COURSE PACK

FOR

Statistics

(For internal use only)

Course Code: 304

Course: BCA

Sem: III

YEAR: 2019-2021

Course Leader: Dr. Indu Rani

Subject Faculty: Dr. INDU

Ms. Bhavika

Forwarded by:

Dr. Daljeet Singh Bawa
(Programme Cordinator-BCA Morning &Afternoon)

Dr. A.K.Srivastav
HOD

Approved By:

Dr. Vikas Nath
Director In Charge

Bharati Vidyapeeth(Deemed to be)University Institute of Management &
Research, New Delhi
An ISO 9001:2015 Certified Institute
"A+" Accreditation by NAAC

Bharati Vidyapeeth (Deemed to be)University Institute of Management and Research,

1.Course Instructor; Indu Rani

2.Course Credit:

3.Number of Sessions: 40

4.Course Overview;-Statistics is the science that deals with the collection, description, analysis, interpretation, and presentation of data. Statistics can be used to describe a particular data set (termed descriptive statistics) as well as to draw conclusions about the population from a particular data set (termed inferential statistics). *Statistics in Business* applies statistical methods in a business context in order to address business related questions and help make evidence based decisions. In *Statistics in Business* you will learn to apply commonly used statistical methods in business contexts and how to interpret analyses performed by others. Business statistics is the science of good decision making in the face of uncertainty and is used in many disciplines such as financial analysis, econometrics, auditing, production and operations including services improvement, and marketing research.

Statistical ideas and methods are essential tools in virtually all areas that rely on data to make decisions and reach conclusions. This includes diverse fields such as medicine, science, technology, government, commerce and manufacturing. In broad terms, statistics is about getting information from data. This includes both the important question of how to obtain suitable data for a given purpose and also how best to extract the information, often in the presence of random variability. This course provides an introduction to the contemporary application of statistics to a wide range of real world situations. It has a strong practical focus using the statistical package SPSS to analyse real data. Topics covered are: organisation, description and presentation of data; design of experiments and surveys; random variables, probability distributions, the binomial distribution and the normal distribution; statistical inference, tests of significance, confidence intervals; inference for means and proportions, one-sample tests, two independent samples, paired data, t-tests, contingency tables; analysis of variance; linear regression, least squares estimation, residuals and transformations, inference for regression coefficients, prediction.

5. Learning Outcomes -Understand graph frequency distribution with histogram, polygon, and ogives.

Learn mean, median, mode and also to describe how data bunch up".

Understand range, variance and standard deviation to describe how data spread out.

Provide a description of the method used for analysis, including a discussion of advantages, disadvantages, and necessary assumptions.

Provide a discussion of the results and of the statistical analysis.

Provide a conclusion to the study including a discussion of limitations of the analysis.

intervals, hypothesis for means and proportions, one-sample tests, two independent samples, paired data, t-tests, contingency tables; analysis of variance; linear regression, least squares estimation, residuals and transformations, inference for regression coefficients, prediction.

5. Learning Outcomes -Understand graph frequency distribution with histogram, polygon, and ogives.

Learn mean, median, mode and also to describe how data bunch up".

Understand range, variance and standard deviation to describe how data spread out.

Provide a description of the method used for analysis, including a discussion of advantages, disadvantages, and necessary assumptions.

Provide a discussion of the results and of the statistical analysis.

Provide a conclusion to the study including a discussion of limitations of the analysis.

5 Pedagogy

The course shall be covered in 40 sessions. The course shall consist of class discussions, presentations, experience sharing, investigative research and case studies.

Table of Contents

6. Evaluation Criteria

Component	Description	Weight	Objective
First Internal	First internal would be of Marks 40 for first three units 1, 2 & 3.	Marks 10	To evaluate student's cognitive skills (Think, read, learn, remember, reason, and pay attention) for first half of the course.
Second Internal	Second internal will be of Marks 40 for rest of the units.	Marks 10	To evaluate student's cognitive skills (Think, read, learn, remember, reason, and pay attention) for rest half of the course.
CES activities 1- Class Test	There would be a class test for which individual assessment will be done.	Marks 5	To recall their subject learning.
2-Class Test		Marks 5	
3- Class Test		Marks 5	
Class participation & 75% attendance	Class notes and involvement of students will be checked by faculty during semester.	Marks 10	To encourage and enhance class participation

Note: All CES activities are mandatory. If any students misses any one CES in that case the weightage of each CES would be 3.33 marks and if a student attempts all 3 CES then his/her best 2 CES will be considered in that case weightage would be 5 marks each.

7. RECOMMENDED/REFERENCE TEXT BOOKS

Text Book	Business Statistics, S P Gupta and M P Gupta by Sultan Chand and Sons
Course reading	1. Fundamentals of Statistics, SC Gupta, Himalaya Publishing House. 2. Business Statistics, ND Vohra, Mac Graw Hill publications. 3. Statistics for Management, Levin Rubin, Pearson.

Statistics

Content of course

Unit I

Introduction to Statistical Methods:

- Definition of statistics, Importance of Statistics, scope of statistics: Economics, Computer Science, Business and Management, Limitations of statistics.

Unit II

Collection and Organisation of Data

- Sources of data: Primary Data and Secondary Data, Tabular Representation of data: Ungrouped and Grouped Frequency Distribution, Graphical representation of data: Histogram, Frequency Polygon, ogives, Diagrammatic Representation: Simple bar, Subdivided bar, pie diagram.

Unit III

Measure of Central Tendency

- Mean- Definition, Formulae, and Computing for grouped and ungrouped data, merits and demerits.
  Median- Definition, Formulae, and Computing for grouped and ungrouped data, merits and demerits.
  Mode- Definition, Formulae, and Computing for grouped and ungrouped data, merits and demerits.

Unit IV

Measure of Dispersion

- Range: Definition, Formulae, and Computing for grouped and ungrouped data, merits and demerits.
- Mean Deviation: Definition, Formulae, and Computing for grouped and ungrouped data, merits and demerits.
- Standard Deviation: Definition, Formulae, and Computing for grouped and ungrouped data, merits and demerits.
- Deciles, percentiles and Quartiles.

Unit V

Regression and Correlation

- **Regression** : Definitions, Regression equations, Regression coefficients, Problems on finding Regression equations and estimations.
- **Correlation**: Definitions, Karl Pearson's correlation coefficient, Spearman's rank correlation with correction factor.

Unit VI

Analysis of Time Series

- Component of time series analysis, fitting a straight line $y=ax+b$, fitting a curve $y= ax^2+bx+c$, 3 yearly and 5 yearly Moving Averages.

Session Plan

Unit	Lecture No.	Topic	Book	Learning Outcome
I	1.	Introduction to Statistics	Statistical Method By (Dr.S.P.Gupta Chapter 1)	LO1
	2.	Importance of Statistics	Statistical Method By (Dr.S.P.Gupta Chapter 1)	Able to know what the subject of statistics involves.
	3.	Importance of Statistics	Statistical Method By (Dr.S.P.Gupta Chapter 1)	LO1
	4.	Scope of Statistics	Statistical Method By (Dr.S.P.Gupta Chapter 1)	Understand the difference between descriptive and inferential statistics.
	5.	Limitations of Statistics	Statistical Method By (Dr.S.P.Gupta Chapter 1)	LO1
	6.	Sources of data: Primary Data	Statistical Method By (Dr.S.P.Gupta Chapter 2 Pg. 30)	LO1
	7.	Sources of data: Secondary Data	Statistical Method By (Dr.S.P.Gupta Chapter 2 Pg. 30)	Know the methods commonly used for collecting data.
	8.	Classification of data – various methods to classify data – discrete and continuous variables	Statistical Method By (Dr.S.P.Gupta Chapter 5)	LO1
	9.	Construction of Frequency Distribution and relative frequency	Statistical Method By (Dr.S.P.Gupta Chapter 5)	LO1
	10.	Construction of cumulative Frequency Distribution	Statistical Method By (Dr.S.P.Gupta Chapter 5)	LO1
	11.	Graphical representation of Frequency distribution: Histogram, Frequency Polygon.	Statistical Method By (Dr.S.P.Gupta Chapter 6)	LO1
	12.	Graphical representation of Frequency distribution: Histogram, Frequency Polygon.	Statistical Method By (Dr.S.P.Gupta Chapter 6)	LO1
	13.	Graphical representation of Frequency distribution: Histogram, Frequency Polygon.	Statistical Method By (Dr.S.P.Gupta Chapter 6)	LO1
	14.	Diagrammatic Representation: Simple bar, Subdivided bar, pie diagram	Statistical Method By (Dr.S.P.Gupta Chapter 6)	LO1
	15.	Diagrammatic Representation: Simple bar, Subdivided bar, pie diagram	Statistical Method By (Dr.S.P.Gupta Chapter 6)	LO1

16.	Introduction to Measure of Central Tendency	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
17.	Arithmetic Mean- Individual series and discrete series	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
18.	Arithmetic Mean- Continuous and cumulative series	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
19.	Median: definition, merits and demerits, Individual series and discrete series	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
20.	Median: Continuous and cumulative series	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
21.	Mode: Individual series and discrete series	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
22.	Mode: Continuous and cumulative series	Statistical Method By (Dr.S.P.Gupta Chapter 7)	LO2
23.	Concept of Dispersion: Absolute and Relative Measure of dispersion	Statistical Method By (Dr.S.P.Gupta Chapter 8)	LO3
24.	Range	Statistical Method By (Dr.S.P.Gupta Chapter 8)	LO3
25.	Mean Deviation	Statistical Method By (Dr.S.P.Gupta Chapter 8)	LO3
26.	Standard Deviation	Statistical Method By (Dr.S.P.Gupta Chapter 8)	LO3
27.	Deciles, Percentiles and quartiles	Statistical Method By (Dr.S.P.Gupta Chapter 8)	LO3
28.	Linear Regression Model, Regression Lines, Y on X and X on Y	Statistical Method By (Dr.S.P.Gupta Chapter 11)	LO4
29.	Regression coefficients, Properties of Regression coefficients	Statistical Method By (Dr.S.P.Gupta Chapter 11)	LO4
30.	Estimation of unknown values to find regression coefficient and correlation coefficient from lines of regression	Statistical Method By (Dr.S.P.Gupta Chapter 11)	LO4
31.	Concept of Correlation and Types of Correlation	Statistical Method By (Dr.S.P.Gupta Chapter 10)	LO4
32.	Karl Pearson's Coefficient	Statistical Method By (Dr.S.P.Gupta Chapter 10)	LO5
33.	Rank Coefficient	Statistical Method By (Dr.S.P.Gupta Chapter 10)	LO5
34.	Properties of Correlation Coefficient	Statistical Method By (Dr.S.P.Gupta Chapter 10)	LO5
35.	Component of time series	Statistical Method By (Dr.S.P.Gupta Chapter 14)	LO6
36.	fitting a straight line $y=ax+b$, fitting a curve $y= ax^2+bx+c$,	Statistical Method By (Dr.S.P.Gupta Chapter 14)	LO6

37.	3 yearly and 5 yearly Moving Averages.	Statistical Method By (Dr.S.P.Gupta Chapter 14)	LO7
38.	3 yearly and 5 yearly Moving Averages.	Statistical Method By (Dr.S.P.Gupta Chapter 14)	LO7

UNIT 1

Introduction to Statistical Methods

Unit I

Statistics: Meaning, Characteristics and Importance

Meaning of Statistics:

The subject Statistics, as it seems, is not a new discipline but it is as old as the human society, itself. It has been used right from the existence of life on this earth, although the sphere of its utility was very much restricted.

In the olden days Statistics was regarded as the 'Science Statecraft' and was the byproduct of the administrative activity of the state. The word Statistics seems to have been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'statistik' or the French word 'statistique' each of which means a political state

Sixteen century saw the application of Statistics for the collection of the data relating to the movements of heavenly bodies—stars and planets—to know about their position and for the prediction of Eclipses. Seventeenth century witnessed the origin of Vital Statistics. Captain John Graunt of London (1620-1674), known as the Father of Vital Statistics, was the first man to make a systematic study of the birth and death statistics. Francis Galton (1822-1921) pioneered the study of 'Regression Analysis' in Biometry; Karl Pearson (1857-1936) who founded the greatest statistical laboratory in England pioneered the study of 'Correlation Analysis'.

His Chi-Square test (X -test) of Goodness of Fit is the first and most important of the tests of significance in Statistics; W.S. Gosset with his t-test ushered in an era of exact (small) sample tests. Perhaps most of the work in the statistical theory during the past few decades can be attributed to a single person Sir Ronald A.

Statistics as Statistical Methods (Singular Sense):

In this category of definitions Statistics is in singular sense. In singular sense statistics is used to describe the principles and methods which are employed in collection, presentation, analysis and interpretation of data. These devices help to simplify the complex data and make it possible for a common man to understand it without much difficulty.

Simple and comprehensive meaning of statistics, in singular sense, can be that a device which is employed for the purpose of collection, classification, presentation, comparison and interpretation of data. The purpose is to make the data simple, lucid and easy to be understood by a common man of mediocre intelligence.

Selligman maintained this view of the term 'statistics'. All this involves a procedure and a method from the primary stage to the final stage of analysis or conclusions etc. So this is quite comprehensive meaning and interpretation of the term statistics. Turtle also defines statistics as ***"the body of principles and techniques of collecting, classifying, presenting, comparing and interpreting quantitative data."***

3. "Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations."
—Bowley A.L.

4. "Statistics is the science of estimates and probabilities."
—Boddington

5. "The science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates."
—King

6. "Statistics is the science which deals with classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon."
—Lovitt

7. "Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry." —Selligman

8. "Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data."
—Croxton and Cowden

9. "Statistics may be regarded as a body of methods for making wise decision in the face of uncertainty."
—Wallis and Roberts

10. "Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks." —Prof. Ya-Lun-Chou

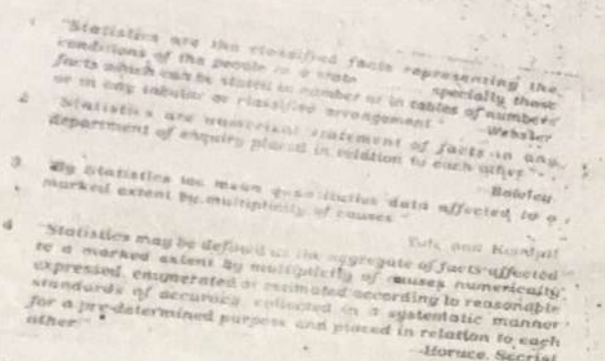
11. "The science and art of handling aggregate of facts—observing, enumeration, recording, classifying and otherwise systematically treating them." —Harlow

The first three definitions of Bowley are inadequate. Boddington's definition also fails to describe the meaning and functions of statistics since it is confined to only probabilities and estimates.

King's definition is also inadequate since it confines statistics only to social sciences. Lovitt's definition is fairly satisfactory, though incomplete. Selligman's definition, though very short and simple, is quite comprehensive. However, the best of all the above definitions seem to be given by Croxton and Cowden.

Statistics as Numerical Data (Plural Sense):

plural sense, statistics is considered as a numerical description of quantitative aspect of ngs. However, we give below some selected definitions of statistics as numerical data.



The definition of Statistics as given by Horace Secrist is most comprehensive and clearly points out certain essential characteristics which must be possessed by numerical data, in order to be called 'Statistics'

The characteristics are stated in following paragraphs:

1. Statistics are Aggregate of Facts:

Only those facts which are capable of being studied in relation to time, place or frequency can be called statistics. Individual, single or unconnected figures are not statistics because they cannot be studied in relation to each other. Due to this reason, only aggregate of facts e.g., data relating to I.Q. of a group of students, academic achievement of students, etc. are called statistics and are studied in relation to each other.

2. Statistics are Affected to a marked Extent by Multiplicity, of Causes:

Statistical data are more related to social sciences and as such, changes are affected to a combined effect of many factors. We cannot study the effect of a particular cause on a phenomenon. It is only in physical sciences that individual causes can be traced and their impact is clearly known. In statistical study of social sciences, we come to know the combined effect of multiple causes.

For example, deterioration of achievement score in academic sphere of some students may ot be only due to lack of interest in school subjects, but may also due to lack of motivation, fective teaching methods, attitude of the students on school subjects, faulty scoring cedure, etc.

larly scores on memory test of a group certainly depend on meaningfulness of ng materials, maturation of the students, methods of learning, motivation, interest

statistics e.g. honesty, goodness, ability, etc. But if we assign numerical expression, it maybe described as statistics.

4. Statistics are Enumerated or estimated according to Reasonable Standards of Accuracy:

The standard of estimation and of accuracy differs from enquiry to enquiry or from purpose to purpose. There cannot be one standard of uniformity for all types of enquiries and for all purposes. A single student cannot be ignored while calculating I Q. of 100 students in group whereas 10 soldiers can be easily ignored while finding out I.Q. of soldiers of whole country.

Similarly we can ignore ten deaths in a country but we cannot ignore even a single death in a family. The amount of time and resources at disposal also determine the amount of accuracy in estimates.

6. Statistics are Collected in a Systematic Manner:

In order to have reasonable standard of accuracy statistics must be collected in a very systematic manner. Any rough and haphazard method of collection will not be desirable for that may lead to improper and wrong conclusion. Accuracy will also be not definite and as such cannot be believed.

6. Statistics for a Pre-determined Purpose:

The investigator must have a purpose beforehand and then should start the work of collection. Data collected without any purpose is of no use. Suppose we want to know intelligence of a section of people, we must not collect data relating to income, attitude and interest. Without having a clear idea about the purpose we will not be in a position to distinguish between necessary data and unnecessary data or relevant data and irrelevant data.

7. Statistics are Capable of being Placed in Relation to each other:

Statistics is a method for the purpose of comparison etc. It must be capable of being compared. otherwise, it will lose much of its value and significance. Comparison can be made only if the data are homogeneous.

Data on memory test can be compared with I.Q. not with salary status of parents. It is with the use of comparison only that we can depict changes which may relate to time, place, frequency or any other character, and statistical devices are used for this purpose.

Concepts in Statistics:

1. Data:

You might be reading a newspaper regularly. Almost every newspaper gives the minimum and the maximum temperature recorded in the city on the previous day. It also indicates the rainfall recorded, and the time of sunrise and sunset. In the school, attendance of the students are recorded in a register regularly.

For a patient, the doctor advises recording of the body temperature at regular intervals. If we record the minimum and maximum temperature, or rainfall, or the time of sunrise and sunset, or attendance of children, or the body temperature of the patient, over a period of time, what we are recording is known as data.

Here we are recording the data of minimum and maximum temperature of the city, data of rainfall, data for the time of sunrise and sunset, and the data pertaining to the attendance of children.

