.167.$$

Example 3.

Find the covariance between x and y for the following observations of (x, y):

(2, 4), (10, 15), (20, 12), (10, 16), (17, 17), (12, 10), (15, 11), (16, 8).

Solution:**CALCULATION OF COVARIANCE**

X	Y	XY
2	4	8
10	15	150
20	12	240
10	16	160
17	17	289
12	10	120
15	11	165
16	8	128
$\Sigma X = 102$	$\Sigma Y = 93$	$\Sigma XY = 1260$

We know that

$$\bar{X} = \frac{\Sigma X}{n} = \frac{102}{8} = 12.75$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{93}{8} = 11.63$$

Since $\bar{x}$ and $\bar{y}$ are not whole number so,

$$\begin{aligned}\text{Cov}(X, Y) &= \frac{1}{n} \Sigma XY - \bar{X} \bar{Y} = \frac{1}{8} (1260) - (12.75) (11.63) \\ &= 9.29\end{aligned}$$

8.5 Karl Pearson's Coefficient of Correlation

A mathematical method for measuring the degree of linear relationship between two variables was given by Karl Pearson (1867 – 1936), a great British Bio-matrician and statistician. Karl Pearson's measure of correlation coefficient between two variables X and Y , usually denoted by r_{XY} or simply r , is a numerical measure of linear relationship between them and is defined as the ratio of the $\text{cov}(X, Y)$ to the product of square roots of variances of X and Y .

$$\therefore r_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X} \sqrt{\text{Var}Y}} \quad (1)$$

$$\text{Since } \text{Cov}(X, Y) = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n}$$

$$\text{Var}(X) = \frac{\Sigma(X - \bar{X})^2}{n}$$

$$\text{and } \text{Var}(Y) = \frac{\Sigma(Y - \bar{Y})^2}{n}$$

From (1), we write

$$\begin{aligned}r_{XY} &= \frac{\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n}}{\sqrt{\frac{\Sigma(X - \bar{X})^2}{n}} \sqrt{\frac{\Sigma(Y - \bar{Y})^2}{n}}} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2} \sqrt{\Sigma(Y - \bar{Y})^2}} \\ &\quad (2)\end{aligned}$$

For simplicity, relation (2) can be written as

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}} \quad (3)$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

From (2) after simplification we can write

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2} \sqrt{N\sum Y^2 - (\sum Y)^2}} \quad (4)$$

This is known as **direct method**. The formula (4) facilitates to compute the coefficient of correlation of the ungrouped data of small numbers directly from the original data.

Example 4.

Find the coefficient of correlation between x and y , when $\text{Cov}(X, Y) = 16.5$, $\text{Var}(X) = 2.89$ and $\text{Var}(Y) = 100$.

Solution:

$$\begin{aligned} \text{We have, } r &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{16.5}{\sqrt{2.89} \sqrt{100}} = \frac{16.5}{\sqrt{289}} \\ &= \frac{16.5}{17} \quad \therefore r = 0.97 \end{aligned}$$

Example 5.

Find out the coefficient of correlation between the heights of the fathers and sons:

Height of fathers(in inches):	65	66	67	68	69	70	71
Height of sons(in inches):	67	68	66	69	72	72	69

Solution:

COMPUTATION OF CORRELATION COEFFICIENT

Height of fathers X	Height of sons Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	66	-1	-3	1	9	3
68	69	0	0	0	0	0
69	72	1	3	1	9	3
70	72	2	3	4	9	6
71	69	3	0	9	0	0
$\Sigma X = 476$	$\Sigma Y = 483$			$\Sigma x^2 = 28$	$\Sigma y^2 = 32$	$\Sigma xy = 20$

Now,

Average height of fathers

$$\bar{X} = \frac{\Sigma X}{N} = \frac{476}{7} = 68$$

Average height of sons

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{483}{7} = 69$$

We know,

$$\begin{aligned} \text{Correlation coefficient } (r_{XY}) &= \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}} \\ &= \frac{20}{\sqrt{28} \sqrt{32}} = \frac{20}{\sqrt{896}} = \frac{20}{29.93} \\ &= 0.67 \end{aligned}$$

Therefore, there is positive correlation between the heights of the fathers and sons.

Example 6.

Calculate the coefficient of correlation between X and Y series from the following data:

	Series X	Series Y
No. of observations	20	20
s.d.	3.25	3.5

$$\Sigma(X - \bar{X})(Y - \bar{Y}) = 120$$

Solution:

We have,

$$\begin{aligned} r &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var } X} \sqrt{\text{Var } Y}} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{n \sigma_X \sigma_Y} \\ &= \frac{120}{20(3.25)(3.5)} = \frac{120}{227.5} = 0.53 \end{aligned}$$

Example 7.