As an example, the class-wise attendance of students, in a school, is as recorded in Table 2.0:

Table 2.0 Class-wise Attendance of Students

Class	No. of Students Present
VI	42
VII	40
VIII	41
IX	35
X	36
XI	32
XII	30
	Total 256

Table 2.0 gives the data for class-wise attendance of students. Here the data comprise 7 observations in all. These observations are, attendance for class VI, VII, and so on. So, data refers to the set of observations, values, elements or objects under consideration. The complete-set of all possible elements or objects is called a population.

Each of the elements is called a piece of data. Data also refers to the known facts or things used as basis for inference or reckoning facts, information, material to be processed or stored.

(ii) Statistics in Mathematics:

Statistics is intimately related to and essentially dependent upon mathematics. The modern theory of Statistics has its foundations on the theory of probability which in turn is a particular branch of more advanced mathematical theory of Measures and Integration. Ever increasing role of mathematics into statistics has led to the development of a new branch of statistics called Mathematical Statistics.

Thus Statistics may be considered to be an important member of the mathematics family. In the words of Connor, "Statistics is a branch of applied mathematics which specialises in data."

(iii) Statistics in Economics:

Statistics and Economics are so intermixed with each other that it looks foolishness to separate them. Development of modern statistical methods has led to an extensive use of statistics in Economics.

All the important branches of Economics—consumption, production, exchange, distribution, public finance—use statistics for the purpose of comparison, presentation, interpretation, etc. Problem of spending of income on and by different sections of the people, production of national wealth, adjustment of demand and supply, effect of economic policies on the economy etc. simply indicate the importance of statistics in the field of economics and in its different branches.

Statistics of Public Finance enables us to impose tax, to provide subsidy, to spend on various heads, amount of money to be borrowed or lent etc. So we cannot think of Statistics without Economics or Economics without Statistics.

(iv) Statistics in Social Sciences:

Every social phenomenon is affected to a marked extent by a multiplicity of factors which bring out the variation in observations from time to time, place to place and object to object. Statistical tools of Regression and Correlation Analysis can be used to study and isolate the effect of each of these factors on the given observation.

mpling Techniques and Estimation Theory are very powerful and indispensable tools for nducting any social survey; pertaining to any strata of society and then analysing the results and ving valid inferences. The most important application of statistics in

sociology is in the field of Demography for studying mortality (death rates), fertility (birth rates), marriages, population growth and so on.

In this context Croxton and Cowden have rightly remarked:

"Without an adequate understanding of the statistical methods, the investigators in the social sciences may be like the blind man groping in a dark room for a black cat that is not there. The methods of statistics are useful in an over-widening range of human activities in any field of thought in which numerical data may be had."

(v) Statistics in Trade:

As already mentioned, statistics is a body of methods to make wise decisions in the face of uncertainties. Business is full of uncertainties and risks. We have to forecast at every step. Speculation is just gaining or losing by way of forecasting. Can we forecast without taking into view the past? Perhaps, no. The future trend of the market can only be expected if we make use of statistics. Failure in anticipation will mean failure of business.

Changes in demand, supply, habits, fashion etc. can be anticipated with the help of statistics. Statistics is of utmost significance in determining prices of the various products, determining the phases of boom and depression etc. Use of statistics helps in smooth running of the business, in reducing the uncertainties and thus contributes towards the success of business.

(vi) Statistics in Research Work:

The job of a research worker is to present the result of his research before the community. The effect of a variable on a particular problem, under differing conditions, can be known by the research worker only if he makes use of statistical methods. Statistics are everywhere basic to research activities. To keep alive his research interests and research activities, the researcher is required to lean upon his knowledge and skills in statistical methods.

Briefly, the advantages of statistical thinking and operations in research are as follows:

1. They permit the most exact kind of description:

The goal of science is description of phenomena. The description should be complete and accurate so that it can be useful to anyone who can understand it when he reads

2. They force us to be definite:

Statistics makes the activities of a researcher definite and exact—both in his procedures and thinking. Statistics systematizes the efforts of a researcher and leads him towards the goal.

3. They help us to summarize the results:

Masses of observations taken by themselves are bewildering and almost meaningless. Statistics enables us to summarize our results in meaningful and convenient form. Before we can see the forest as well as the trees, order must be given to the data. Statistics provides an unrivalled device for bringing order out of chaos, of seeing the general picture in one's results.

4. They enable us to draw general conclusions:

And the process of extracting conclusions is carried out according to accepted rules. Furthermore, by means of statistical steps, we can say about how much faith should be placed in any conclusion and about how far we may extend our generalisation.

5. They enable us to make predictions:

of "how much" of a thing will happen under conditions we know and have measured. For example, we can predict the probable mark a freshman will earn in college algebra if we know his score in a general academic aptitude test, his score in a special algebra-aptitude test, his average mark in high-school mathematics, etc. Our prediction may be somewhat in error, but statistical method will tell us about how much margin of error to allow in making predictions.

6. They enable us to analyze some of the casual factors of complex and otherwise bewildering events.

The Student's Aims in his Study of Statistics:

1. To master the vocabulary of statistics:

In order to read and understand a foreign language, there is always the necessity of building up an adequate vocabulary. To the beginner, statistics should be regarded as a foreign language. The vocabulary consists of concepts that are symbolized by words and by letter symbols.

2. To acquire, or to revive, and to extend skill in computation:

Statistics aims at developing computational skills within the students. The understanding of statistical concepts comes largely through applying them in computing operations.

3. To learn to interpret statistical results correctly

Statistical results can be useful only to the extent that they are correctly interpreted. With full and proper interpretations extracted from data, statistical results are the most powerful source of meaning and significance. Inadequately interpreted, they may represent something worse than wasted effort. Erroneously understood, they are worse than useless.

4. To grasp the logic of statistics:

Statistics provides a way of thinking as well as a vocabulary and a language. It is a logical system, like all mathematics, which is peculiarly adaptable to the handling of scientific problems. Guilford has rightly remarked, "Well-planned investigations always include in their design clear considerations of the specific statistical operations to be employed."

5. To learn where to apply statistics and where not to:

While all statistical devices can illuminate data, each has its limitations. It is in this respect that the average student will probably suffer most from lack of mathematical background, whether he realizes it or not. Every statistics is developed as a purely mathematical idea. As such, it rests upon certain assumptions. If those assumptions are true of the particular data with which we have to deal, the statistics may be appropriately applied.

6. To understand the underlying mathematics of statistics:

This objective will not apply to all students. But it should apply to more than those with unusual previous mathematical training. This will give him a more than commonsense understanding of what goes on in the use of formulas.

Statistics in Psychology and Education:

Statistics has been used very widely in Psychology and education too e.g., in the scaling of mental tests and other psychological data; for measuring the reliability and validity of tests scores ; for determining the Intelligence Quotient; in item analysis and factor

new discipline called ...

Modern problems and needs are forcing statistical methods and ideas more and more to the fore. There are so many things we wish to know which cannot be discovered by a single observation or by a single measurement. We wish to envisage the behaviour of a man who, like all men, is rather a variable quantity, and must be observed repeatedly and not once for all. We wish to study the social group, composed of individuals differing from one another.

We should like to be able to compare one group with another, one race with another, as well as one individual with another individual, or the individual with the norm for his age, race or class. We wish to trace the curve which pictures the growth of a child, or of a population. We wish to disentangle the interwoven factors of heredity and environment. The only solution is the application of statistics in above mentioned areas.

Knowledge of statistics is particularly useful to the students of psychology and education for the following reasons:

1. It helps in understanding the modern literature in these subjects. Most books and articles in research journals in these subjects use statistical terminology and present the results in a statistical form, which cannot be understood without adequate knowledge of statistics.

2. It helps in conducting research investigations for which sample survey or experimental approach has to be used. Knowledge of sample survey methods, design of experiments and statistical methods of data analysis is essential for the advanced students who have to conduct their own investigations.

3. It forms the basis of scientific approach to problems, in which inductive inference is commonly used. Students of psychology and education cannot afford to remain ignorant of the scientific method of approach to problem-solving in their disciplines.

4. It helps the professional psychologist, whether a counsellor, a guidance worker or a clinical psychologist, in doing his work efficiently, since in the course of his work he has to administer tests, interprete test scores and maintain a record of a number of cases (which constitute data requiring statistical analysis for proper interpretation). For all this the knowledge of statistics is essential.

5. It provides basic tools of data analysis to educationists who are engaged in planning and administration of an educational system. They need to know statistics in order to study past trends of enrolment, to estimate teacher requirements, to plan new schools and for many other such purposes.

6. It helps the teachers and school administrators in evaluating the performance of students and schools. They have to know some statistics in order to deal with examination data, test scores of students and quantitative data used for different types of evaluation.

7. In psychology and education, quantitative methods are being increasingly used to study various phenomena, for which statistical techniques are indispensable. In psychophysics, one studies relationship between measurements obtained by instruments and by human judgement.

Attempts are made to measure human ability (e.g. intelligence, scholastic aptitude, creativity, personality, interest, behaviour, attitude, etc.) by tests. The knowledge of statistical techniques is required for understanding and solving problems in all these situations, which are quite common in the fields of psychology and education.

Statistics Helps a Teacher:

The best part of statistics is that, 'how does it help a teacher in meeting the instructional or teaching objectives in classroom situations?' Thus, in the real sense, it is concerned with the organisation, analysis and interpretation of test scores and other numerical data.

It is necessary for a teacher to know all the statistical techniques which help him to:

(i) Analyse and describe the results of measurement obtained in his classroom.

(ii) Understand the statistics used in test manuals and research reports.

(iii) And interprete the various types of derived scores used in testing.

If a teacher has an elementary knowledge of statistical measures and its uses, certainly he will improve his effectiveness in teaching and thus statistics shall be a great help to him in his mission.

In the light of above discussions the statistics has to achieve the following set of objectives in Psychology and Education as well as in general type of

(i) In gathering information's of various aspects to test assumptions or test hypothesis.

(ii) In observation, selection, collection, organisation and analysis of facts and data of various nature.

(iii) In underlining or deriving different methodology for uses.

(iv) In knowing the central tendency of a group, variations in its folds and norms of its structure or consolidation.

(v) In testing the reliability, validity, usability and comprehensiveness of a test result.

(vi) In deciding the procedures and techniques of a test preparation and its uses.

(vii) In deducing results and conclusions.

Data Analysis:

Statistical data are analysed in two ways which are stated below:

(i) Descriptive Statistics:

The descriptive statistics is concerned with describing or summarising the numerical properties of data. The methodology of descriptive statistics includes classification, tabulation, graphical representation and calculation of certain indicators such as mean, median, range, etc. which summarise certain important features of data.

It restricts to generalisation and to specifically a particular group of individuals being observed. No conclusion can be drawn beyond this group. The data describe only one group on which these have been collected. Many such action researches involve descriptive analysis. These researches provide worthy information's regarding the nature of a specific group of individuals.

(ii) Inferential Statistics:

Inferential statistics, which is also referred to as statistical inference, is concerned with derivation of scientific inference about generalisation of results from the study of a few particular cases.

Technically speaking, the methods of statistical inference help in generalising the results of a sample to the entire population from which the sample is drawn. It should be kept in mind while selecting a sample that it should approximately represent the larger group of population. Thus the characteristics of the sample will represent the characteristics of the total groups.

The nature of inference is inductive in the sense that we make general statements from the study of a few cases. Inferential statistics provides us the tools of making inductive inference scientific and rigorous. In such inference, it is presumed that the generalisation cannot be made with certainty.

Some uncertainty is inevitable since in some cases the inference drawn from the data of a sample survey or an experiment can be wrong. However, the degree of uncertainty is itself measurable and one can make rigorous statements about the uncertainty (or the chance of being wrong) associated with a particular inference. This uncertainty in inference is dealt with by applying the theory of probability, which is the backbone of statistical inference.

It is a branch of mathematical statistics that deals with measurement of the extent of certainty of events whose occurrence depends on chance.

Statistics in business and management:

1. *Marketing:* Statistical analysis are frequently used in providing information for making decision in the field of marketing it is necessary first to find out what can be sold and the to evolve suitable strategy, so that the goods which to the ultimate consumer. A skill full analysis of data on production purchasing power, man power, habits of compotators, habits of consumer, transportation cost should be consider to take any attempt to establish a new market.

2. *Production:* In the field of production statistical data and method play a very important role. The decision about what to produce? How to produce? When to produce? For whom to produce is based largely on statistical analysis.

3. *Finance:* The financial organization discharging their finance function effectively depend very heavily on statistical analysis of peat and tigers.

4. *Banking:* Banking institute have found if increasingly to establish research department within their organization for the purpose of gathering and analysis information, not only regarding their own business but also regarding general economic situation and every segment of business in which they may have interest.

5. *Investment:* Statistics greatly assists investors in making clear and valued judgment in his investment decision in selecting securities which are safe and have the best prospects of yielding a good income.

6. *Purchase:* the purchase department in discharging their function makes use of statistical data to frame suitable purchase policies such as what to buy? What quantity to buy? What time to buy? Where to buy? Whom to buy?

7. *Accounting:* statistical data are also employer in accounting particularly in auditing function. the technique of sampling and destination is frequently used.

8. *Control:* the management control process combines statistical and accounting method in making the overall budget for the coming year including sales, materials, labor and other costs and net profits and capital requirement.

Limitations of Statistics:

1. Qualitative Aspect Ignored:

The statistical methods don't study the nature of phenomenon which cannot be expressed in quantitative terms. Such phenomena cannot be a part of the study of statistics. These include health, riches, intelligence etc. It needs conversion of qualitative data into quantitative data. So experiments are being undertaken to measure the reactions of a man through data. Now a days statistics is used in all the aspects of the life as well as universal activities.

2. It does not deal with individual items:

It is clear from the definition given by Prof. Horace Sacrist, "By statistics we mean aggregates of facts.... and placed in relation to each other", that statistics deals with only aggregates of facts or items and it does not recognize any individual item. Thus, individual terms as death of 6 persons in a accident, 85% results of a class of a school in a particular year, will not amount to statistics as they are not placed in a group of similar items. It does not deal with the individual items, however, important they may be.

3. It does not depict entire story of phenomenon:

When even phenomena happen, that is due to many causes, but all these causes can not be expressed in terms of data. So we cannot reach at the correct conclusions. Development of a group depends upon many social factors like, parents' economic condition, education, culture, region, administration by government etc. But all these factors cannot be placed in data. So we analyse only that data we find quantitatively and not qualitatively. So results or conclusion are not 100% correct because many aspects are ignored.

4. It is liable to be miscued:

As W.I. King points out, "One of the shortcomings of statistics is that do not bear on their face the label of their quality." So we can say that we can check the data and procedures of its approaching to conclusions. But these data may have been collected by inexperienced persons or they may have been dishonest or biased. As it is a delicate science and can be easily misused by an unscrupulous person. So data must be used with a caution. Otherwise results may prove to be disastrous.

5. Laws are not exact:

As far as two fundamental laws are concerned with statistics:
(i) Law of inertia of large numbers and
(ii) Law of statistical regularity, are not as good as their science laws.
They are based on probability. So these results will not always be as good as of scientific laws. On the basis of probability or interpolation, we can only estimate the production of paddy in 2008 but cannot make a claim that it would be exactly 100 %. Here only approximations are made.

6. Results are true only on average: As discussed above, here the results are interpolated for which time series or regression or probability can be used. These are not absolutely true. If average of two sections of students in statistics is same, it does not mean that all the 50 students is section A has got same marks as in B. There may be much variation between the two. So we get average results.

"Statistics largely deals with averages and these averages may be made up of individual items radically different from each other." —W.L King

7. To Many methods to study problems:

In this subject we use so many methods to find a single result. Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case. "It must not be assumed that the statistics is the only method to use in research; neither should this method of considered the best attack for the problem." —Croxten and Cowden

8. Statistical results are not always beyond doubt:

Statistics deals only with measurable aspects of things and therefore, can seldom give the complet olution to problem. They provide a basis for judgment but not the whole judgment." —Prof. L.R. Conn

though we use many laws and formulae in statistics but still the results achieved are not final a nclusive. As they are unable to give complete solution to a problem, the result must be taken and u h much wisdom.

ure /Features /Characteristics of statistics

It is an aggregate of facts.
Analysis of multiplicity of causes.
It is numerically expressed.

- It is estimated according to reasonable standard of accuracy.
- It is collected for pre-determined purpose.
- It is collected in a systematic manner.

Division of Statistics:

- **Theoretical:** Mathematical theory which is the basis of the science of statistics is called theoretical statistics.
- **Statistical Methods:** By this method we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

Few Methods are:-
(1) Collection of Data (2) Classification (3) Tabulation (4) Presentation (5) Analysis (6) Interpretation (7) Forecasting.

- **Applied:** - It deals with the application of rules and principles developed for specific problem in different disciplines.

Eg: - Time series, Sampling, Statistical Quality control, design of experiments.

Discuss the functions and importance/utility of Statistics.

Ans.: Statistical methods are used not only in the social, economic and political fields but in every field of science and knowledge. Statistical analysis has become more significant in global relations and in the age of fast developing information technology.

According to **Prof. Bowley, *"The proper function of statistics is to enlarge individual experiences".***

Following are some of the **important functions of Statistics :**

a) To provide numerical facts.
b) To simplify complex facts.
c) To enlarge human knowledge and experience.
d) Helps in formulation of policies.
e) To provide comparison.
f) To establish mutual relations.
g) Helps in forecasting.
h) Test the accuracy of scientific theories.
i) To study extensively and intensively.

The use of statistics has become almost essential in order to clearly understand and solve a problem. Statistics proves to be much useful in **unfamiliar fields of application and complex situations** such as :-

a) Planning
b) Administration
c) Economics
d) Trade & Commerce
e) Production management
f) Quality control
g) Helpful in inspection
h) Insurance business
i) Railways & transport Co
a) Banking Institutions
b) Speculation and Gambling
c) Underwriters and Investors
d) Politicians & social workers.

How many types of Series are there on the basis of Quantitative Classification? Give the difference between Exclusive and Inclusive Series.

There are three types of frequency distributions :

(i) Individual Series: In individual series, the frequency of each item or value is only one for example; marks scored by 10 students of a class are written individually.

(ii) Discrete Series: A discrete series is that in which the individual values are different from each other by a different amount.

For example:

Daily wages	5	10	15	20
No. of workers	6	9	8	5

(iii) Continuous Series : When the number of items are placed within the limits of the class, the series obtained by classification of such data is known as continuous series.

For example

Marks obtained	0-10	10-20	20-30	30-40
No.of students	10	18	22	25

Difference between Exclusive and Inclusive Series

Exclusive Series	Inclusive Series
The two limits are not equal. class is equal to the lower limit of next class.	Limits Upper limit of one
The value equal to the upper limit is included in the next class.	Both upper & lower limits are included in the same class
It does not require any conversion	Inclusive series is converted into exclusive series for calculation purpose
It is suitable in all situations.	It is suitable only when the values are in integers.

TYPES OF DATA AND DATA SOURCES

Statistical data are the basic raw material of statistics. Data may relate to an activity of our interest, a phenomenon, or a problem situation under study. They derive as a result of the process of measuring, counting and/or observing. Statistical data, therefore, refer to those aspects of a problem situation that can be measured, quantified, counted, or classified. Any object subject phenomenon, or activity that generates data through this process is termed as a variable. In other words, a variable is one that shows a degree of variability when successive measurements are recorded. In statistics, data are classified into two broad categories: quantitative data and qualitative data. This classification is based on the kind of characteristics that are measured.

Quantitative data are those that can be quantified in definite units of measurement. These refer to characteristics whose successive measurements yield quantifiable observations. Depending on the nature of the variable observed for measurement, quantitative data can be further categorized as continuous and discrete data.

Obviously, a variable may be a continuous variable or a discrete variable.

(i) Continuous data represent the numerical values of a continuous variable. A continuous variable is the one that can assume any value between any two points on a line segment, thus representing an interval of values. The values are quite precise and close to each other, yet distinguishably different. All characteristics such as weight, length, height, thickness, velocity, temperature, tensile strength, etc., represent continuous variables. Thus, the data recorded on these and similar other characteristics are called continuous data. It may be noted that a continuous variable assumes the finest unit of measurement. Finest in the sense that it enables measurements to the maximum degree of precision.

(ii) Discrete data are the values assumed by a discrete variable. A discrete variable is the one whose outcomes are measured in fixed numbers. Such data are essentially count data. These are derived from a process of counting, such as the number of items possessing or not possessing a certain characteristic. The number of customers visiting a departmental store everyday, the incoming flights at an airport, and the defective items in a consignment received for sale, are all examples of discrete data.

Qualitative data refer to qualitative characteristics of a subject or an object. A characteristic is qualitative in nature when its observations are defined and noted in terms of the presence or absence of a certain attribute in discrete numbers. These data are further classified as nominal and rank data.

Nominal data are the outcome of classification into two or more categories of items or units comprising a sample or a population according to some quality characteristic. Classification of students according to sex (as males and females), of workers according to skill (as skilled, semi-skilled, and unskilled), and of employees according to the level of education (as matriculates, undergraduates, and post-graduates), all result into nominal data. Given any such basis of classification, it is always possible to assign each item to a particular class and make a summation of items belonging to each class. The count data so obtained are called nominal data.

Rank data, on the other hand, are the result of assigning ranks to specify order in terms of the integers 1,2,3, ..., n. Ranks may be assigned according to the level of performance in a test, a contest, a competition, an interview, or a show. The candidates appearing in an interview, for example, may be assigned ranks in integers ranging from I to n, depending on their performance in the interview. Ranks so assigned can be viewed as the continuous values of a variable involving performance as the quality characteristic.

Data sources could be seen as of two types, viz., secondary and primary. The two can be defined as under:

(i) Secondary data: They already exist in some form: published or unpublished - in an identifiable secondary source. They are, generally, available from published source(s), though not necessarily in the form actually required. **(ii) Primary data:** Those data which do not already exist in any form, and thus have to be collected for the first time from the primary source(s). By their very nature, these data require fresh and first-time collection covering the whole population or a sample drawn from it.

Collection of data

For studying a problem statistically, first of all, the data relevant thereto must be collected. The numerical facts constitute the raw material of the statistical process. The interpretation of the ultimate conclusion and the decisions depend upon the accuracy with which the data are collected. Unless the data are collected with sufficient care and are as accurate as is necessary for the purposes of the inquiry, the result obtained cannot be expected to be valid or reliable. Before starting the collection of the data, it is necessary to know the sources from which the data are to be collected.

Primary and Secondary Sources

The original compiler of the data is the primary source. For example, the office of the Registrar General will be the primary source of the decennial population census figures.

A secondary source is the one that furnishes the data that were originally compiled by someone else.

If the population census figures issued by the office of the Registrar-General are published in the Indian year Book, this publication will be the secondary source of the population data.

The source of data also are classified according to the character of the data yielded by them. Thus the data which are gathered from the primary source is known as primary data and the one gathered from the secondary source is known as secondary data. When an investigator is making use of figures which he has obtained by field enumeration, he is said to be using primary data and when he is making use of figures which he has obtained from some other source, he is said to be using secondary data.

Choice between Primary and Secondary Data

An investigator has to decide whether he will collect fresh (primary) data or he will compile data from the published sources. The former is reliable per se but the latter can be relied upon only by examining the following factors :—

(i) source from which they have been obtained;
(ii) their true significance;
(iii) completeness and
(iv) method to collection.

In addition to the above factors, there are other factors to be considered while making choice between the primary or secondary data :
(i) Nature and scope of enquiry.
(ii) Availability of time and money.
(iii) Degree of accuracy required and

(iv) The status of the investigator i.e., individual, Pvt. Co., Govt. etc.

However, it may be pointed out that in certain investigations both primary and secondary data may have to be used, one may be supplement to the other.

Methods of Collection of Primary Data

The primary methods of collection of statistical information are the following :

1. Direct Personal Observation,
2. Indirect Personal Observation,
3. Schedules to be filled in by informants
4. Information from Correspondents, and
5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

1. Direct Personal Observation

According to this method, the investigator obtains the data by personal observation. The method is adopted when the field of inquiry is small. Since the investigator is closely connected with the collection of data, it is bound to be more accurate. Thus, for example, if an inquiry is to be conducted into the family budgets and giving conditions of industrial labour, the investigation himself live in the industrial area as one of the industrial workers, mix with other residents and make patience and careful personal observation regarding how they spend, work and live.

2. Indirect Personal Observation

According to this method, the investigator interviews several persons who are either directly or indirectly in possession of the information sought to be collected. It may be distinguished form the first method in which information is collected directly from the persons who are involved in the inquiry. In the case of indirect personal observation, the persons from whom the information is being collected are known as witnesses or informants. However it should be ascertained that the informants really passes the knowledge and they are not prejudiced in favour of or against a particular view point. This method is adopted in the following situations :

1. Where the information to be collected is of a complete nature.
2. When investigation has to be made over a wide area.
3. Where the persons involved in the inquiry would be reluctant to part with the information.

This method is generally adopted by enquiry committee or commissions appointed by government.

3. Schedules to be filled in by the informants

Under this method properly drawn up schedules or blank forms are distributed among the persons from whom the necessary figure are to be obtained. The informants would fill in the forms and return them to the officer incharge of investigation. The Government of India issued slips for the special enumeration of scientific and technical personnel at the time of census. These slips are good examples of schedules to be filled in by the informants.

The merit of this method is its simplicity and lesser degree of trouble and pain for the investigator. Its greatest drawback is that the informants may not send back the schedules duly filled in.

4. Information from Correspondents

Under this method certain correspondent are appointed in different parts of the field of enquiry, who submit their reports to the Central Office in their own manner. For example, estimates of agricultural wages may be periodically furnished to the Government by village school teachers.

The local correspondents being on the spot of the enquiry are capable of giving reliable information.

But it is not always advisable to place much reliance on correspondents, who have often got their own personal prejudices. However, by this method, a rough and approximate estimate is obtained at a very low cost. This method is also adopted by various departments of the government in such cases where regular information is to be collected from a wide area.

Questionnaire incharge of Enumerations

A questionnaire is a list of questions directly or indirectly connected with the work of the enquiry. The answers to these questions would provide all the information sought. The questionnaire is put in the charge of trained investigators whose duty is to go to all persons or selected persons connected with the enquiry. This method is usually adopted in case of large inquiries. The method of collecting data is relatively cheap. Also the information obtained is that of good quality.

The main drawback of this method is that the enumerator (i.e., investigator in charge of the questionnaire) may be a biased one and may not enter the answer given by the information. Where there are many enumerators, they may interpret various terms in questionnaire according to their whims. To that extent the information supplied may be either inaccurate or inadequate or not comparable.

This drawback can be removed to a great extent by training the investigators before the enquiry begins. The meaning of different questions may be explained to them so that they do not interpret them according to their whims.

Sources of Secondary Data

There are number of sources from which secondary data may be obtained. They may be classified as follow. :

1. Published sources, and
2. Unpublished sources.

1. Published Sources

The various sources of published data are :

1. Reports and official publications of-

(a) International bodies such as the International Monetary Fund, International Finance Corporation, and United Nations Organisation.

(b) Central and State Governments- such as the Report of the Patel Committee, etc.

2. Semi Official Publication. Various local bodies such as Municipal Corporation, and Districts Boards.

3. Private Publication of—

(a) Trade and professional bodies such as the Federation of India, Chamber of Commerce and Institute of Chartered Accountants of India.

(b) Financial and Economic Journals such as "Commerce", 'Capital' etc.

(c) Annual Reports of Joint Stock Companies.

(d) Publication brought out by research agendas, research scholars, etc.

2. Unpublished Sources

There are various sources of unpublished data such as records maintained by various government and private offices, studies made by research institutions, scholars, etc. such source can also be used where necessary.

UNIT 2

Collection and Organization of Data

Unit 2
(Data presentations and graphically display)

ortance of Diagrammatic and Graphic Representation of Data
ram
quency Polygon
ency Curve
ves Curves/cumulative frequency curves:

antages of Diagrams and Graphs
llowing are the advantages of diagrams and graphs:

Diagram gives an alternative and elegant presentation: Diagrammatic presentations of the data directly attract people, give delight to the eye and add to the statistics.
Diagrams leave good visual impact: The visual impact of the diagram stresses the mind of the people to think about the situation of the statistics.
They facilitate comparison: It makes easy to compare the data by visualizing the fact in front of the observer.
Saves time: Diagrams present the set of data in such a way that their significance known without loss of much time.
Diagrams simplify complexity and depict the characteristics of the data in simple manner.
Graphs reveal trend of the data series which is helpful for simple forecasting.

Frequency Histogram

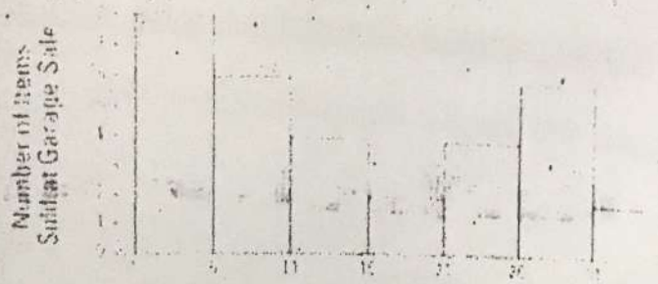
One of the more commonly used pictorials in statistics is the frequency histogram, which in some ways is similar to a bar chart and tells how many items are in each numerical category. For example, suppose that after a garage sale, you want to determine which items were the most popular: the high-priced items, the low-priced items, and so forth. Let's say you sold a total of 32 items for the following prices: 1, 2, 2, 2, 5, 5, 5, 5, 7, 8, 10, 10, 10, 10, 11, 15, 15, 15, 19, 20, 21, 21, 25, 25, 29, 29, 29, 30, 30, 30, 35, and 35.

items sold *ranged* in price from 1 to 35. First, divide this **range** of 1 to 35 into a number of categories, led **class intervals**. Typically, no fewer than 5 and no more than 20 class intervals work best for a frequency ogram.

ose the first class interval to include your lowest (smallest value) data and make sure that no *overlap* exists that one piece of data does not fall into two class intervals. For example, you would not have your first class val be 1 to 5 and your second class interval be 5 to 10 because the four items that sold for 5 would belong both the first and the second intervals. Instead, use 1 to 5 for the first interval and 6 to 10 for the second. ss intervals are mutually exclusive.

t, make a table of how your data is distributed (see Table 1). The number of observations that falls into each lass interval is called the **class frequency**.

Class	Interval	Frequency
1	$1 to $5	8
2	$6 to $10	6
3	$11 to $15	4
4	$16 to $20	2
5	$21 to $25	4
6	$26 to $30	6
7	$31 to $35	2



Definition of Frequency Polygons

Statistics deals with the collection of data and information for a particular purpose. The tabulation of each run for each ball in cricket gives the statistics of the game. Tables, graphs, pie-charts, bar graphs, histograms, frequency polygons etc. are used to represent statistical data pictorially.
In the upcoming discussion let us discuss about frequency polygons. Frequency polygons are visually substantial method of representing quantitative data and its frequencies.

To draw frequency polygons, we begin with, drawing histograms and follow the following steps:

- Choose the class interval and mark the values on the horizontal axes
- Mark the mid value of each interval on the horizontal axes.
- Mark the frequency of the class on the vertical axes.
- Corresponding to the frequency of each class interval, mark a point at the height in the middle of the class interval
- Connect these points using line segment..
- The obtained representation is a frequency polygon.

Let us consider an example to understand frequency polygons better.

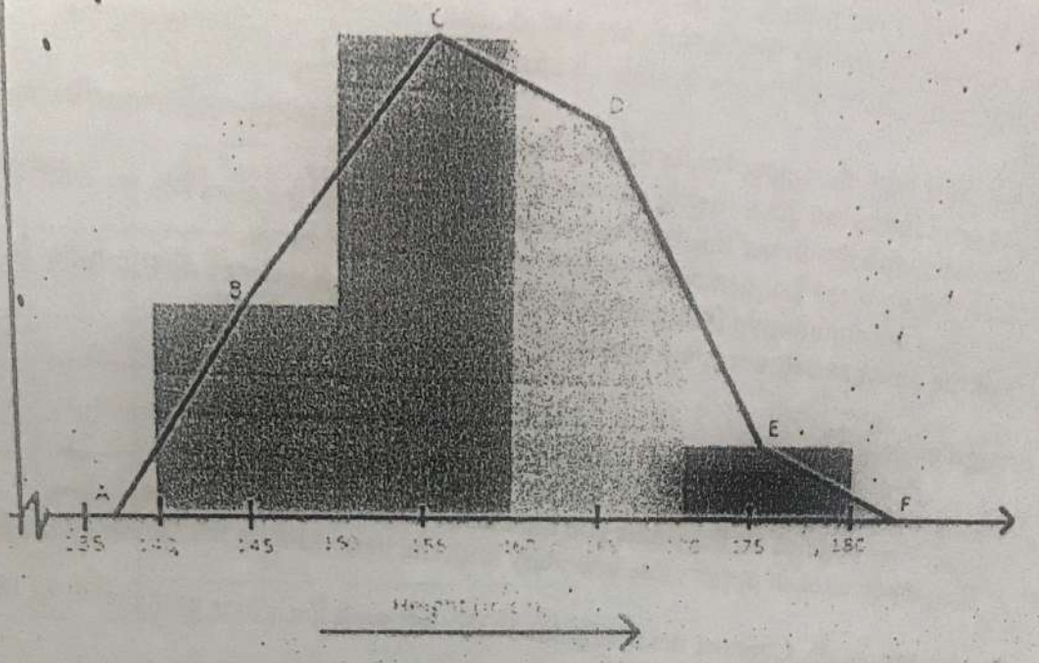
Example 1: In a batch of 400 students, the height of students is given in the following table. Represent it through a frequency polygon.

Height (in cm)	Number of Students(Frequency)
140 – 150	74
150 – 160	163
160 – 170	135
170 – 180	28
Total	400

Solution: Following steps are to be followed to construct a histogram from the given data:

- The heights are represented on the horizontal axes on a suitable scale as shown.
- The number of students is represented on the vertical axes on a suitable scale as shown.
- Now rectangular bars of widths equal to the class- size and the length of the bars corresponding to frequency of the class interval is drawn.

ABCDEF represents the given data graphically in form of frequency polygon as:

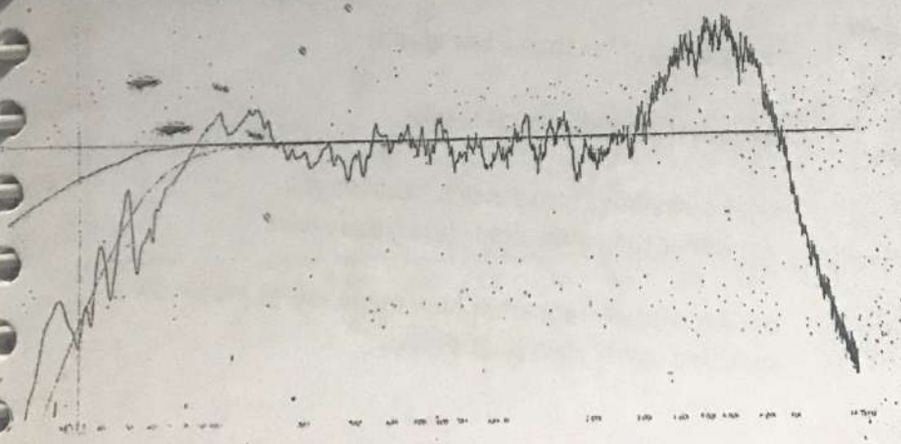


ency polygons can also be drawn independently without drawing histograms. For this the mid points of the ss intervals known as class marks are used to plot the points.

$$ss\ Mark = \frac{Upper\ Limit + Lower\ Limit}{2}$$

at is frequency curve? How do we draw it?

uency curve is obtained by joining the points of frequency polygon by a freehand smoothed curve. Unlike equency polygon, where the points we joined by straight lines, we make use of free hand joining of those points rder to get a smoothed frequency curve. It is used to remove the ruggedness of polygon and to present it in a ood form or shape. We smoothen the angularities of the polygon only without making any basic change in the hape of the curve. In this case also the curve begins and ends at base line, as is in case of polygon. Area under curve must remain almost the same as in the case of polygon.



At times we are interested in knowing how many workers of a factory earn less than Rs. 700 per month, or how many workers earn more than Rs: 1000 per month, or percentage of students who have failed, etc.
So, it is necessary to add the frequencies, when frequencies are added. They are called cumulative Frequency.
The curve abstained by plotting cumulative frequency is called a cumulative frequency curve or ogives.
There are two types of ogives

(1) Less than ogives:- we start with the upper limits of the classes and go on additing the frequencies, where this frequencies are plotted we get a rising curve.
(2) More than ogives:- we start with the lower limits of the classes and from the frequencies, we subtract the frequency of each class. When these frequency are plotted we get a decline curve.

An ogive is a graph that represents the cumulative frequencies for the classes in a frequency distribution. It shows how many of values of the data are below certain boundary.