Calculate the coefficient of correlation from the following data:

X	5	2	10	4	7	8	17	6
Y	9	13	5	7	11	7	7	12

Solution:

COMPUTATION OF CORRELATION COEFFICIENT

X	Y	X ²	Y ²	XY
5	9	25	81	45
2	13	4	169	26
10	5	100	25	50
4	7	16	49	28
7	11	49	121	77
8	7	64	49	56
17	7	289	49	119
6	12	36	144	72
$\Sigma X = 59$	$\Sigma Y = 71$	$\Sigma X^2 = 583$	$\Sigma Y^2 = 687$	$\Sigma XY = 473$

$$\begin{aligned}
 \text{Now, } r_{XY} &= \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \\
 &= \frac{8 \times 473 - 59 \times 71}{\sqrt{8 \times 583 - (59)^2} \sqrt{8 \times 687 - (71)^2}} \\
 &= -0.55
 \end{aligned}$$

Step deviation method

If $\bar{X}$ and $\bar{Y}$ are fractional and X and Y assume large values, the above formula are not generally used for numerical problems. In such a case, the step deviation method where we take the deviations of the variables X and Y from any arbitrary numbers, is used. For this we use the following formula,

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

The values of u and v are the step deviations of the values of X and Y respectively. To calculate r_{XY} , the following steps should be adopted

- (a) Write the values of X and Y. Find deviation $X - A$ and $Y - B$ from suitable constants A and B respectively. If there is some common factor $h(>1)$ in the value of $X - A$ then find $\frac{X-A}{h}$ in

the next column and put $\frac{X - A}{h} = u$. If there is no common factor other than 1 in the value of $X - A$, then put this itself as u . In the similar way, find the value of the variable v as $v = \frac{Y - B}{k}$.

(b) Make the columns with headings uv , u^2 and v^2 .

(c) Find the sum of values of u , v , uv , u^2 and v^2 .

(d) Calculate $r = \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2} \sqrt{n\sum v^2 - (\sum v)^2}}$

Example 8.

Calculate the coefficient of correlation from the following data:

X	14	16	19	22	24	30
Y	24	22	20	24	23	26

Solution:

COMPUTATIONS OF CORRELATION COEFFICIENT

X	Y	$u = X - 19$	$v = Y - 23$	u^2	v^2	uv
14	24	-5	1	25	1	-5
16	22	-3	-1	9	1	3
19	20	0	-3	0	9	0
22	24	3	1	9	1	3
24	23	5	0	25	0	0
30	26	11	3	121	9	33
		$\sum u = 11$	$\sum v = 1$	$\sum u^2 = 189$	$\sum v^2 = 21$	$\sum uv = 34$

We know that,

$$\begin{aligned}
 r &= \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2} \sqrt{n\sum v^2 - (\sum v)^2}} = \frac{6 \times 34 - 11 \times 1}{\sqrt{6 \times 189 - (11)^2} \sqrt{6 \times 21 - (1)^2}} \\
 &= \frac{204 - 11}{\sqrt{1134 - 121} \sqrt{126 - 1}} = \frac{193}{\sqrt{1013} \sqrt{125}} = \frac{193}{\sqrt{126625}} = \frac{193}{355.84} \\
 \therefore r &= 0.542
 \end{aligned}$$

8.6 Correlation Coefficient of Bi-variate Distribution

If values of two variables (or series) X and Y are grouped with their corresponding given frequencies of different groups which are presented combinably in a table is known as "Bivariate frequency table" or "Correlation table". This table is constructed if the number of observations is too large and the data have been classified according as two measurements. The correlation coefficient of bi-variate distribution is obtained by using the formula as:

$$r = \frac{N\sum f_{uv} - (\sum f_u)(\sum f_v)}{\sqrt{N\sum f_u^2 - (\sum f_u)^2} \sqrt{N\sum f_v^2 - (\sum f_v)^2}} \text{ where,}$$