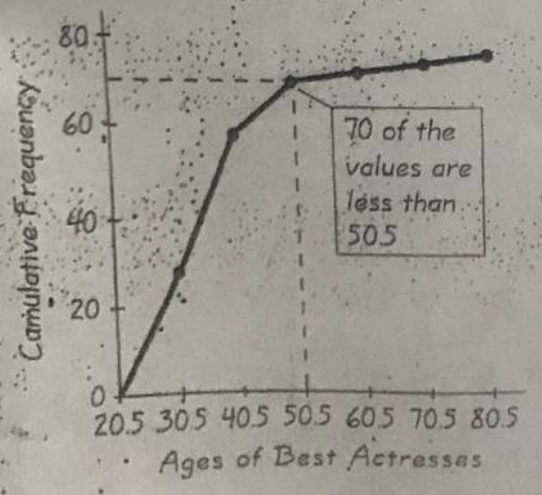
Steps for constructing an ogive:

1. Draw and label the x (horizontal) and the y (vertical) axes.
2. Represent the cumulative frequencies on the y axis and the class boundaries on the x axis.
3. Plot the cumulative frequency at each upper class boundary with the height being the corresponding cumulative frequency.
4. Connect the points with segments. Connect the first point on the left with the x axis at the level of the lowest lower class boundary.

Note: For the ogive we need the class boundaries and the cumulative frequencies

Example: Construct a histogram for the frequency distribution for the record high temperatures of the 50 states.

Class Limits	Class Boundaries	Frequency (f)	Cumulative Frequency	Midpoints (Xm)
100-104	99.5-104.5	2	2	102
105-109	104.5-109.5	8	10	107
110-114	109.5-114.5	18	28	112
115-119	114.5-119.5	13	41	117
120-124	119.5-124.5	7	48	122
125-129	124.5-129.5	1	49	127
130-134	129.5-134.5	1	50	132



Line graph (rather than a bar graph)

Uses class boundaries on x-axis

Uses cumulative frequencies (total as you go) rather than individual class frequencies

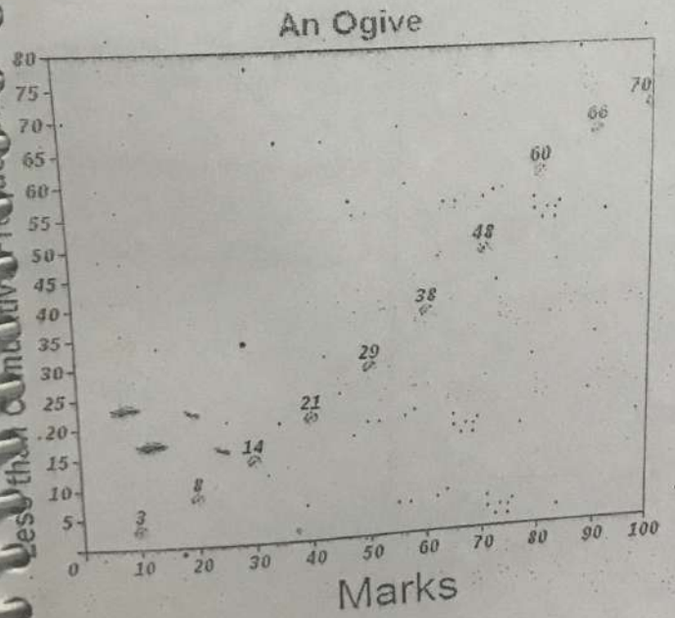
Used to visually represent how many values are below a specified upper class boundary

n: For the data given below, construct a less than cumulative frequency table and plot its ogive.

	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90	90 -100
quency	3	5	6	7	8	9	10	12	6	4

ks	Frequency	Less than cumulative frequency
10	3	3
) - 20	5	8
30	6	14
) - 40	7	21
50	8	29
) - 60	9	38
70	10	48
) - 80	12	60
90	6	66
0 - 100	4	70

the points having abscissa as upper limits and ordinates as the cumulative frequencies (10, 3), (20, 8), (30, 14),), (50, 29), (60,38), (70, 48), (80, 60), (90, 66), (100, 70) and join the points by a smooth curve.

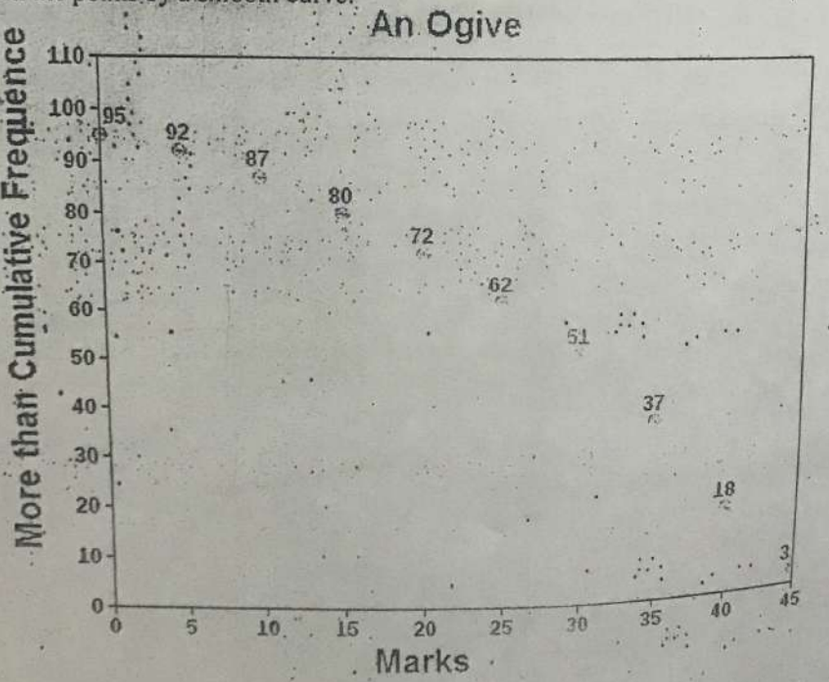


Marks	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50
Frequency	3	5	7	8	10	11	14	19	15	13

Solution:

Marks	Frequency	More than cumulative frequency
0 - 5	3	95
5 - 10	5	95 - 3 = 92
10 - 15	7	92 - 5 = 87
15 - 20	8	87 - 7 = 80
20 - 25	10	80 - 8 = 72
25 - 30	11	72 - 10 = 62
30 - 35	14	62 - 11 = 51
35 - 40	19	51 - 14 = 37
40 - 45	15	37 - 19 = 18
45 - 50	13	18 - 15 = 3

On the graph, plot the points (0, 95), (5, 92), (10, 87), (15, 80), (20, 72), (25, 62), (30, 51), (35, 37), (40, 18), (45, 3) and join the points by a smooth curve.



Draw 'more than' and 'less than' ogive curves for the following data:

Class Interval	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50	50 - 55	55 - 60	60 - 65	65 - 70	70 - 75
Frequency	2	5	8	10	13	17	20	16	12	18	19	20

Solution:

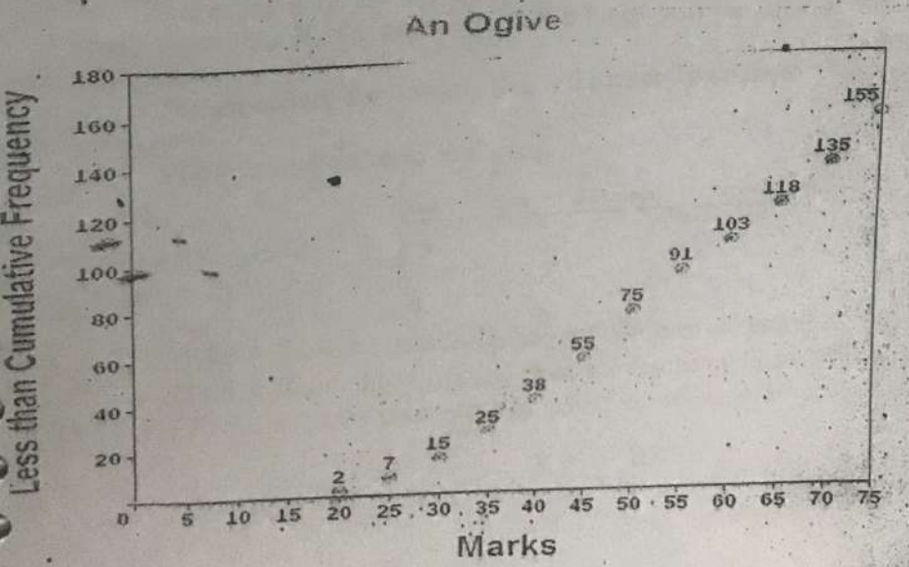
Let us calculate cumulative frequencies as follows:

Class Interval	Frequency	Less than cumulative frequency	More than cumulative frequency
15 - 20	2	2	155
20 - 25	5	2 + 5 = 7	155 - 2 = 153
25 - 30	8	7 + 8 = 15	153 - 5 = 148
30 - 35	10	15 + 10 = 25	148 - 8 = 140
35 - 40	13	25 + 13 = 38	140 - 10 = 130
40 - 45	17	38 + 17 = 55	130 - 13 = 117
45 - 50	20	55 + 20 = 75	117 - 17 = 100
50 -55	16	75 + 16 = 91	100 - 20 = 80
55 - 60	12	91 + 12 = 103	80 - 16 = 64
60 - 65	15	103 + 15 = 118	64 - 12 = 52
65 - 70	17	118 + 17 = 135	52 - 15 = 37
70 - 75	20	135 + 20 = 155	37 - 17 = 20

Less than ogive:

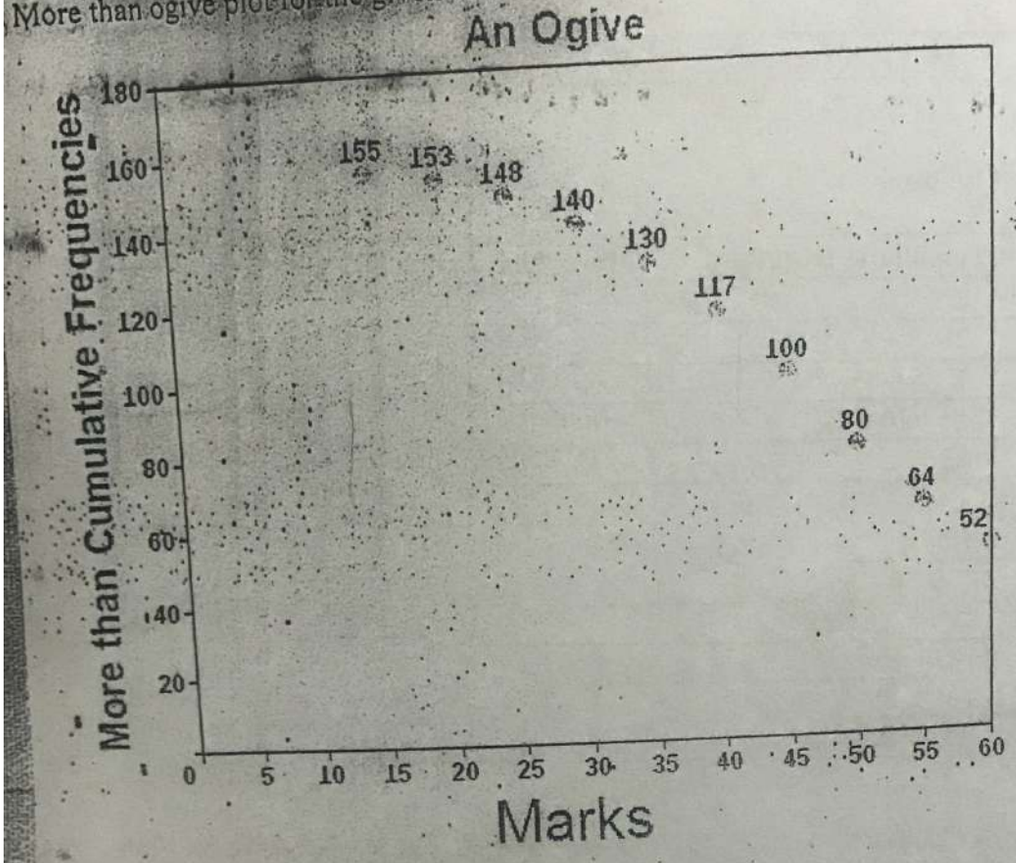
For **less than ogive**, plot the points. (20, 2), (25, 7), (30, 15), (35, 25), (40, 38), (45, 55), (50, 75), (55, 91), (60, 103), (65, 118), (70, 135), (75, 155) and join the points by smooth curve.

Less than ogive plot for the given data is.



80), (55, 64), (60, 52), (65, 37), (70, 20) and join the points by a [illegible]

More than ogive plot for the given data is as follow



Frequency distribution and its types

In order to describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a ***frequency distribution***.

A ***frequency distribution*** is the organizing of raw data in table form, using classes and frequencies.

There are three basic types of frequency distributions, and there are specific procedures for constructing each type. The three types are ***categorical, ungrouped,*** and ***grouped frequency distributions.***

1- Categorical frequency distribution: - The categorical frequency distribution is used for data that can be placed in specific categories, such as nominal - or ordinal-level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

Example -1-

Twenty-five army inductees were given a blood test to determine their blood type. The data set is as follows:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

Solution

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

The procedure for constructing a frequency distribution for categorical data is given next.

STEP 1: Make a table as shown.

A	B	C	D
Class	Tally	Frequency	Percent
A			
B			
O			
AB			

STEP 2: Tally the data and place the results in column B.

STEP 3: Count the tallies and place the results in column C.

STEP 4: Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100\%$$

where

f = frequency of the class
n = total number of values

For example, in the class of type A blood, the percentage is

$$\%A = \frac{5}{25} \cdot 100\% = 20\%$$

STEP 5: Find the totals for columns C and D (see the completed table that follows).

Class	Tally	Frequency	Percent
A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
	Total	25	100

2- Ungrouped frequency distribution: - When the data are numerical instead of categorical, the procedure for constructing a frequency distribution is somewhat more complicated.

Example -2-

A psychologist administered a test of manual dexterity to 25 third-grade students. The times, in minutes, required to complete the test are given below. Construct a frequency distribution for the data.

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

Solution

STEP 1 Find the rage of the data. The **range**, R, is defined as

$$R = \text{highest value} - \text{lowest value}$$

For this data set, the range is $11 - 4$, or 7. Since the range is small, classes consisting of a single data value can be used. They are 4, 5, 6, 7, 8, 9, 10, and 11.

STEP 2 Make a table as shown next.

STEP 3 Tally the data.

STEP 4 Complete the frequency column.

Class	Tally	Frequency
4	//	2
5	///	3
6	/	1
7	𝍸	5
8	𝍸 //	7
9	////	4
10	//	2
11	/	1

Construct boundaries for each class by subtracting 0.5 from each class value and adding 0.5 to each class value, as shown next.

Class	Class boundaries	Tally	Frequency
4	3.5–4.5		2
5	4.5–5.5		3
6	5.5–6.5		1
7	6.5–7.5		5
8	7.5–8.5		7
9	8.5–9.5		4
10	9.5–10.5		2
11	10.5–11.5		1

Add a cumulative frequency (cf) to the frequency distribution shown above by adding the frequency in each class to the total of the frequencies of the classes above that class, as shown next:

Class	Class boundaries	Frequency	Cumulative frequency
4	3.5–4.5	2	0 + 2 = 2
5	4.5–5.5	3	2 + 3 = 5
6	5.5–6.5	1	5 + 1 = 6
7	6.5–7.5	5	6 + 5 = 11
8	7.5–8.5	7	11 + 7 = 18
9	8.5–9.5	4	18 + 4 = 22
10	9.5–10.5	2	22 + 2 = 24
11	10.5–11.5	1	24 + 1 = 25

Cumulative frequencies are used to show how many values are accumulated up to and including a specific class. For example, 18 students successfully completed the test in 8 minutes or less; 24 students completed the test in 10 minutes or less.

3- Grouped frequency distribution: - When the range of the data is large, the data must be grouped into classes that are more than one unit in width. For example, a distribution of the number of hour's boat batteries lasted is as follows:

Class limits	Class boundaries	Tally	Frequency	Cumulative frequency
24–30	23.5–30.5		3	3
31–37	30.5–37.5		1	4
38–44	37.5–44.5		5	9
45–51	44.5–51.5		9	18
52–58	51.5–58.5		6	24
59–65	58.5–65.5		1	25
			25	

In this distribution, the values 24 and 30 of the first class are called *class limits*. The *lower class limit* is 24; it represents the smallest data value that can be included in the class. The *upper class limit* is 30; it represents the largest data value that can be included in the class. The numbers in the second column are called class boundaries. These numbers are used to separate the classes so that there are no gaps in the

frequency distribution. The gaps are due to the limits; for example, there is a gap between 30 and 31.

(lower limit)	31 − 0.5 = 30.5	(lower boundary)
(upper limit)	37 + 0.5 = 37.5	(upper boundary)

Finally, the class width for a class in a frequency distribution is found by subtracting lower (or upper) class limit of one class minus the lower (or upper) class limit of the previous class. For example, the class width in the preceding distribution is 7, found by subtracting 31-24 = 7.

The researcher must decide how many classes to use and the width of each class.

To construct a frequency distribution, follow these rules.

1- *There should be between 5 and 20 classes.* A student would not be in error for having less than 5 classes or more than 20 classes; however, statisticians generally agree on these numbers.

2- *The class width should be an odd number.* This ensures that the midpoint of each class has the same place value as the data. The class midpoint X is obtained by adding the lower and upper boundaries a nd dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the midpoint of the first class is

$$\frac{24 + 30}{2} = 27 \quad \text{or} \quad \frac{23.5 + 30.5}{2} = 27$$

The midpoint is the numerical location of the center of the class. Midpoints are necessary for graphing and are used in computing the mean and standard deviation.

3- *The classes must be mutually exclusive.* Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes. Many times, frequency distributions such as

Age
10–20
21–31
32–42
43–53

4- *The classes must be continuous*. Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last class. A class with a zero frequency at either end can be omitted without affecting the distribution.

5- *The classes must be exhaustive*. There should be enough classes to accommodate all the data.

6- *The classes must be equal in width*. This avoids a distorted view of the data.

One exception occurs when a distribution is ***open-ended***—i.e., it has no specific beginning value or no specific ending value. Following are the class limits for two open-ended distributions.

Age	Minutes
10–20	Below 110
21–31	110–114
32–42	115–119
43–53	120–124
54–64	125–129
65 and above	130–134

The frequency distribution for age is open-ended for the last class, which means that anybody who is 65 years or older will be tallied in the last class. The distribution for minutes is open-ended for the first class, meaning that any minute values below 110 will be tallied in that class.

Example -3-

The following data represent the record high temperatures for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Solution

The procedure for constructing a grouped frequency distribution for numerical data follows.

STEP 1 Find the highest value and lowest value: $H = 134$ and $L = 100$.

STEP 2 Find the range: R = highest value − lowest value.

$$R = 134 - 100 = 34$$

STEP 3 Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.

STEP 4 Find the class width by dividing the range by the number of classes.

$$\text{width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9$$

Round the answer up to the nearest whole number if there is a remainder: 4.9 ≈ 5. (Rounding up is different from rounding off. A number is rounded up if there is any decimal remainder when dividing. For example, $85 \div 6 = 14.167$ and is rounded up to 15. Also, $53 \div 4 = 13.25$ and is rounded up to 14.)

STEP 5 Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. Add the width to the lowest score taken as the starting point to get the

lower limit of the next class. Keep adding until there are 7 classes, as shown.

100
105
110
115
120
125
130

STEP 6 Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

$105 - 1 = 104$

So the first class is 100 – 104.

Class limits
100–104
105–109
110–114
115–119
120–124
125–129
130–134

STEP 7 Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to the upper class limit, as shown.

Class boundaries
99.5–104.5
104.5–109.5
109.5–114.5
114.5–119.5
119.5–124.5
124.5–129.5
129.5–134.5

STEP 8 Tally the data, write the numerical values for the tallies in the frequency column, and find the cumulative frequencies.

The completed frequency distribution follows:

Class limits	Class boundaries	Tally	Frequency	Cumulative frequency
100–104	99.5–104.5	//	2	2
105–109	104.5–109.5	𝍸 ///	8	10
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18	28
115–119	114.5–119.5	𝍸 𝍸 ///	13	41
120–124	119.5–124.5	𝍸 //	7	48
125–129	124.5–129.5	/	1	49
130–134	129.5–134.5	/	1	50

The frequency distribution shows that the class 109.5–114.5 contains the largest number of temperatures (18) followed by the class

114.5-119.5 with 13 temperatures. Hence, most of the temperatures (31) fall between 109.5° and 119.5°.

The procedure for constructing a grouped frequency distribution is summarized in Procedure Table below:

Constructing a Grouped Frequency Distribution

1. Find the highest and lowest value.
2. Find the range.
3. Select the number of classes desired.
4. Find the width by dividing the range by the number of classes and rounding up.
5. Select a starting point (usually the lowest value); add the width to get the lower limits.
6. Find the upper class limits.
7. Find the boundaries.
8. Tally the data, find the frequencies, and find the cumulative frequency.

The reasons for constructing a frequency distribution follow:

1- To organize the data in a meaningful, intelligible way.

2- To enable the reader to determine the nature or shape of the distribution.

3- To facilitate computational procedures for measures of average and spread.

4- To enable the researcher to draw charts and graphs for the presentation of data.

5- To enable the reader to make comparisons among different data sets.

Exercises

1- Find class boundaries, midpoints, and widths for each class?

a) 11-15, b) 17-39, c) 293-353, d) 11.8-14.7, e) 3.13-3.93

2- The following zip codes were obtained from the respondents to a mail survey. Construct a frequency distribution for the data?

15132	15130	15132	15130
15130	15131	15134	15133
15131	15133	15133	15133
15130	15131	15132	15130
15133	15134	15133	15133

3- *At a college financial aid office, students who applied for a scholarship were classified according to their class rank: Fr = freshman, So = sophomore, Jr = Junior, Se = senior. Construct a frequency distribution for the data?*

Fr	Fr	Fr	Fr	Fr
Jr	Fr	Fr	So	Fr
Fr	So	Jr	So	Fr
So	Fr	Fr	Fr	So
Se	Jr	Jr	So	Fr
Fr	Fr	Fr	Fr	So
Se	Se	Jr	Jr	Se
So	So	So	So	So

3- *The numbers of games won by the pitchers who were inducted into the Baseball Hall of Fame through 1992 are shown below. Construct a frequency distribution for the data using 12 classes?*

373	254	237	243	308
210	266	253	201	266
239	114	224	373	286
329	236	284	247	273
198	361	416	207	243
326	251	160	360	311
215	189	344	268	363
21	270	165	240	48
150	300	207	314	197
209	210	260	327	

UNIT 3

Measures of Central Tendency

CHAPTER 3

Measures of Central Tendency

OBJECTIVES

After studying the material of this chapter, you should be able to:

- understand the meaning of the term central tendency
- know the requirements of an ideal measure of central tendency
- identify different types of measures of central tendency
- calculate various measures of central tendency such as the arithmetic mean, median, mode, geometric mean and harmonic mean for ungrouped and grouped data
- know the merits and demerits of different measures of central tendency
- appreciate the use of different measures of central tendency in solving real life problems.

3.1 INTRODUCTION

Tables and graphical representation give a general description of data. However, it is often convenient and useful to define statistical measures that describe important features of the data. Some of these statistical measures define, in some sense, the centre of a set of data and consequently are called *measures of central tendency* or *measures of central location*. The most commonly used measures of central tendency are the *mean, median, mode, geometric mean and harmonic mean.*

3.2 AVERAGE

An *average* is a single value within the range of the 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 a *measure of central value.*

Some of the important definitions of an average given by different statisticians from time to time are given below:

1. *Average is an attempt to find one single figure to describe whole of figures.* —*Clark*
2. *An average is a single value selected from a group of values to represent them in:*

some way a value which is supposed to stand for the whole group of which it is a part, as typical of all the values in the group. —A.E. Waugh

3. Averages are statistical constants which enable us to comprehend in a single effort the significance of the whole. —A.L. Bowley

4. An average is a typical value in the sense that it is sometimes employed to represent all individual values in a series or of a variable. —Ya-Lun-Chou

5. An average is sometimes called a measure of central tendency because individual values of the variable usually cluster around it. Averages are useful, however, for certain types of data in which there is little or no central tendency. —Crum and Smith

6. Statistical analysis seeks to develop concise summary figures which describe a large body of quantitative data. One of the most widely used set of summary figures is known as measures of location, which are often referred to as averages, measures of central tendency or central location. The purpose for computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp simply and quickly. The single value is the point or location around which the individuals items cluster. —Lawrence J. Kaplan

3.3 REQUISITES OF A GOOD AVERAGE

Since an average is a single value that represents all of the values in the series, it is desired that such a value satisfies the following properties:

1. ***It should be rigidly defined.*** An average should be rigidly defined so that it has one and only one interpretation. In other words, the definition should not leave anything to the discretion of the investigator.
2. ***It should be easy to understand and simple to calculate.*** An average should be readily understood and should not involve complex calculations. However, it should not be accomplished at the expense of other advantages. For example, if in the interest of greater accuracy, use of a more difficult average is required, one should prefer that.
3. ***It should be based on all the observations.*** It is logical that an average, which is supposed to stand for the whole group of values, should be based on all the observations.
4. ***It should be capable of further algebraic treatment.*** We would like to have an average that can be used for further mathematical treatment to enhance its utility. For example, if we are given the averages and sizes of a number of different groups, we should be able to compute the average of the combined group.

It should not be unduly affected by extreme observations. An ideal average is one which should not be affected unduly by the presence of one or two very small or very large observations.

It should have sampling stability. By sampling stability we mean that if we take different random samples of the same size from a given population and compute the average of each sample we should expect to get approximately the

It does not mean, however, that there can be no difference in the values of average for different samples.

3.4 ARITHMETIC MEAN

The **arithmetic mean** or simply the **mean** is the most popular and commonly used measure of central tendency. It is what we commonly call the *average*.

Definition. The **arithmetic mean** of a set of observations is equal to the sum of all the observations divided by the total number of observations.

For example, the arithmetic mean of a set of 5 observations 14, 16, 19, 25 and 21 is

$$\frac{14+16+19+25+21}{5} = \frac{95}{5} = 19$$

Calculation of Arithmetic Mean : Individual Observations

The arithmetic mean (A.M.) of a set of n observations $X_1, X_2, \ldots, X_n$ (not necessarily all distinct), denoted by $\bar{X}$, is given by

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

The summation notation $\sum X$ is an abbreviated form of the more general notation $\sum_{i=1}^{n} X_i$.

We will use the abbreviated form $\sum X$ to indicate the sum of all the numbers being considered.

EXAMPLE 1. *The following figures give the marks of 10 students in a class test:*

Marks obtained : 12 8 17 13 15 9 18 11 6 1

Find the arithmetic mean.

SOLUTION. The arithmetic mean of the marks is determined by finding the sum of all the marks and then dividing this total by 10. Thus

$$\bar{X} = \frac{\sum X}{n} = \frac{12+8+17+13+15+9+18+11+6+1}{10} = \frac{110}{10} = 11$$

Short-cut Method. It may be pointed out that if the values of X are very large, the computation of arithmetic mean can be done by using what is known as *short-cut method*. The various steps involved in the computation of arithmetic mean by short-cut method are as follows:

Step 1. Choose an arbitrary number A, called an assumed mean. Any number can be chosen as an assumed mean. However, it is usually taken as the value of X which corresponds to the middle part of the distribution. Moreover, A need not necessarily be one of the values of X.

Step 2. Find the deviations of X from A. Algebraic signs are to be taken

3.4

Step 3. The arithmetic mean is given by

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

EXAMPLE 2. *The following figures show the heights in cms of 7 students chosen at random:*

164, 159, 167, 169, 165, 170, 168.

Calculate the arithmetic mean of heights by (a) Direct method (b) Short-cut method.

SOLUTION. CALCULATION OF ARITHMETIC MEAN

1		164	-1
2		159	-6
3		167	2
4		169	4
5		165	0
6		170	5
7		168	3
$n=7$		$\sum X = 1162$	$\sum d = 7$

(a) Direct Method $\bar{X} = \frac{\sum X}{n} = \frac{1162}{7} = 166$ cm.

(b) Short-cut Method $\bar{X} = A + \frac{\sum d}{n} = 165 + \frac{7}{7} = 165 + 1 = 166$ cm.

Calculation of Arithmetic Mean - Discrete Series

In case of discrete series where the variable X takes the values $X_1, X_2, \ldots, X_n$ with respective frequencies $f_1, f_2, \ldots, f_n$, the arithmetic mean can be calculated by applying

(i) Direct Method, or

(ii) Short-cut Method.

Direct Method. According to this method, the *A.M.* is given by

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \ldots + f_nX_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N}$$

where $N = \sum f$ = total frequency.

Short-cut Method. According to this method, arithmetic mean is given by

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

where A = assumed mean, $d = X - A$ and $N = \sum f$.

EXAMPLE 3. *Calculate the arithmetic mean for the following discrete frequency distribution:*

X	20	30	40	50	60	70
f	8	12	20	10	6	4

SOLUTION.

CALCULATION OF ARITHMETIC MEAN

X	f	fX
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
	$N = \sum f = 60$	$\sum fX = 2460$

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{2460}{60} = 41$$

EXAMPLE 4. *The following data give the daily earnings (in Rs.) of 20 workers in a factory:*

Daily earnings (in Rs.)	100	140	170	200	250
No. of workers	5	2	6	4	3

Calculate the average daily earnings using: (a) Direct Method (b) Short-cut Method.

SOLUTION.

CALCULATION OF ARITHMETIC MEAN

Daily earnings	No. of workers				
X	f	fX	d	fd
100	5	500	−70	−350
140	2	280	−30	−60
170	6	1020	0	0
200	4	800	30	120
250	3	750	80	240
	$N = \sum f = 20$	$\sum fX = 3350$		$\sum fd = -50$

(a) ***Direct Method :*** According to this method, the average daily earnings is:

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{3350}{20} = \text{Rs. } 167.50$$

(b) ***Short-cut Method :*** According to this method, the average daily earnings is:

$$\bar{X} = A + \frac{\sum fd}{N} = 170 + \frac{-50}{20} = 170 - 2.5 = \text{Rs. } 167.50$$

Calculation of Arithmetic Mean-Continuous Series

In case of continuous series, the arithmetic mean may be computed by applying any

3.6

(i) Direct Method

(ii) Short-cut Method

(iii) Step-deviation Method

Direct Method. If $X_1, X_2, \ldots, X_n$ are the class marks (or mid-values) of a set of grouped data with corresponding class frequencies $f_1, f_2, \ldots, f_n$, then according to the direct method, arithmetic mean is given by

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \ldots + f_nX_n}{f_1 + f_2 + \ldots + f_n} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{N}$$

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

Short-cut Method. According to this method, arithmetic mean is given by

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

where A = assumed mean, $d = X - A$, deviation of mid-value from assumed mean, and $N = \Sigma f$ = total frequency.

Step-deviation Method. In case of grouped or continuous frequency distribution with class intervals of equal size, the calculation of arithmetic mean can further be simplified by taking

$$u = \frac{X - A}{h}$$

where X is the mid-value and h is the common size (or width) of the class intervals. According to this method, the arithmetic mean is given by

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

EXAMPLE 5. *Compute the arithmetic mean from the following frequency distribution:*

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of students	5	7	8	14	10	6

SOLUTION.

CALCULATION OF ARITHMETIC MEAN

Marks	Mid-value X	No. of students f	fX
0 – 10	5	5	25
10 – 20	15	7	105
20 – 30	25	8	200
30 – 40	35	14	490
40 – 50	45	10	450
50 – 60	55	6	330
		$N = \Sigma f = 50$	$\Sigma fX = 1625$

∴ Arithmetic Mean is : $\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{1625}{50} = 32.5$

EXAMPLE 6. *Calculate the arithmetic mean from the following frequency distribution:*

Marks	:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of students	:	10	9	25	30	16	10

SOLUTION.

CALCULATION OF ARITHMETIC MEAN

Marks	Mid-value X	No. of students f	fX	$A = 35$, $d = X - A$	$h = 10$, $u = \frac{X - A}{h}$	fd	fu
0 – 10	5	10	50	– 30	– 3	– 300	– 30
10 – 20	15	9	135	– 20	– 2	– 180	– 18
20 – 30	25	25	625	– 10	– 1	– 250	– 25
30 – 40	35	30	1050	0	0	0	0
40 – 50	45	16	720	10	1	160	16
50 – 60	55	10	550	20	2	200	20
		$N = \Sigma f = 100$	$\Sigma fX = 3130$			$\Sigma fd = -370$	$\Sigma fu = -37$

Direct Method : $\bar{X} = \frac{\Sigma fX}{N} = \frac{3130}{100} = 31.30$

Short-cut Method : $\bar{X} = A + \frac{\Sigma fd}{N} = 35 + \frac{-370}{100} = 35 - 3.70 = 31.30$

Step-deviation Method : $\bar{X} = A + \frac{\Sigma fu}{N} \times h = 35 + \frac{-37}{100} \times 10$

$= 35 - 3.70 = 31.30.$

EXAMPLE 7. *Calculate mean from the following data :*

Marks	*No. of students*	*Marks*	*No. of students*
Less than 10	4	*Less than 50*	96
Less than 20	16	*Less than 60*	112
Less than 30	40	*Less than 70*	120
Less than 40	76	*Less than 80*	125

SOLUTION. We are given 'less than' cumulative frequency distribution. We shall first convert it into an ordinary frequency distribution and then calculate mean.

3.8 CALCULATION OF ARITHMETIC MEAN

0 – 10	5	4	–4	–16
10 – 20	15	12	–3	–36
20 – 30	25	24	–2	–48
30 – 40	35	36	–1	–36
40 – 50	45	20	0	0
50 – 60	55	16	1	16
60 – 70	65	8	2	16
70 – 80	75	5	3	15
			$N = \sum f = 125$	$\sum fu = -105$

$$\bar{X} = A + \frac{\sum fu}{N} \times h = 45 + \frac{-105}{125} \times 10 = 45 - 8.4 = 36.60$$

EXAMPLE 8. *The following table gives the life-time in hours of 400 radio tubes of a certain make.*

Life-time (in hours)	No. of tubes	Life-time (in hours)	No. of tubes
Less than 300	0	Less than 800	265
Less than 400	20	Less than 900	324
Less than 500	60	Less than 1000	374
Less than 600	116	Less than 1100	392
Less than 700	194	Less than 1200	400

Calculate the mean life-time of radio tubes.

SOLUTION. The data is given in the form of a cumulative frequency distribution. To calculate the mean, we shall first convert it into an ordinary frequency distribution as shown below:

CALCULATION OF ARITHMETIC MEAN

Class Interval	Frequency (f)	Mid-value (X)	$u = \frac{X - 750}{100}$ $(A = 750, h = 100)$	fu
300 – 400	20 – 0 = 20	350	– 4	– 80
400 – 500	60 – 20 = 40	450	– 3	– 120
500 – 600	116 – 60 = 56	550	– 2	– 112
600 – 700	194 – 116 = 78	650	– 1	– 78
700 – 800	265 – 194 = 71	750	0	0
800 – 900	324 – 265 = 59	850	1	59
900 – 1000	374 – 324 = 50	950	2	100
1000 – 1100	392 – 374 = 18	1050	3	54
1100 – 1200	400 – 392 = 8	1150	4	32
	$N = \sum f = 400$			$\sum fu = -145$

$$\bar{X} = A + \frac{\sum fu}{N} \times h = 750 + \frac{-145}{400} \times 100 = 750 - 36.25 = 713.5 \text{ hours}$$

Hence the average life-time of radio tubes is 713.5 hours.

EXAMPLE 9. *A market with 168 operating firms has the following distribution of average number of employees in various income groups :*

Income groups	No. of firms	Average no. of employees
1500 – 3000	40	8
3000 – 5000	32	12
5000 – 8000	26	7.5
8000 – 12000	28	8.5
12000 – 18000	42	4

Calculate the average salary paid in the whole market.

SOLUTION. In this problem the total number of employees (*i.e.*, frequencies) working in different income groups are not given. The frequencies can be obtained by multiplying the number of firms with the average number of employees.

CALCULATION OF AVERAGE SALARY

Income groups	No. of firms	Average No. of Employees	Frequency (f)	Mid-value (X)	fX
1500 – 3000	40	8	320	2250	7,20,000
3000 – 5000	32	12	384	4000	15,36,000
5000 – 8000	26	7.5	195	6500	12,67,500
8000 – 12000	28	8.5	238	10000	23,80,000
12000 – 18000	42	4	168	15000	25,20,000
			$N = \sum f$ = 1305		$\sum fX$ = 84,23,500

The average salary is given by :

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{8423500}{1305} = \text{Rs. } 6454.79$$

REMARK. It may be remarked that even if the data is given in the form of a grouped frequency distribution with *'inclusive type'* classes, it is not necessary to adjust the classes for calculating arithmetic mean because the mid-values remain the same whether or not the adjustment is made.