$$N = \sum f, u = \frac{x-a}{h} \text{ and } v = \frac{y-b}{k}.$$

Steps to obtain the correlation coefficient of bi-variate distribution

1. Obtain middle values of each class group of X and Y and denote it respectively by x and y.
2. Obtain $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$ where a, b are two assumed means of mid-values x, y and h, k be the class intervals of two series X, Y respectively.
3. Find total frequency of all classes of y goes horizontally, kept in lower left corner of each small rectangular cell for each interval of X and again similarly find total frequency of all class intervals of X goes vertically, kept in upper right corner of each small rectangular cell.
4. Find f_v, f_v^2 and kept horizontally below each class interval of Y and find f_u, f_u^2 and kept vertically, right to each class interval of X again find their corresponding sums also.
5. Find f_{uv} for each class interval of Y taking vertical sum of the values kept at the upper right corner of each cell and again find f_{uv} for each class interval of X taking horizontal sum of the values kept at the upper right corner of each cell also find their corresponding sums horizontally and vertically respectively.
6. Obtain the correlation coefficient (r) of bi-variate distribution by using the above formula.

Example 9.

From the data of marks obtained by students in Math and Physics, find the Karl Pearson correlation coefficient and interpret the result.

Mark in Math	Marks in Physics				
	20-30	30-40	40-50	50-60	60-70
25-35	5	9	3	—	—
35-45	2	10	25	2	—
45-55	1	2	12	2	3
55-65	—	—	4	16	5
65-75	—	—	—	4	2

[CTEVT 2073 R/B]

Solution:**Computation of Correlation Coefficient**

x↓ y⇒		Marks in Physics	20-30	30-40	40-50	50-60	60-70				
		Mid-value of Y	25	35	45	55	65				
Marks in Maths	Mid-value of X	u	v					f	fu	fu ²	fuv
25-35	30	-2	20	18	0	—	—	17	-34	68	38
35-45	40	-1	4	10	0	-2	—	39	-39	39	12
45-55	50	0	0	0	0	0	0	20	0	0	0
55-65	60	1	—	—	4	16	10	25	25	25	26
65-75	70	2	—	—	—	8	8	6	12	24	16
f			8	21	44	24	10	107	-36	156	92
fv			-16	-21	0	24	20	7			
fv ²			32	21	0	24	40	117			
fuv			24	28	0	22	18	92			

$$r = \frac{N \sum fuv - (\sum fu)(\sum fv)}{\sqrt{N \sum fu^2 - (\sum fu)^2} \sqrt{N \sum fv^2 - (\sum fv)^2}}$$

$$\begin{aligned}
 &= \frac{107 \times 92 - (-36)(7)}{\sqrt{107 \times 156 - (-36)^2} \sqrt{107 \times 117 - (7)^2}} \\
 &= \frac{9844 + 252}{\sqrt{16692 - 1296} \sqrt{12519 - 49}} \\
 &= \frac{10096}{\sqrt{15396} \sqrt{12470}} \\
 &= \frac{10672}{124.081 \times 111.67} = \frac{10672}{13856.020} \\
 &= 0.770
 \end{aligned}$$

∴ The interpretation of result is that there is high degree positive correlation between the marks of Math and Physics.

8.7 Properties of Correlation Coefficient

Some important properties of the correlation coefficients and their proof are as follows:

1. The correlation coefficient between two variables lies in between -1 and $+1$. i.e. $-1 \leq r \leq 1$

Proof:

Let us consider the sum of squares

$$\sum \left[\frac{x - \bar{x}}{\sigma_x} \pm \frac{y - \bar{y}}{\sigma_y} \right]^2, \text{ which is always non-negative.}$$

$$\text{i.e. } \sum \left[\frac{x - \bar{x}}{\sigma_x} \pm \frac{y - \bar{y}}{\sigma_y} \right]^2 \geq 0$$

$$\Rightarrow \sum \left[\left(\frac{x - \bar{x}}{\sigma_x} \right)^2 + \left(\frac{y - \bar{y}}{\sigma_y} \right)^2 \pm 2 \left(\frac{x - \bar{x}}{\sigma_x} \right) \left(\frac{y - \bar{y}}{\sigma_y} \right) \right] \geq 0$$

$$\Rightarrow \frac{\sum (x - \bar{x})^2}{\sigma_x^2} + \frac{\sum (y - \bar{y})^2}{\sigma_y^2} \pm \frac{2 \sum (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} \geq 0$$