EXAMPLE 10. *Given below is the distribution of marks obtained by 140 students in an examination:*

3.10 Find the mean of the distribution. —

SOLUTION. CALCULATION OF ARITHMETIC MEAN

10 – 19	14.5	7	– 4	– 28
20 – 29	24.5	15	– 3	– 45
30 – 39	34.5	18	– 2	– 36
40 – 49	44.5	25	– 1	– 25
50 – 59	54.5	30	0	0
60 – 69	64.5	20	1	20
70 – 79	74.5	16	2	32
80 – 89	84.5	7	3	21
90 – 99	94.5	2	4	8
		$N = \sum f = 140$		$\sum fu = -53$

Mean of the distribution is:

$$\bar{X} = A + \frac{\sum fu}{N} \times h = 54.5 + \frac{-53}{140} \times 10 = 54.5 - 3.79 = 50.71.$$

The mean of the following frequency distribution is 50. But the frequencies f_1.

EXAMPLE 11. *The mean of the following frequency distribution is 50. But the frequencies f_1 and f_2 in classes 20 – 40 and 60 – 80 are missing. Find the missing frequencies.*

Class :	*0 – 20*	*20 – 40*	*40 – 60*	*60 – 80*	*80 – 100*	*Total*
Frequency :	*17*	f_1	*32*	f_2	*19*	*120*

[*Delhi Univ. B. Com., 1999*]

SOLUTION. CALCULATION OF MISSING FREQUENCIES

Class	Mid-value X	Frequency f	$u = \frac{X - 50}{20}$ $(A = 50, h = 20)$	fu
0 – 20	10	17	–2	–34
20 – 40	30	f_1	–1	$-f_1$
40 – 60	50	32	0	0
60 – 80	70	f_2	1	f_2
80 – 100	90	19	2	38
		$N = \sum f = 68 + f_1 + f_2$		$\sum fu = 4 - f_1 + f_2$

We are given $N = 120 \Rightarrow 68 + f_1 + f_2 = 120 \Rightarrow f_1 + f_2 = 52$ ... (1)

Using the step-deviation method for calculating mean, we obtain

$$\bar{X} = A + \frac{\sum fu}{N} \times h$$

i.e., $$50 = 50 + \frac{4 - f_1 + f_2}{120} \times 20$$

$$\Rightarrow \quad \frac{4 - f_1 + f_2}{120} = 0 \quad \because \Rightarrow \quad 4 - f_1 + f_2 = 0 \quad \Rightarrow \quad f_1 - f_2 = 4 \qquad \ldots (2)$$

Adding (1) and (2), we get $2f_1 = 56 \Rightarrow f_1 = 28$.

Substituting $f_1 = 28$ in (1), we get $f_2 = 24$.

EXAMPLE 12. *For the following data find the missing frequency if the arithmetic mean is 33.*

Marks	*0 – 10*	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*
No. of students	*10*	*15*	*30*	*x*	*25*	*20*

SOLUTION. Let the missing frequency be x. **[C.A. Foundation, Nov. 2000]**

Marks	Mid-value X	No. of Students f	$u = \frac{X - 35}{10}$ $(A = 35, h = 10)$	fu
0 – 10	5	10	-3	-30
10 – 20	15	15	-2	-30
20 – 30	25	30	-1	-30
30 – 40	35	x	0	0
40 – 50	45	25	1	25
50 – 60	55	20	2	40
—		$N = \sum f = 100 + x$		$\sum fu = -25$

Using the step-deviation method $\bar{X} = A + \frac{\sum fu}{N} \times h$

for calculating the mean, we obtain $33 = 35 + \frac{-25}{100 + x} \times 10 \Rightarrow \frac{250}{100 + x} = 35 - 33 = 2$

$$\Rightarrow \quad 200 + 2x = 250 \quad \Rightarrow \quad 2x = 50 \quad \Rightarrow \quad x = 25$$

Thus the missing frequency is 25.

EXAMPLE 13. *For the two frequency distributions given below, the mean calculated from the first was 25.4 and that from the second was 32.5. Find the values of x and y.*

Class	*Distribution I*	*Distribution II*
10 – 20	*20*	*4*
20 – 30	*15*	*8*
30 – 40	*10*	*4*
40 – 50	*x*	*2x*
50 – 60	*y*	*y*

3.12 CALCULATION FOR MISSING FREQUENCIES

SOLUTION.

Class	x	$X = \frac{x-35}{10}$	Distribution I: f	fu	Distribution II: f'	$f'u$
10 – 20	15	–2	20	–40	4	–8
20 – 30	25	–1	15	–15	8	–8
30 – 40	35	0	10	0	4	0
40 – 50	45	1	x	x	$2x$	$2x$
50 – 60	55	2	y	$2y$	y	$2y$
			$\Sigma f =$ $45 + x + y$	$\Sigma fu =$ $x + 2y - 55$	$\Sigma f' =$ $16 + 2x + y$	$\Sigma f'u =$ $2x + 2y - 16$

Mean of the first distribution is :

$$\bar{X}_1 = 35 + \frac{\Sigma fu}{\Sigma f} \times 10$$

$$\Rightarrow \quad 25.4 = 35 + \frac{x+2y-55}{45+x+y} \times 10 \qquad (\because \bar{X}_1 = 25.4)$$

$$\Rightarrow \quad \frac{10(x+2y-55)}{45+x+y} = 25.4 - 35 = -9.6$$

$$\Rightarrow \quad 10x + 20y - 550 = -9.6(45 + x + y) = -432 - 9.6x - 9.6y$$

or $\qquad 19.6x + 29.6y - 118 = 0 \qquad \ldots (1)$

Mean of the second distribution is :

$$\bar{X}_2 = 35 + \frac{\Sigma f'u}{\Sigma f'} \times 10$$

or, $$32.5 = 35 + \frac{2x + 2y - 16}{16 + 2x + y} \times 10$$

$$\Rightarrow \quad \frac{10(2x + 2y - 16)}{16 + 2x + y} = 32.5 - 35 = -2.5$$

$$\Rightarrow \quad 20x + 20y - 160 = -2.5(16 + 2x + y) = -40 - 5x - 2.5y$$

$$\Rightarrow \quad 25x + 22.5y - 120 = 0 \qquad \ldots (2)$$

Solving eqs. (1) and (2) simultaneously for x and y, we obtain $x = 3$ and $y = 2$.

Hence the missing frequencies are : $x = 3$ and $y = 2$.

EXAMPLE 14. *In the following grouped data, X are the mid values of the class intervals and C is a constant. If the arithmetic mean of the original distribution is 35.84, find its class intervals.*

63

3.9 MEDIAN

Median is another measure of central tendency. As distinct from the arithmetic mean, which is based on all the items of the distribution, the median is what is called a *positional average*. The term 'position' refers to the place of a value in the distribution. The position of the median in a distribution is such that the number of observations above it is equal to the number of observations below it.

Definition. The median (*Md*) of a set of observations arranged in an ascending or descending order of magnitude is defined as the middle value or the arithmetic mean of two middle values according as the number of observations is odd or even respectively.

Thus median of a distribution is that value of the variable which exceeds and is exceeded by the same number of observations.

Calculation of Median - Individual Observations

For ungrouped data consisting of n observations, the calculation of median involves the following steps:

Step 1. *Arrange the given set of observations in an ascending or descending order of magnitude.*

Step 2. *The median is given by*

(i) *the value of* $\left(\frac{n+1}{2}\right)$*th observation, when n is odd*

(ii) *the arithmetic mean of the values of* $\left(\frac{n}{2}\right)$*th and* $\left(\frac{n}{2}+1\right)$*th observations, when n is even.*

EXAMPLE 40. *Find the median for the following data:*

(i) 18 12 17 22 20

(ii) 85 69 74 60 59 84

SOLUTION. (i) Arranging the data in ascending order of magnitude, we get

12 17 18 20 22

Here, n = the number of observations = 5, an odd number

median = size of $\left(\frac{n+1}{2}\right)$th observation = size of 3rd observation = 18

(ii) Arranging the data in ascending order of magnitude, we get

59 60 69 74 84 85

Here, n = the number of observations = 6, an even number

median = Arithmetic mean of two middle terms

= Arithmetic mean of 3rd and 4th terms = $\frac{1}{2}(69 + 74) = 71.5$

Calculation of Median — Discrete Series

In the case of discrete series, where the variable takes the values $X_1, X_2, \ldots, X_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ with $\Sigma f = N$, median is the size of $\left(\frac{N+1}{2}\right)$th observation. In this case, the calculation of median involves the following steps :

Step 1. Prepare the 'less than' cumulative frequency (c.f.) distribution.

Step 2. Find $\frac{N+1}{2}$

Step 3. See the c.f. just greater than or equal to $\frac{N+1}{2}$.

Step 4. The value of the variable corresponding to the c.f. obtained in Step 3 gives the required median.

EXAMPLE 41. *Calculate median from the following data:*

X	10	20	30	40	50	60	70
f	1	5	12	20	19	9	4

SOLUTION.

CALCULATION OF MEDIAN

X	f	Less than c.f.
10	1	1
20	5	6
30	12	18
40	20	38
50	19	57
60	9	66
70	4	70
	$N = \Sigma f = 70$	

We have $\frac{N+1}{2} = \frac{71}{2} = 35.5$ and the c.f. just greater than or equal to 35.5 is 38. The corresponding value of X is 40.

$\therefore$ Median = 40.

Calculation of Median — Continuous Series

In the case of continuous series, median is the size of $\frac{N}{2}$ th observation, where $N = \Sigma f$ is the total frequency. The calculation of median in this case involves the following steps:

Step 1. *Prepare the 'less than' cumulative frequency (c.f.) distribution.*

Step 2. *Find* $\frac{N}{2}$.

Step 3. *See the c.f. just greater than or equal to* $\frac{N}{2}$.

Step 4. *Find the class corresponding to the c.f. obtained in Step 3. This is called the median class.*

Step 5. *Apply the following interpolation formula for calculating the median:*

$$\text{Median} = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where l = *lower limit of the median class,*

f = *frequency of the median class,*

C = *cumulative frequency of the class preceding the median class, and*

h = *size or width of the median class.*

NOTE. It may be noted that the interpolation formula used to obtain median is based on the following assumptions:

1. *The distribution of the variable under consideration is continuous with exclusive type classes without any gap.*
2. *There is an orderly and even distribution of observations within each class.*

EXAMPLE 42. *The marks obtained by 100 students in a certain examination are given below:*

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
No. of Students	10	9	25	30	16	10

Calculate the median marks.

SOLUTION. **CALCULATION OF MEDIAN**

Marks	No. of Students (f)	C.f. (less than)	
0 – 10	10	10	
10 – 20	9	19	
20 – 30	25	44	
30 – 40	30	74	← Median class
40 – 50	16	90	
50 – 60	10	100	

$N = \Sigma f = 100$

We have $\frac{N}{2} = \frac{100}{2} = 50$. The cumulative frequency just greater than or equal to 50 is 74 and the corresponding class interval is 30 – 40. Thus the median class is 30 – 40. The median is given by the formula

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h$$

where l = lower limit of the median class = 30
f = frequency of the median class = 30
C = cumulative frequency of the class preceding the median class = 44
h = size of the median class = 10

$$\therefore \quad Md = 30 + \frac{50 - 44}{30} \times 10 = 30 + 2 = 32 \text{ marks.}$$

EXAMPLE 43. *Given below is the distribution of marks obtained by 140 students in an examination*

Marks	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
No. of Students	7	15	18	25	30	20	16	7	2

Find the median of the distribution. **[C.A. PEE-I, May 2004]**

SOLUTION. Since data is given as a grouped frequency distribution with inclusive type classes, the first step involved in computation of median is to convert the given data into a continuous frequency distribution with exclusive type classes as shown below:

CALCULATION OF MEDIAN

Class Boundaries	Frequency (f)	c.f. (less than)
9.5 – 19.5	7	7
19.5 – 29.5	15	22
29.5 – 39.5	18	40
39.5 – 49.5	25	65
49.5 – 59.5	30	95
59.5 – 69.5	20	115
69.5 – 79.5	16	131
79.5 – 89.5	7	138
89.5 – 99.5	2	140
	$N = \sum f = 140$	

We have $\frac{N}{2} = \frac{140}{2} = 70$. The cumulative frequency just greater than or equal to 70 is 95 and the corresponding class interval is 49.5 – 59.5. Thus the median class is 49.5 – 59.5. The median is given by the formula

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 49.5$, $C = 65$, $f = 30$ and $h = 10$

$$\therefore \quad Md = 49.5 + \frac{70-65}{30} \times 10 = 49.5 + 1.67 = 51.17.$$

EXAMPLE 44. *Calculate median from the following data :*

Age	No. of persons	Age	No. of persons
55 – 60	7	35 – 40	30
50 – 55	13	30 – 35	33
45 – 50	15	25 – 30	58
40 – 45	20	20 – 25	14

SOLUTION. We first arrange the series in ascending order as shown in the following table :

CALCULATION OF MEDIAN

Age	No. of persons (f)	c.f. less than
20 – 25	14	14
25 – 30	28	42
30 – 35	33	75
35 – 40	30	105
40 – 45	20	125
45 – 50	15	140
50 – 55	13	153
55 – 60	7	160
	$N = \sum f = 160$	

Since $\frac{N}{2} = \frac{160}{2} = 80$ and c.f. just greater than or equal to 80 is 105, therefore median lies in the class 35 – 40 and is given by

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 35, \quad C = 75, \quad f = 30 \quad \text{and} \quad h = 5$

$$\therefore \quad Md = 35 + \frac{80-75}{30} \times 5 = 35 + 0.83 = 35.83$$

REMARK. *Calculation of Median when Class Intervals are Unequal*

It may be remarked that even if class intervals are unequal in size, the frequencies need not be adjusted to make the class intervals equal and the same interpolation formula can be applied for calculating median as discussed before.

EXAMPLE 45. *Calculate median from the following data:*

Marks :	0 – 10	10 – 30	30 – 50	50 – 60	60 – 80	80 – 90
No. of Students :	5	15	20	10	8	2

CALCULATION OF MEDIAN

	No. of Students (f)	c.f. (less than)
0 – 10	5	5
10 – 30	15	20
30 – 50	20	40
50 – 60	10	50
60 – 80	8	58
80 – 90	2	60
	$N = \Sigma f = 60$	

Since $\frac{N}{2} = 30$ and c.f. just greater than or equal to 30 is 40, therefore median lies in the class 30 – 50. Using the following formula for median,

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 30$, $C = 20$, $f = 20$ and $h = 20$, we get

$$Md = 30 + \frac{30 - 20}{20} \times 10 = 40.$$

REMARK. ***Calculation of Median when mid-values of class intervals are given***

When the mid-values of class intervals are given, the class intervals can be obtained by first finding out the difference between two mid-values and then subtracting half of it from each mid-value to find the lower limit and adding it to each mid-value to find the upper limit of the class intervals.

EXAMPLE 46. *Compute median from the following data :*

Mid-values :	115	125	135	145	155	165	175	185	195
Frequency :	6	25	48	72	116	60	38	22	3

SOLUTION. In this problem we are given the mid-values of class intervals of a continuous frequency distribution. The difference between two mid-values is 10. Thus $\frac{10}{2} = 5$ is subtracted from each mid-value to find the lower limit and the same is added to find the upper limit of a class. The classes are thus 110 – 120, 120 – 130, ..., 190 – 200.

CALCULATION OF MEDIAN

Class Intervals	Frequency	c.f. (less than)
110 – 120	6	6
120 – 130	25	31
130 – 140	48	79
140 – 150	72	151
150 – 160	116	267
160 – 170	60	327
170 – 180	38	365
180 – 190	22	387
190 – 200	3	390 = N

Since $\frac{N}{2} = \frac{390}{2} = 195$ and c.f. just greater than or equal to 195 is 267, therefore median lies in the class 150 – 160. The median is given by the formula :

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 150,\quad C = 151,\quad f = 116$ and $h = 10$

$$\therefore \quad Md = 150 + \frac{195 - 151}{116} \times 10 = 150 + \frac{44}{116} \times 10 = 150 + 3.79 = 153.79.$$

EXAMPLE 47. *An incomplete distribution is given below:*

Class	*0 – 10*	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*	*60 – 70*	*Total*
Frequency	*10*	*20*	*?*	*40*	*?*	*25*	*15*	*170*

Find out missing frequencies if median value is 35. **[Kerala Univ. B.Com. 2004]**

SOLUTION. Let the missing frequencies for the classes 20 – 30 and 40 – 50 be f_1 and f_2 respectively. To find f_1 and f_2, we prepare the following table.

Class Interval	*f*	*c.f.*
0 – 10	10	10
10 – 20	20	30
20 – 30	f_1	$30 + f_1$
30 – 40	40	$70 + f_1$
40 – 50	f_2	$70 + f_1 + f_2$
50 – 60	25	$95 + f_1 + f_2$
60 – 70	15	$110 + f_1 + f_2$
	$N = \sum f = 110 + f_1 + f_2$	

We are given $\quad N = \sum f = 170$

$$\Rightarrow \quad 110 + f_1 + f_2 = 170 \quad \Rightarrow \quad f_1 + f_2 = 60. \qquad \text{... (1)}$$

Since median is given to be 35, which lies in the class 30 – 40, therefore 30 – 40 is the median class. Applying the formula for computing median, we get

$$35 = 30 + \frac{85 - (30 + f_1)}{40} \times 10 \quad \Rightarrow \quad 5 = \frac{55 - f_1}{4}$$

$$\Rightarrow \quad 55 - f_1 = 20 \quad \text{or} \quad f_1 = 35$$

Substituting $f_1 = 35$ in (1), we get $f_2 = 25$. Hence missing frequencies are 35 and 25 respectively.

EXAMPLE 48. *An incomplete distribution is given below:—*

Class Interval:	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*	*60 – 70*	*70 – 80*	*Total*
Frequency:	*12*	*30*	*?*	*65*	*?*	*25*	*18*	*200*

You are given that the median value is 46.

(i) Using the median formula fill up the missing frequencies.

(ii) Calculate the arithmetic mean of the completed table.

SOLUTION. (*i*) Let the missing frequencies for the classes 30 – 40 and 50 – 60 be f_1 and f_2 respectively. To find f_1 and f_2, we prepare the following table.