Dividing throughout by n , the number of pairs of observations, we get

$$\begin{aligned}
& \frac{1}{\sigma_x^2} \cdot \frac{1}{n} \Sigma (x - \bar{x})^2 + \frac{1}{\sigma_y^2} \cdot \frac{1}{n} \Sigma (y - \bar{y})^2 \pm \\
& \quad \frac{2}{\sigma_x \sigma_y} \cdot \frac{1}{n} \Sigma (x - \bar{x})(y - \bar{y}) \geq 0 \\
\Rightarrow & \frac{1}{\sigma_x^2} \cdot \sigma_x^2 + \frac{1}{\sigma_y^2} \cdot \sigma_y^2 \pm \frac{2}{\sigma_x \sigma_y} \text{Cov}(x, y) \geq 0 \\
\text{But } r = & \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}
\end{aligned}$$

$$\Rightarrow 1 + 1 \pm 2r \geq 0$$

$$\Rightarrow 2 \pm 2r \geq 0$$

$$\Rightarrow 1 \pm r \geq 0$$

Taking +ve sign

$$1 + r \geq 0$$

$$\Rightarrow -1 \leq r \quad (1)$$

Taking -ve sign

$$1 - r \geq 0$$

$$\Rightarrow r \leq 1 \quad (2)$$

Combining (1) and (2), we get

$$-1 \leq r \leq 1$$

2. The correlation coefficient between two variables is independent of change of origin and scale.

Proof:

Let X and Y be two given variables and r_{XY} be the correlation coefficient between them.

$$\text{Then, } r_{XY} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2} \sqrt{\Sigma (Y - \bar{Y})^2}} \quad (1)$$

Let the original variables X and Y be transformed to the new variables U and V by the change of origin and scale.

$$\text{i.e., } U = \frac{X - A}{h} \text{ and } V = \frac{Y - B}{k} \quad (2)$$

where A , B , h and k are constant and $h > 0$, $k > 0$. We have to show that the correlation coefficient between X and Y is same as the correlation coefficient between U and V .

$$\text{i.e., } r_{XY} = r_{UV}$$

From (2),

$$X = A + hU \text{ and } Y = B + kV \quad (3)$$

Summing both sides and dividing by n , we get

$$\bar{X} = A + h\bar{U} \quad \text{and} \quad \bar{Y} = B + k\bar{V} \quad (4)$$

Subtracting (4) from (3), we get

$$X - \bar{X} = h(U - \bar{U}) \text{ and } Y - \bar{Y} = k(V - \bar{V})$$

Substituting these values in (1), we get

$$\begin{aligned} r_{XY} &= \frac{\Sigma h(U - \bar{U}) \times k(V - \bar{V})}{\sqrt{\Sigma h^2 (U - \bar{U})^2} \sqrt{\Sigma k^2 (V - \bar{V})^2}} \\ &= \frac{hk \Sigma (U - \bar{U})(V - \bar{V})}{hk \sqrt{\Sigma (U - \bar{U})^2} \sqrt{\Sigma (V - \bar{V})^2}} \\ &= \frac{\Sigma (U - \bar{U})(V - \bar{V})}{\sqrt{\Sigma (U - \bar{U})^2} \sqrt{\Sigma (V - \bar{V})^2}} \\ &= r_{UV}. \text{ Hence proved.} \end{aligned}$$

3. The correlation coefficient between the two variables X and Y is symmetrical. i.e. $r_{XY} = r_{YX}$.

8.8 Interpretation of Correlation Coefficient

The value of Karl person's coefficient of correlation can be interpreted as follows:

1. If $r = +1$ then there is perfect positive linear correlation.
2. If $0.75 \leq r < 1$ then there is high degree positive linear correlation.
3. If $0.50 \leq r < 0.75$ then there is the moderate degree positive linear correlation.
4. If $0 < r < 0.50$ then there is low degree positive linear correlation.
5. If $r = 0$ then there is no linear correlation.

For the negative value of r can be interpreted as negative correlation as mentioned above.

Merits and demerits

The following are the merits and demerits of the Karl Pearson's correlation coefficient.

Merits:

- (a) The coefficient of correlation indicate magnitude as well as direction of linear correlation between the variables.
- (b) This is a mathematical method and there is no chance of personal bias in its computation.
- (c) This method is considered as the best method of computing simple linear correlation.

Demerits:

- (a) This is difficult to compute.
- (b) This is applicable only when the correlation between the variables is linear.
- (c) This is affected by the extreme items.

EXERCISE - 8(I)

1. What do you mean by correlation? Explain the different types of correlation.
2. What is scatter diagram? How does it help to give the idea about the correlation between two variables?
3. Find $\text{cov}(X, Y)$ between X and Y from the following data:
 - (a) $\Sigma X = 20$, $\Sigma Y = 45$, $\Sigma XY = 148$ and $n = 5$
 - (b) $\Sigma X = 55$, $\Sigma Y = 74$, $\Sigma XY = 411$ and $n = 10$
4. Find the coefficient of correlation between X and Y when $\text{cov}(X, Y) = -16.5$, $\text{var}(X) = 2.89$ and $\text{var}(Y) = 100$.
5. Compute the correlation coefficient when $\text{cov}(X, Y) = 18.5$, $\text{var}(X) = 15$ and $\text{var}(Y) = 48$.
6. If the covariance between the variables X and Y is 20 and the variances of X and Y are 25 and 49 respectively, find the coefficient of correlation between them.
7. If the covariance between the two variables X and Y is 10 and standard deviations are 4 and 3 respectively, find the coefficient of correlation.
8. If the covariance between the two variables X and Y is 15 and standard deviations are 5 and 7 respectively. Find the coefficient of correlation.
9. Calculate the Karl Pearson's coefficient of correlation from the data given below:

X	2	3	5	7	3
Y	15	17	4	5	4

10. Find the coefficient of correlation for the following data:

X	6	2	10	4	8
Y	9	11	5	8	7

11. Find the coefficient of correlation for the following data:

X	17	10	23	18	22
Y	6	8	10	6	5

12. From the following data, calculate the coefficient of correlation:

X	6	2	10	4	8
Y	9	11	?	8	7

It being given that the arithmetic mean of Y is 8.

13. Find the Karl Pearson's coefficient of correlation between X and Y for the following data:

X	10	7	12	15	9	15	8
Y	6	4	7	10	11	8	10

14. Calculate the coefficient of correlation between the marks in Botany and Zoology obtained by 7 students.

Marks in Botany	60	57	72	85	79	75	88
Marks in Zoology	66	64	67	70	71	78	80

15. Calculate Karl Pearson's coefficient of correlation between X and Y for the following data:

X	78	89	96	69	59	79	68	62
Y	125	137	156	112	107	136	123	104

16. Calculate the Karl Pearson's coefficient of correlation between the age and blood pressure of 10 patients.

Age	23	48	43	68	70	28	35	26	56	51
Blood pressure	115	127	123	140	145	118	121	120	133	130

Assuming that the average age and blood pressure of 10 patients are 45 and 127 respectively.

17. Calculate the Karl Pearson's coefficient of correlation between height and weight from the following data:

Height(cms)	161	160	163	167	164	166
Weight(kg)	61	63	64	70	66	68

18. Compute the correlation coefficient between blood pressure and age of patients.

Blood Pressure (mmHg)	120	150	145	100	120	120	155	130	140	120
Age of Patients	20	35	32	40	43	45	30	25	50	40

19. Ten students got the following percentage of marks in Mathematics and Physics:

Mathematics	8	36	93	35	62	65	25	75	92	82
Physics	84	51	91	49	58	35	60	68	36	62

Calculate the correlation coefficient.

20. The following table gives bivariate frequency distribution of 50 clerks according to the age in years and pay in rupees.

Age	Pay			
	250-300	300-350	350-400	400-450
20-30	8	3	—	—
30-40	2	5	2	2
40-50	—	2	9	6
50-60	—	—	5	6

Calculate the coefficient of correlation. Also interpret on the result.
[CTEVT 2075 R/B]

21. Calculate the coefficient of correlation from the following data. Also interpret on the result.

Ages of wives	Ages of husband				
	20-30	30-40	40-50	50-60	60-70
15-25	5	9	3	—	—
25-35	—	10	25	2	—
35-45	—	1	12	2	—
45-55	—	—	4	16	5
55-65	—	—	—	4	2

[CTEVT 2072 R/B]

ANSWERS

- | | | |
|---|-----------|-----------|
| 3. (a) -6.4 | (b) 0.4 | 4. -0.97 |
| 5. 0.69 | 6. 0.57 | 7. 0.83 |
| 8. 0.43 | 9. -0.62 | 10. -0.92 |
| 11. 0 | 12. -0.92 | 13. 0.26 |
| 14. 0.74 | 15. 0.81 | 16. 0.97 |
| 17. 0.94 | 18. -0.17 | 19. -0.09 |
| 20. 0.763, high degree positive correlation | | |
| 21. 0.761, high degree positive correlation | | |

8.9 Line Regression

The regression analysis is used to determine the strength (i.e., degree) of the relationship (or correlation) between two variables. The word regression is stepping (or returning) back to the original position. It is a statistical device. It helps to estimate (or predict) the value of one variable if the value of other is known. The unknown variable is called dependent and that of the variable whose value is known is the independent variable. The average relationship between two variables is known as simple linear regression analysis.