CALCULATION FOR MISSING FREQUENCIES

Class Interval	f	c.f.
10 – 20	12	12
20 – 30	30	42
30 – 40	f_1	$42 + f_1$
40 – 50	65	$107 + f_1$
50 – 60	f_2	$107 + f_1 + f_2$
60 – 70	25	$132 + f_1 + f_2$
70 – 80	18	$150 + f_1 + f_2$
	$N = \sum f = 150 + f_1 + f_2$	

We are given $N = \sum f = 229$

$\Rightarrow \quad 150 + f_1 + f_2 = 229 \quad \Rightarrow \quad f_1 + f_2 = 79 \quad \ldots (1)$

Since median is given to be 46, which lies in the class 40 – 50, therefore 40 – 50 is the median class. Using the median formula, we get

$$46 = 40 + \frac{114.5 - (42 + f_1)}{65} \times 10 \quad \Rightarrow \quad 6 = \frac{72.5 - f_1}{65} \times 10$$

$$\Rightarrow \quad 390 = 725 - 10 f_1 \quad \Rightarrow \quad 10 f_1 = 335$$

$\Rightarrow \quad f_1 = 33.5$ or 34 (*Since the frequency can not be in fraction*)

Substituting $f_1 = 34$ in (1), we get $f_2 = 45$. Hence missing frequencies are 34 and 45 respectively.

(*ii*) Now mean can be calculated by completing the series by putting the values of f_1 and f_2. The calculation is shown in the following table :

CALCULATION OF ARITHMATIC MEAN

Class Interval	Mid-Value X	Frequency f	$u = \frac{X - 45}{10}$ $(A = 45, h = 10)$	fu
10 – 20	15	12	– 3	– 36
20 – 30	25	30	– 2	– 60
30 – 40	35	34	– 1	– 34
40 – 50	45	65	0	0
50 – 60	55	45	1	45
60 – 70	65	25	2	50
70 – 80	75	18	2	54
		$N = 229$		$\sum fu = 19$

$$A.M. = A + \frac{\sum fu}{N} \times h = 45 + \frac{19}{229} \times 10 = 45 + 0.83 = 45.83.$$

REMARK. In a distribution with open-end classes the value of median can be calculated without making assumptions regarding the size of the class interval of the open-end classes.

EXAMPLE 49. *The following are the marks obtained by 80 students in a class test. Find the median marks :*

Marks	*below 10*	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 60*	*60 – 80*	*above 80*
No. of Students :	*8*	*10*	*8*	*16*	*23*	*10*	*5*

SOLUTION.

CALCULATION OF MEDIAN

Marks	No. of Students (f)	
below 10	8	8
10 – 20	10	18
20 – 30	8	26
30 – 40	16	42
40 – 60	23	65
60 – 80	10	75
above 80	5	80
	$N = 80$	

$$\text{Median} = \text{size of } \frac{N}{2} \textit{th item} = \text{size of } 40\textit{th item}$$

∴ Median lies in the class 30 – 40 and is given by

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h = 30 + \frac{40 - 26}{16} \times 10$$

$$= 30 + 8.75 = 38.75$$

EXAMPLE 50. *You are given below a certain statistical distribution :*

Value	*Frequency*
Less than 100	40
100 – 200	89
200 – 300	148
300 – 400	64
400 and above	39
Total	380

Calculate the must suitable average giving reasons for your choice.

SOLUTION. Since the distribution has open-end classes, median would be the most suitable choice for computing the average.

CALCULATION OF MEDIAN

	Frequency (f)	c.f. (less than)
Less than 100	40	40
100 – 200	89	129
200 – 300	148	277
300 – 400	64	341
400 and above	39	380 = N

Since $\frac{N}{2} = \frac{380}{2} = 190$ and *c.f.* just greater than or equal to 190 is 277, therefore median lies in the class 200 – 300. Applying the following formula for median :

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 200$, $C = 129$, $f = 148$ and $h = 100$, we get

$$Md = 200 + \frac{190 - 129}{148} \times 100 = 200 + 41.2 = 241.2.$$

Calculation of Median in Cumulative Series

If the data is given in the form of a cumulative frequency distribution, it has to be first arranged in an ordinary frequency distribution in order to find out the frequency of the median class which is needed in the calculation of median. Once it is done, the rest of the procedure is same as in any other continuous series.

EXAMPLE 51. *Following is the distribution of marks obtained by 125 students in a Business Statistics paper :*

Marks (less than) :	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*
No. of Students :	*4*	*16*	*40*	*76*	*96*	*112*	*120*	*125*

Calculate the median marks.

SOLUTION. Since the data is given in the form of cumulative frequency distribution, it has to be arranged in a frequency distribution as shown in the following table :

CALCULATION OF MEDIAN

Marks	Frequency	c.f.
0 – 10	4	4
10 – 20	16 – 4 = 12	16
20 – 30	40 – 16 = 24	40
30 – 40	76 – 40 = 36	76
40 – 50	96 – 76 = 20	96
50 – 60	112 – 96 = 16	112
60 – 70	120 – 112 = 8	120
70 – 80	125 – 120 = 5	125 = N

Since $\frac{N}{2} = \frac{125}{2} = 62.5$ and *c.f.* just greater than or equal to 62.5 is 76, therefore median lies in the class 30 – 40 and is given by

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 30$, $C = 40$, $f = 36$ and $h = 10$

$$\therefore \quad Md = 30 + \frac{62.5 - 40}{36} \times 10 = 30 + 62.5 = 36.25.$$

EXAMPLE 52. *Following is the distribution of marks obtained by 65 students in statistics paper:*

Marks (more than):	20	30	40	50	60	70
No. of Students:	65	63	40	40	18	7

Calculate the median marks.

SOLUTION. Since the data is given in the form of cumulative frequency distribution, it has to be arranged in a frequency distribution as shown in the following table:

CALCULATION OF MEDIAN

Marks	*Frequency (f)*	*c.f. (less than)*
20 – 30	65 – 63 = 2	2
30 – 40	63 – 40 = 23	25
40 – 50	40 – 40 = 0	25
50 – 60	40 – 18 = 22	47 ← *Median class*
60 – 70	18 – 7 = 11	58
70 and above	7	65 = *N*

$$\text{Median} = \text{size of } \frac{N}{2} \text{th item} = \text{size of } 32.5 \text{ item}$$

Median lies in the class 50 – 60 and is given by

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h = 50 + \frac{32.5 - 25}{22} \times 10 = 50 + 3.41 = 53.41.$$

53. Ten percent of the workers in a firm employing a total of 1000 workers earn les

where $l = 99.995$, $C = 300$, $f = 300$ and $h = 50$, we get

$$\therefore \quad Md = 99.995 + \frac{500 - 300}{300} \times 50 = 99.995 + 33.333 = \text{Rs. } 133.328.$$

3.10 MERITS AND DEMERITS OF MEDIAN

Merits. Median possesses the following merits :

1. *It is rigidly defined.*
2. *It is easy to calculate and simple to understand.*
3. *It can be computed while dealing with a distribution with open end classes.*
4. *Being a positional average, it is not much affected by extreme observations. For example, the median of 10, 15, 20, 25 and 130 is 20 whereas the mean is 40. Hence very often when extreme values are present in a set of observations, the median is more satisfactory measure of central tendency than the mean.*
5. *It is the most appropriate average to be used while dealing with qualitative data.*
6. *It can sometimes be located by inspection and can also be determined graphically.*

Demerits. Median has the following limitations :

1. *Median, being a positional average, is not based on each and every item of the distribution.*
2. *It is not suitable for further mathematical treatment. For example, it is not possible to find the combined median of two or more groups.*
3. *It can not be determined exactly for an ungrouped data consisting of an even number of observation. It is determined approximately as the mid-point of two middle observations.*

In comparison to arithmetic mean, it is much affected by sampling fluctuations.

For calculating median, it is necessary to arrange the data in order of magnitude.

3.13 MODE

The mode (*Mo*) of a set of observations is that value which appears most frequently or with the greatest frequency. If two or more values appear with the same greatest frequency, each is a mode. If no value is repeated, there is no mode.

EXAMPLE 69. *(i) The mode of the numbers*

1 2 2 3 5 6 6 6

is 6, since it appears three times and no other value appears more than twice.

(ii) The set of numbers

1 2 2 2 4 5 6 6 6 8

has two modes, viz., 2 and 6, since each appears with the same greatest frequency. Such a

(iii) The set of numbers

1 2 3 5 7 9

has no mode, since none of the numbers is repeated.

Calculation of Mode - Ungrouped Data

For determining mode in the case of ungrouped data, count the number of times the various values repeat themselves and the value occurring the maximum number of times is the mode.

For example, the mode of the set of numbers

3 4 4 5 6 7 7 7 9

is 7, since it appears three times and no other value appears more than twice.

Calculation of Mode - Discrete Series

In discrete frequency distribution, mode can be determined just by inspection. It is the value of the variable corresponding to the maximum frequency. However, this method is applicable only if the distribution is 'unimodal', *i.e.*, if it has only one mode. For example, consider the following distribution:

X	1	2	3	4	5	6	7
f	1	4	12	7	2	3	1

Since the value of X corresponding to the maximum frequency is 3, the mode is 3.

NOTE. While determining mode by inspection in the case of discrete frequency distribution, an error of judgment is possible when the difference between the greatest frequency and the frequency preceding it or succeeding it is very small and the values are heavily concentrated on either side. In such cases, it is desirable to locate the mode by what is called the *method of grouping*.

Method of Grouping

The method of grouping involves preparing a grouping table. A grouping table has six columns. In column (1), we write down the original frequencies. The greatest frequency in this column is put in a circle or marked by bold type. In column (2), frequencies are grouped in two's. In column (3), we leave the first frequency and then group the remaining frequencies in two's. In column (4), frequencies are grouped in three's. In column (5), we leave the first frequency and then group the remaining frequencies in three's. In column (6), we leave the first two frequencies and then group the remaining frequencies in three's. In each of these columns, the highest total is put in a circle or marked by bold type.

After completing the grouping table, we prepare an analysis table. In the analysis table, column numbers are put on the left-hand side and the various probable values of mode are put on the right-hand side. The values against which frequencies are highest are entered by means of a bar in the relevant box corresponding to the values they represent. The value which is repeated the maximum number of times represents the mode.

The method of preparing grouping table and analysis table is best illustrated in the following example.

EXAMPLE 70. *Calculate mode from the following data.*

Height in inches	56	58	59	60	61	62	63	64	66	68
No. of Persons	3	7	6	9	20	22	24	5	3	1

SOLUTION. By inspection one is likely to say that the mode is 63 since it occurs the maximum number of times, *i.e.*, 24. However, the difference between the maximum frequency and the frequency preceding it is very small, we prepare a grouping table and an analysis table as shown below :

GROUPING TABLE

56	3				
58	7	10			
59			13		
	6			16	
60	9	15			22
61			29		
	20			51	
62	22	42			35
63			46		
	24				66
64	5	29			
			8	32	
66	3				51
		4			9
68	1				

ANALYSIS TABLE

Col. No.	56	58	59	60	61	62	63	64	66	68
1							1			
2					1	1				
3						1	1			
4				1	1	1				
5					1	1	1			
6						1	1	1		
Total			—	1	3	5	4	1		

Since the value 62 has occurred the maximum number of times, i.e., 5, the mode is

f_2 = frequency of the class succeeding the modal class = 15
h = size of the modal class = 8

$$\therefore \quad \text{Mode} = 24 + \frac{24-16}{48-16-15} \times 8 = 24 + \frac{8}{17} \times 8 = 24 + 3.76 = 27.76$$

EXAMPLE 72. *Given below is the distribution of weights of a group of 60 students in a class:*

Weights (in kg) :	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64
No. of Students :	3	5	12	18	14	6	2

Find the mode of the distribution.

SOLUTION. Since the formula for mode requires the distribution to be continuous with 'exclusive type' classes, we first convert the classes into class boundaries as shown in the following table:

CALCULATION OF MODE

Weights (in kg)	Class Boundaries	Frequency (f)
30 – 34	29.5 – 34.5	3
35 – 39	34.5 – 39.5	5
40 – 44	39.5 – 44.5	12
45 – 49	44.5 – 49.5	18
50 – 54	49.5 – 54.5	14
55 – 59	54.5 – 59.5	6
60 – 64	59.5 – 64.5	2

Since the maximum frequency is 18, therefore the corresponding class 44.5 – 49.5 is the modal class. Applying the modal formula:

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

where, $l = 44.5$, $f_1 = 18$, $f_0 = 12$, $f_2 = 14$ and $h = 5$, we get

$$\text{Mode} = 44.5 + \frac{18-12}{36-12-14} \times 5 = 44.5 + \frac{6}{10} \times 5 = 44.5 + 3 = 47.5.$$

Mode when Class intervals are unequal. The formula for calculating the value of mode given above is applicable only where there are equal class intervals. If the class intervals are unequal, then we must make them equal before we start computing the value of mode. The class interval should be made equal and frequencies adjusted on the assumption that they are equally distributed throughout the class.

EXAMPLE 73. *Calculate mode from the following data :*

Marks :	0 – 10	10 – 20	20 – 40	40 – 50	50 – 70
No. of Students :	2	7	18	15	8

CALCULATION OF MODE

0 – 10	2
10 – 20	1
20 – 30	9
30 – 40	9
40 – 50	15
50 – 60	4
60 – 70	4

By inspection the class 40 – 50 is the modal class. Applying the mode formula:

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

where, $l = 40$, $f_1 = 15$, $f_0 = 9$, $f_2 = 4$ and $h = 10$, we get

$$\text{Mode} = 40 + \frac{15-9}{30-9-4} \times 10 = 40 + \frac{6}{17} \times 10 = 40 + \frac{60}{17} = 40 + 3.53 = 43.53$$

NOTE. If we had not made any adjustment, the value of mode would have been

$$\text{Mode} = 20 + \frac{18-7}{36-7-15} \times 20 = 20 + \frac{11}{14} \times 20 = 20 + \frac{110}{14} = 20 + 15.71 = 35.71,$$

which is not possible, since mode cannot be less than 40.

EXAMPLE 74. *From the following wage distribution find the missing frequencies if median and mode are given as Rs. 33.5 and Rs. 34 respectively.*

Wages (Rs.) :	*0 – 10*	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*	*60 – 70*	*Total*
Frequencies :	*4*	*16*	*?*	*?*	*?*	*6*	*4*	*230*

[Delhi Univ. B. Com. (H) 1993]

SOLUTION. Let the missing frequencies for the classes 20 – 30, 30 – 40 and 40 – 50 be f_1, f_2 and f_3 respectively. To find f_1, f_2 and f_3, we prepare the following table :

CALCULATION FOR MISSING FREQUENCIES

0 – 20	4	4
10 – 20	16	20
20 – 30	f_1	$20 + f_1$
30 – 40	f_2	$20 + f_1 + f_2$
40 – 50	f_3	$20 + f_1 + f_2 + f_3$
50 – 60	6	$26 + f_1 + f_2 + f_3$
60 – 70	4	$30 + f_1 + f_2 + f_3$

$$N = \Sigma f = 30 + f_1 + f_2 + f_3$$

We are given $N = \Sigma f = 230$

$\Rightarrow \quad 30 + f_1 + f_2 + f_3 = 230 \quad \Rightarrow \quad f_1 + f_2 + f_3 = 200 \quad \ldots (1)$

Since median is given to be 33.5, which lies in the class 30 – 40, therefore 30 – 40 is the median class. Applying the formula for computing median, we get

$$33.5 = 30 + \frac{115 - (20 + f_1)}{f_2} \times 10$$

$$\Rightarrow \quad 33.5 - 30 = \frac{95 - f_1}{f_2} \times 10 \quad \Rightarrow \quad 3.5 f_2 = 950 - 10 f_1$$

$$\Rightarrow \quad 10 f_1 + 3.5 f_2 = 950 \quad \ldots (2)$$

Further, mode is given to be 34, which lies in the class 30 – 40, therefore 30 – 40 is the modal class. Applying the formula for computing mode, we get

$$34 = 30 + \frac{f_2 - f_1}{2f_2 - f_1 - f_3} \times 10$$

$$\Rightarrow \quad 34 - 30 = \frac{f_2 - f_1}{2f_2 - f_1 - f_3} \times 10$$

$$\Rightarrow \quad 4(2f_2 - f_1 - f_3) = 10 f_2 - 10 f_1 \quad \Rightarrow \quad 6f_1 - 2f_2 - 4f_3 = 0 \quad \ldots (3)$$

Multiplying Eq. (1) by 4 and adding to Eq. (3), we get

$$10 f_1 + 2 f_2 = 800 \quad \ldots (4)$$

Subtracting Eq. (4) from Eq. (2), we get

$$1.5 f_2 = 150 \quad \Rightarrow \quad f_2 = 100$$

Substituting $f_2 = 100$ in Eq. (4), we get

$$10 f_1 + 200 = 800 \quad \Rightarrow \quad 10 f_1 = 600 \quad \Rightarrow \quad f_1 = 60$$

Now, substituting $f_1 = 60$ and $f_2 = 100$ in Eq. (1), we get

$$60 + 100 + f_3 = 200 \quad \Rightarrow \quad f_3 = 40$$

Hence missing frequencies are $f_1 = 60$, $f_2 = 100$ and $f_3 = 40$.

EXAMPLE 75. *Calculate the value of mode by the usual formula (after regrouping if necessary).*

Class-Interval	Frequency	Class-Interval	Frequency
10 – 20	4	60 – 70	22
20 – 30	6	70 – 80	24
30 – 40	5	80 – 90	6
40 – 50	10	90 – 100	2
50 – 60	20	100 – 110	1

SOLUTION. By preparing a grouping table and an analysis table, it may be checked that the modal class is 60 – 70 even though the class 70 – 80 has a higher frequency. Applying the

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

where $l = 60$, $f_1 = 22$, $f_0 = 20$, $f_2 = 24$ and $h = 10$, we get

$$\text{Mode} = 60 + \frac{22 - 20}{44 - 20 - 24} \times 10 = 60 + \left(\frac{2}{0} \times 10\right),$$

which is meaningless because $2f_1 - f_0 - f_2 = 0$. In such cases where $2f_1 - f_0 - f_2 = 0$, there are two alternative methods, *viz*, either to use absolute values or to regroup the series.