8.10 Regression Equations and Regression Coefficients

The regression lines, given by two variables (independent and dependent), which are expressed in terms of algebraic relations are known as regression equations. There are two types of regression lines (equations): (a) The regression equation of y on x and (b) The regression equation of x on y.

(a) **Regression equation of y on x:** The regression equation of y on x is

$$y = a + bx \quad (1)$$

where a and b are constants to be determined. The values of a and b can be obtained by solving the equations

$$\Sigma y = na + b\Sigma x \quad (2)$$

$$\text{and } \Sigma xy = a\Sigma x + b\Sigma x^2 \quad (3)$$

For the values of a and b the equation (1) gives the required line of regression of y on x where b is the regression coefficient of y on x denoted by b_{yx} . It gives the best estimate, of y for the given value of x.

(b) **Regression equation of x on y:** The regression equation of x on y is

$$x = a + by \quad (2)$$

where a and b are constants to be determined. The values of a and b can be obtained by solving the equations

$$\Sigma x = na + b\Sigma y \quad (4)$$

$$\text{and } \Sigma xy = a\Sigma y + b\Sigma y^2 \quad (5)$$

For the values of a and b the equations (2) gives the required line of regression of x on y where b is the regression coefficient of x on y . It is denoted by b_{xy} . It gives the best estimate of x for the given value of y .

Notes:

- (i) If two lines of regression coincide with each other then there is perfect correlation between them.
- (ii) If two lines of regression intersect at right angle then there is no correlation between them.
- (iii) Nearer the two lines of regression implies the high degree of correlation and farther (or more wider between) two lines of regression implies the low degree of correlation.

Example 1.

The following table gives the normal weight of a baby during the first 6 months of his life. Find the regression equation of y on x . Also estimate the weight of the baby at the age of 4 months.

Age (in month)	0	2	3	5	6
Weight (in lbs)	5	7	8	10	12

Solution:

Computation of regression equation

Age (x)	Weight (y)	xy	x^2	y^2
0	5	0	0	25
2	7	14	4	49
3	8	24	9	64
5	10	50	25	100
6	12	72	36	144
$\Sigma x = 16$	$\Sigma y = 42$	$\Sigma xy = 160$	$\Sigma x^2 = 74$	$\Sigma y^2 = 382$

Here, $\Sigma x = 16$, $\Sigma y = 42$, $\Sigma xy = 160$, $\Sigma x^2 = 74$, $\Sigma y^2 = 382$
 $n(\text{no. of items}) = 5$

The regression equation of y on x is

$$y = a + bx \quad (1)$$

The values of a and b can be obtained on solving,

$$\Sigma y = na + b\Sigma x$$

$$\text{i.e. } 42 = 5a + 16b \quad (2)$$

$$\text{and } \Sigma xy = a\Sigma x + b\Sigma x^2$$

$$\text{i.e. } 160 = 16a + 74b$$

$$\Rightarrow 80 = 8a + 37b \quad (3)$$

We get, $a = 4.81, b = 1.12$

Putting the values of a and b in equation (1) we get, $y = 4.81 + 1.12x$ which is the required regression equation of y on x.

Again, the regression equation of x on y is given by,

$$x = a + by \quad (4)$$

The values of a and b can be obtained on solving the equations given by,

$$\Sigma x = na + b\Sigma y$$

$$16 = 5a + 42b \quad (5)$$

$$\text{and } \Sigma xy = a\Sigma y + b\Sigma y^2$$

$$160 = 42a + 382b$$

$$\text{i.e. } 80 = 21a + 191b \quad (6)$$

Solving (5) and (6) we get,

$$a = -0.17 \text{ and } b = 0.035$$

Putting the values of a and b in equation (4) we get,

$$x = -0.17 + 0.035y$$

which is the required regression equation of x on y. Again we have, the regression equation of y on x is

$$y = 4.81 + 1.12x$$

Now, to estimate the required weight (y) for $x = 4$ months, it gives,

$$y = 4.81 + 1.12(4)$$

$$= 4.81 + 4.48 = 9.29$$

$\therefore$ The weight of the baby in 4 month will be 9.29 lbs

Further, the lines of regression can also be obtained on taking the deviation of the items from the assumed mean. In such case the two regression equations can be obtained as below.

The lines of regression of Y on X and x on Y are respectively given by

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \text{ and } X - \bar{X} = b_{xy}(Y - \bar{Y})$$

where $\bar{x}$ and $\bar{y}$ are the arithmetic means of x and y series respectively. Here the regression coefficient of y on x is given by,

$$b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = r \cdot \frac{\sigma_y}{\sigma_x}$$

and the regression coefficient of x on y is given by

$$b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(y - \bar{y})^2} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = r \cdot \frac{\sigma_x}{\sigma_y}$$

where σ_x and σ_y are the standard deviation of the series x and y respectively. and r be the correlation coefficient between the series x and y . Thus for the assumed mean we have,

$$b_{yx} = \frac{n\Sigma uv - \Sigma u \cdot \Sigma v}{n\Sigma u^2 - (\Sigma u)^2} \text{ and } b_{xy} = \frac{n\Sigma uv - \Sigma u \cdot \Sigma v}{n\Sigma v^2 - (\Sigma v)^2}$$

where $u = x - a$ and $v = y - b$.

Example 2.

Find two regression equations from the distributions below:

x	10	12	13	12	16	15
y	40	38	43	45	37	43

Solution:

Computation of regression equations

x	y	$u = x - 13$	$v = y - 40$	u^2	v^2	uv
10	40	-3	0	9	0	0
12	38	-1	-2	1	4	2
13	43	0	3	0	9	0
12	45	-1	5	1	25	-5
16	37	3	-3	9	9	-9
15	43	2	3	4	9	6
		$\Sigma u = 0$	$\Sigma v = 6$	$\Sigma u^2 = 24$	$\Sigma v^2 = 56$	$\Sigma uv = -6$

$$\text{Here, } \bar{x} = A + \frac{\sum u}{n} = 13 + \frac{0}{6} = 13 + 0 = 13$$

$$\bar{y} = B + \frac{\sum v}{n} = 40 + \frac{6}{6} = 40 + 1 = 41$$

$$\text{Again, } b_{yx} = \frac{n\sum uv - \sum u \cdot \sum v}{n\sum u^2 - (\sum u)^2} = \frac{6(-6) - 0(6)}{6(24) - 0} = -0.25$$

$$\text{And, } b_{xy} = \frac{n\sum uv - \sum u \cdot \sum v}{n\sum v^2 - (\sum v)^2} = \frac{6(-6) - 0(6)}{6(56) - (6)^2} = -0.12$$

Now, the regression equation of y on x is given by,

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or, } y - 41 = (-0.25)(x - 13)$$

$$\text{or, } y - 41 = -0.25x + 3.25$$

$$\begin{aligned} \text{or, } y &= -0.25x + 3.25 + 41 \\ &= 0.25x + 44.25 \end{aligned}$$

$$\therefore y = -0.25x + 44.25$$

which is the required regression equation of y on x.

Again, the regression equation of x on y is given by,

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$\text{i.e. } x - 13 = -0.12(y - 41)$$

$$\text{or, } x - 13 = -0.12y + 0.12(41)$$

$$\text{or, } x - 13 = -0.12y + 4.92$$

$$\therefore x = -0.12y + 17.92$$

is the required regression equation of x on y.

Notes:

- (i) The two regression coefficient must have same sign.
- (ii) The sign of regression coefficients is same as the sign of correlation coefficient.
- (iii) $b_{yx} \cdot b_{xy} \leq 1$.
- (iv) $r = \sqrt{b_{yx} \cdot b_{xy}}$ where r be correlation coefficient.

(v) Regression coefficients are independent of the change of origin but not of scale i.e. $b_{yx} = b_{vu}$ for $u = x - A$, $v = y - B$ and $b_{yx} \neq b_{vu}$ for $u = \frac{x}{h}$,

$$v = \frac{y}{k}.$$

(vi) If $u = \frac{x-A}{h}$ and $v = \frac{y-B}{k}$ then $b_{yx} = \frac{n\sum uv - \sum u \cdot \sum v}{n\sum u^2 - (\sum u)^2} \times \frac{k}{h}$ and

$$b_{xy} = \frac{n\sum uv - \sum u \cdot \sum v}{n\sum v^2 - (\sum v)^2} \times \frac{h}{k}.$$

Example 3.

From the given informations below, estimate the most likely production corresponding to the rainfall of 50 mm.

	Rainfall (mm)	Production (m. ton)
Average	30	500
Standard deviation	5	100
Correlation coefficient	0.8	

Solution:

Suppose that rainfall and production quantities are denoted by x and y respectively.

$$\therefore \text{Here, } \bar{x} = 30, \bar{y} = 500$$

$$\sigma_x = 5, \sigma_y = 100 \text{ and } r = 0.8$$

To predict the production we need to find the regression equation of y on x (as the value of x is given) which is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad (1)$$

$$\text{where } b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x} = 0.8 \times \frac{100}{5}$$

$$\text{or, } b_{yx} = 0.8 \times 20 = 16 \text{ i.e. } b_{yx} = 16$$

The equation (1) becomes as

$$y - 500 = 16(x - 30)$$

$$\text{or, } y - 500 = 16x - 480$$

$$\text{or, } y = 16x + 500 - 480$$

$$\therefore y = 16x + 20 \text{ which is the regression equation of } y \text{ on } x.$$

$$\text{When } x = 40, y = 16(40) + 20$$

$$\text{or, } y = 640 + 20 = 660$$

That is, the volume of production corresponding to 40 mm rainfall is 660 m. tons.

8.11 Difference between Correlation and Regression

The differences between correlation and regression are as below.

Correlation	Regression
1. Since correlation is the relationship between two variables, so the change in the value of one variable affects in the value of next variable.	1. Regression means returning back to the average value.
2. There is no need to imply cause and effect between the relationship of two variables in correlation.	2. In regression analysis, it clearly indicates the cause and effect of relationship between two variables.
3. It is independent of change of origin and scale both.	3. Regression coefficients are independent of change of origin but not of scale.
4. Its value lies between ± 1 .	4. The product of two regression coefficients is less or equal to 1 (i.e. $b_{yx} \times b_{xy} \leq 1$).
5. The correlation coefficient between two variables is the geometric mean between two regression coefficient. i.e. $r = \sqrt{b_{yx} \cdot b_{xy}}$	5. The arithmetic average of two regression coefficients is greater than or equal to the correlation coefficient i.e. $\frac{b_{yx} + b_{xy}}{2} \geq r$.
6. It presents the degree of relationship between two variables and also possesses the direction of their movements.	6. It studies the nature of relationship between two variables and it makes possible to predict the value of one variable if the value of other variable is known.
7. The sign of correlation coefficients depends according as both regression coefficients are positive or both negative.	7. Two regression coefficients have always same signs.

EXERCISE - 8(II)

1. Find the lines of regression from the following data:

(a)

x	10	12	13	17	18
y	5	6	7	9	13

(b)

x	10	12	14	16	18
y	5	6	7	10	10

2. Find the regression equations of y on x and x on y from the data given below and estimate the probable value of x if y = 17.

x	5	9	13	17	21
y	3	8	13	18	23

3. Find the two line of regressions and estimate the income of the worker who has served for 12 years from the following data:

Service year (x)	2	5	8	6	10	11	7
Income (Rs. y)	6	5	9	7	11	10	8

4. From the given informations below find the two lines of regressions and estimate probable sales corresponding to the advertisement expenditure of Rs. 15 lakhs. Also estimate the advertisement expenditure corresponding to the sales target of Rs. 120 lakhs.

	Advertising expenditure (Rs. in lakhs) (x)	Sales (Rs. in lakhs) (y)
Mean	10	90
S.d.	3	12

Correlation coefficient = 0.8

5. For ten observations on price (x) and supply (y) the following results were obtained $\Sigma x = 130$, $\Sigma y = 220$, $\Sigma xy = 3467$, $\Sigma x^2 = 2288$ and $\Sigma y^2 = 5506$. Obtain the line of regression of y on x and estimate the supply if the price is 16 units. Also find the correlation coefficient between price (x) and supply (y).
6. The ages of ten couples of husbands and wives are listed below. Find
- the lines of two regression equation.
 - the most likely age of husband for the age of 25 years of wife.
 - the coefficient of correlation between the ages of husbands and wives.

Age of husbands (x)	22	23	23	24	26	27	27	28	30	30
Age of wives (y)	18	20	21	20	21	22	23	24	25	26

ANSWERS

1. (a) $y = 0.87x - 4.17$, $x = y + 6$
(b) $y = 0.7x - 2.2$, $x = 1.32y + 3.96$
2. $y = 1.25x - 3.25$, $x = 0.8y + 2.6$, 16.2
3. $y = 0.63x + 3.63$, $x = 1.25y - 3$, 11.13
4. $y = 3.2x + 58$, $x = 0.2y - 8$, Rs. 106 lakhs and Rs. 16 lakhs
5. $y = 1.01x + 8.81$, 25.045, 0.962
6. $y = 0.82x + 0.78$, $x = 1.11y + 1.65$ (b) 29.32 yrs (c) 0.95

●●●