If the first alternative is used,

$$\text{Mode} = l + \frac{f_1 - f_0}{|f_1 - f_0| + |f_1 - f_2|} \times h = 60 + \frac{2}{|22 - 20| + |22 - 24|} \times 10$$

$$= 6 + \frac{2}{2+2} \times 10 = 65$$

Alternatively, we can regroup the series as follows :

Class Interval :	10 – 30	30 – 50	50 – 70	70 – 90	90 – 110
Frequency :	10	15	42	30	3

Now, by inspection, the modal class is 50 – 70.

$$\therefore \quad \text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h,$$

where $l = 50$, $f_1 = 42$, $f_0 = 15$, $f_2 = 30$ and $h = 20$

$$\therefore \quad \text{Mode} = 50 + \frac{42 - 15}{84 - 15 - 30} \times 20 = 50 + \frac{27}{39} \times 20 = 50 + 13.85 = 63.85$$

EXAMPLE 76. *The distribution of age of patients turned out in a hospital on January 1, 2005 was as under :*

	No. of Patients
more than 10	*148*
more than 20	*124*
more than 30	*109*
more than 40	*71*
more than 50	*30*
more than 60	*16*
more than 70 and upto 80	*01*

Find the median age and modal age of the patients. **[C.A. PEE-I, May 2005]**

SOLUTION. Since the data is given in the form of a cumulative frequency distribution, it has to be first arranged in a frequency distribution as shown in the following table :

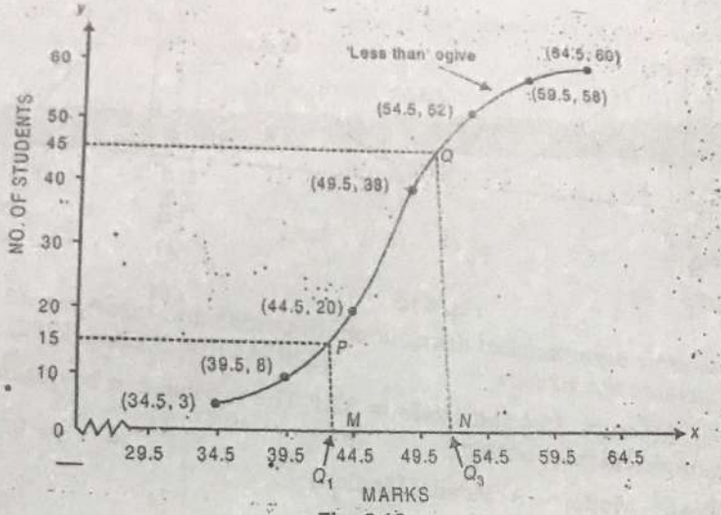


Fig. 3.10

Marks limits of the middle 50% students = $Q_1 - Q_3$

$$Q_1 = \text{size of } \frac{N}{4} \text{ th item} = 15\text{th item}$$

$$Q_3 = \text{size of } \frac{3N}{4} \text{ th item} = 45\text{th item}$$

To find Q_1, locate 15 on the y-axis and from it draw a horizontal line to meet the ogive, say at P. From P draw a perpendicular on the x-axis. The point M where it meets the x-axis gives the value of Q_1. It is clear from the graph that the value of Q_1 is 42.5. Similarly, it can be seen that the value of Q_3 is 52.

Thus, the marks limits of the middle 50% students = 42.5 – 52.

Empirical Relationship among Mean, Median and Mode

For a symmetrical distribution, the mean, mode and median all coincide (see Fig. 3.11).

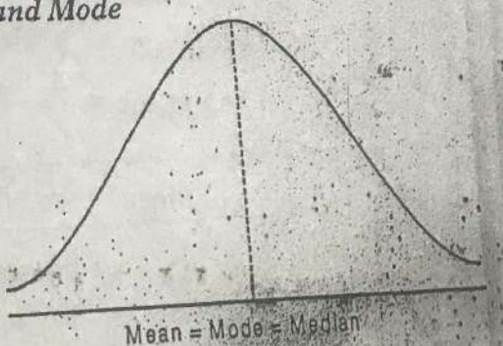


Fig. 3.11

However, for a moderately skewed or asymmetrical distribution, the mean tends to lie on the same side of the mode as the longer side and the median lies in between them (see Fig. 3.12) and they obey the following *Empirical relation*, given by Prof. Karl Pearson.

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

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

or

Knowing any two values out of the three, we can compute the third from these relationships.

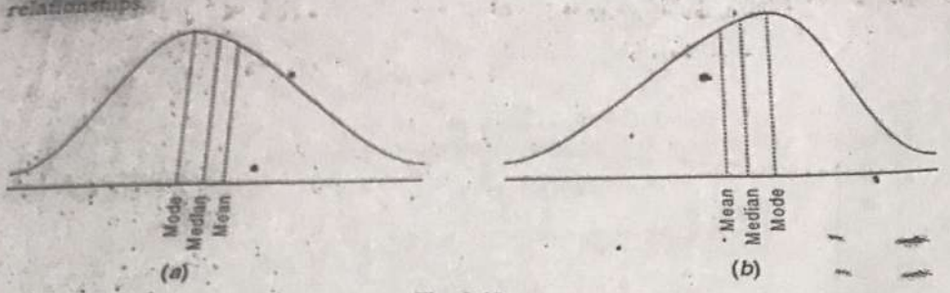


Fig. 3.12

EXAMPLE 85. *In a moderately asymmetrical distribution, the mode and mean are 32.1 and 35.4 respectively. Calculate the median.* **[Delhi Univ. B.Com. 1982, 2005]**

SOLUTION. We are given : Mean = 35.4 and Mode = 32.1. The median can be obtained by using the following empirical relationship:

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

$$\Rightarrow \quad 35.4 - 32.1 = 3\,(35.4 - \text{Median})$$

$$\Rightarrow \quad 3.3 = 106.2 - 3\,\text{Median}$$

or $$3\,\text{Median} = 106.2 - 3.3 = 102.9$$

$$\Rightarrow \quad \text{Median} = \frac{102.9}{3} = 34.3.$$

EXAMPLE 86. *In a moderately skewed distribution, the mode and median are 20 and 24 respectively. Locate the value of mean.* **[Delhi Univ. B.Com. 1979]**

SOLUTION. We are given : Mode = 20 and Median = 24. Substituting these values in the following empirical relationship among mean, median and mode :

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

we get $$\text{Mean} - 20 = 3\,(\text{Mean} - 24) = 3\,\text{Mean} - 72$$

$$\Rightarrow \quad 2\,\text{Mean} = 72 - 20 = 52$$

$$\Rightarrow \quad \text{Mean} = \frac{52}{2} = 26.$$

EXAMPLE 87. *From the following data, calculate the value of mode.*

Class Interval :	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180
Frequency	4	6	20	32	33	17	8	2

SOLUTION. By inspection it is difficult to say which is the modal class. Hence we prepare a grouping table and an analysis table as follows :

GROUPING TABLE

Marks of pupils	Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
100 – 110	4					
		10				
110 – 120	6					
			26			
120 – 130	20					
		52				
130 – 140	32					
			65			
140 – 150	33					
		50				
150 – 160	17					
			25			
160 – 170	8					
		10				
170 – 180	2					

ANALYSIS TABLE

Col. No.	120 – 130			
1				1
2	1		1	
3			1	1
4			1	1
5	1		1	1
6	1		1	1
Total	3		5	5

From the analysis table, we find that this is a bi-modal distribution. Hence we shall apply the following empirical relationship to compute the value of mode :

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

We shall first calculate the mean. For this we prepare the following table :

COMPUTATION OF MEAN

100 – 110	105	4	–4	–16
110 – 120	115	6	–3	–18
120 – 130	125	20	–2	–40
130 – 140	135	32	–1	–32
140 – 150	145	33	0	0
150 – 160	155	17	1	17
160 – 170	165	8	2	16
170 – 180	175	2	3	6
		$N = \sum f = 122$		$\sum fu = -67$

Calculation of Mean : The mean is given by the formula

$$\text{Mean} = A + \frac{\sum fu}{N} \times h = 145 + \left(\frac{-67}{122}\right) \times 10 = 145 - 5.49 = 139.51$$

Calculation of Median: To find median, we prepare the following table:

COMPUTATION OF MEDIAN

100 – 110	4	4
110 – 120	6	10
120 – 130	20	30
130 – 140	32	62
140 – 150	33	95
150 – 160	17	112
160 – 170	8	120
170 – 180	2	122
	$N = \sum f = 122$	

We have $\frac{N}{2} = \frac{122}{2} = 61$. The *c.f.* just greater than or equal to 61 is 62 and the corresponding class interval is 130 – 140. Thus the median class is 130 – 140. The median is given by the formula :

$$\text{Median} = l + \frac{\frac{N}{2} - C}{f} \times h,$$

where $l = 130$, $C = 30$, $f = 32$, and $h = 10$. Substituting these values, we obtain

$$\text{Median} = 130 + \frac{61-30}{32} \times 10 = 130 + 9.68 = 139.68.$$

Calculation of Mode: Mode = 3 Median – 2 Mean

$= 3 \times 139.68 - 2 \times 139.51 = 419.04 - 279.02 = 140.02.$

3.15 MERITS AND DEMERITS OF MODE

Merits. Mode possesses the following merits:

1. *It is simple to understand and easy to calculate.*
2. *In some cases it can be located merely by inspection.*
3. *It can be determined graphically from a histogram.*
4. *It is not at all affected by extreme observations and can be calculated even if extreme values are not known.*
5. *It can be conveniently determined for distribution with open end classes.*

Demerits. Mode has the following drawbacks:

1. *It is not rigidly defined.*
2. *It is not based on all the observations.*
3. *It is not suitable for further mathematical treatment.*
4. *As compared to mean, mode is affected to a greater extent by the fluctuations of sampling.*
5. *The value of mode cannot always be determined. In some cases, we may have a bi-modal distribution.*

EXERCISE 3.3

1. Define mode. What is modal class?
2. When is mode said to be ill-defined?
3. Point out the merits and demerits of mode as a measure of central tendency.
4. Give the formula for computing mode in a continuous

UNIT 4
Measures of Dispersion

CHAPTER 4

Measures of Dispersion

OBJECTIVES

After studying the material in this chapter you should be able to:

- Understand the meaning of the term dispersion.
- Know the qualities of an ideal measure of dispersion.
- Differentiate between relative and absolute measures of dispersion.
- Identify different types of measures of dispersion.
- Calculate various measures of dispersion such as the range, quartile deviation, mean deviation, standard deviation and their values.
- Know the merits and demerits of different measures of dispersion.
- Appreciate properties of standard deviation.
- Understand the concept of coefficient of variation.
- Describe the graphic method of studying dispersion (Lorenz curve).

4.1 INTRODUCTION

The various measures of central tendency discussed in the previous chapter do not by themselves give an adequate description of the data. Regardless of the measure that may be used, it does not tell us whether the data are grouped closely together or spread over a large range of values. It is quite possible to have two or more sets of observations with the same mean but still differ considerably in the variability of their measurements about the average. For example, consider the following three sets of observations, each containing 9 items :

										Total	*Mean*
Set A	:	10	10	10	10	10	10	10	10	10	90	10
Set B	:	6	7	8	9	10	11	12	13	14	90	10
Set C	:	2	4	6	8	10	12	14	16	18	90	10

All the three sets of data have the same mean but otherwise they are quite different. Each set has a sum of 90, but in the first set this sum is dispersed among all the values. In fact, none of the observations in set *A* deviates from the mean and we say that there is no

4.2

dispersion at all in set A. However, comparing the observations in sets B and C, it is quite obvious that set B is more uniform than set C. We say that the variability or the dispersion of the observations from the average is less for set B than for set C. Thus we need some measure of dispersion of a set of numbers. The dispersion of a data measures the degree to which numerical data tend to spread about an average value.

The measures of central tendency are, therefore, inadequate to describe the data completely. They must be supported and supplemented by some other measures.

4.2 DISPERSION

In the following we shall give some important definitions of dispersion.

1. *Dispersion is the measure of the variations of the items.* — **A.L. Bowley**
2. *Dispersion is the measure of the extent to which the individual items vary.* —**L.R. Connor**
3. *The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data.* — **Spiegel**
4. *Dispersion or spread is the degree of the scatter or variation of the variables about a central value.* —**Brooks and Dick**
5. *The term dispersion is used to indicate the facts that within a given group, the items differ from one another in size or in other words, there is a lack of uniformity in their sizes.* —**W.I. King**

It is clear from these definitions that dispersion (also called scatter, spread or variation) measures the extent to which the individual items vary from a central value.

4.3 MEASURES OF DISPERSION

The degree to which numerical data tend to spread about an average value is called the **variation** or **dispersion** of the data. There are various measures of dispersion, the most important being the *range*, the *quartile deviation*, the *mean deviation* and the *standard deviation*. The measures of dispersion can be classified as follows:

(i) *Absolute measures of dispersion.*

(ii) *Relative measures of dispersion.*

Absolute Measures of Dispersion. The measures of dispersion which are expressed in the same statistical unit in which the original data are given are termed as **absolute measures of dispersion.** They have a serious drawback because the dispersion obtained from these measures cannot be used to compare the variability of two distributions expressed in different units.

Relative Measures of Dispersion. These measures are pure numbers independent of the units of measurement and can be used to compare the variability of two distributions expressed in different units.

4.4 PROPERTIES OF A GOOD MEASURE OF DISPERSION

A good measure of dispersion should possess the following properties:

1. *It should be simple to understand.*
2. *It should be easy to calculate.*
3. *It should be rigidly defined.*
4. *It should be based on all the observations.*
5. *It should be amenable to further mathematical treatment.*
6. *It should have sampling stability.*
7. *It should not be unduly affected by extreme observations.*

4.5 RANGE

Definition. The range of a set of data is defined as the difference, between the largest and the smallest value in the set. Symbolically,

$$\text{Range} = L - S$$

where L = largest value and S = smallest value.

In case of a grouped frequency distribution, range is defined as the difference between the upper limit of the highest class and the lower limit of the smallest class.

Coefficient of Range. The range is an absolute measure of dispersion and is expressed in the unit of measurement of values of a distribution. Hence it cannot be used to compare two distributions expressed in different units. To overcome this difficulty we need a relative measure which is independent of the units of measurement. This relative measure, called the **coefficient of range**, is defined as follows:

$$\text{Coefficient of range} = \frac{\text{Range}}{\text{Sum of the largest and the lowest values}} = \frac{L-S}{L+S}$$

EXAMPLE 1. *Find the range and the coefficient of range for the following observations:*
65, 70, 82, 59, 81, 76, 57, 60, 55 and 50. [*CA PEE-I, Nov. 2003*]

SOLUTION. Range $= L - S$

Here, $L = 82$ and $S = 50$

$\therefore$ Range $= 82 - 50 = 32$

$$\text{Coefficient of range} = \frac{L-S}{L+S} = \frac{82-50}{82+50} = \frac{32}{132} = 0.24$$

EXAMPLE 2. *Calculate range and coefficient of range from the following data:*

Marks :	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90
No. of Students :	2	6	12	18	25	20	10	7

SOLUTION. Range $= L - S = 90 - 10 = 80$

$L - S$ 90 – 10 80

NOTE. It should be noted that in the calculation of 'Range' only the values of the variable are taken into account and the frequencies are completely ignored.

4.6 MERITS AND DEMERITS OF RANGE

Merits. The range possesses the following merits:

1. *It is simple to understand and easy to calculate.*
2. *It requires minimum time to calculate the value of range.*

Demerits. The range has the following drawbacks:

1. *It is not based on all the observations.*
2. *Range is a poor measure of variation. It considers only the extreme values and tells us nothing about the distribution of numbers in between. Consider, for example, the following two sets of data, both with a range of 12:*

A	:	8	9	10	11	13	14	15	17	20
B	:	8	12	12	12	13	13	13	14	20

In set A the mean and median are both 13, but the numbers vary over the entire interval from 8 to 20. In set B the mean and median are also 13, but most of the values are closer to the center of the data.

3. *It is very much affected by fluctuations of sampling. Its value varies widely from sample to sample.*
4. *It cannot be calculated for grouped frequency distribution with open-end classes.*
5. *It is not suitable for further mathematical treatment.*

Uses of Range. Although the range is a poor measure of variation, it does have some useful applications in the following areas:

1. ***Industry.*** It is used in industry where the range for measurements on items coming off an assembly line might be specified in advance. As long as all measurements fall within the specified range, the process is said to be in control.
2. ***Weather Forecasts.*** It is used by the meteorological department for weather forecasts. The range in determining the difference between the minimum temperature and the maximum temperature is of great importance to the general public because they come to know the limits within which the temperature is likely to vary on a particular day.
3. ***Stock Exchange.*** It is useful in studying the fluctuations in the share prices.
4. ***Day-to-day Living.*** The range is a most widely used measure of dispersion in our day-to-day life. For example, questions such as "What are the monthly wages of workers in the factory?" "What is the daily sales in a departmental store?" "How much do you travel in a month by your own car" are all answered in the form of range.

Frequency	10	15	20	25	10	12	8

ANSWERS

6. Rs. 120 7. 47, 0.43 8. 6, 0.054 9. 20, 0.07
10. 0.714 11. 0.35 12. 15, 0.11 13. 160
14. 1087.5, 0.062 15. 35, 0.042 16. 3.1b, 0.0236 17. 6, 0.25
18. 0.333 19. 10.71, 0.287 20. (i) 38.75; (ii) 19.375; (iii) 0.3647
21. 41.25, 0.55 22. 2.85, 0.164 23. 1651.754, 0.2948 24. 12.5, 0.2475

4.9 MEAN DEVIATION OR AVERAGE DEVIATION

As discussed earlier, the two measures of dispersion, *viz.*, range and quartile deviation are not based on all the observations. Moreover, they do not show any scatterness around an average and thus completely ignore the composition of the series. However, if we wish to measure variation in the sense of showing the scatterness around an average, we must include the deviations of each and every item from an average. *Mean deviation* or he *average deviation* helps us in achieving this goal. As the name suggests, this measure f dispersion is obtained on taking the average (arithmetic mean) of the deviations of the iven values from a measure of central tendency. According to Clark and Schkade:

Average deviation is the average amount of scatter of the items in a distribution from ther the mean or the median, ignoring the signs of the deviations. The average that is ken of the scatter is an arithmetic mean which accounts for the fact that this measure is ten called the mean deviation.".

mputation of Mean Deviation – Individual Observations

r a given set of n observations $X_1, X_2, \ldots, X_n$, the mean deviation (M.D.) about an rage, say A, is given by

$$\text{Mean Deviation (about an average } A) = \frac{\sum|X-A|}{n} = \frac{\sum|D|}{n}$$

re $|D| = |X-A|$ (read as mod $(X-A)$ is the modulus value (or absolute value) of the ation of X from A, ignoring $\pm$ signs.

mean deviation is computed about mean or median. However, there is an advantage of computing mean deviation about median because *the sum of the deviations of items from median is least when signs are ignored.* Nevertheless, in practice, the mean is more frequently used in computing the average deviation and this is the reason why it is more commonly referred to as mean deviation.

Procedure for Computing the Mean Deviation

We now outline the procedure for computing the mean deviation:

Step 1. *Calculate the average A about which mean deviation is to be computed, by the methods discussed earlier.*

Step 2. *Find the deviation of each observation X from A and denote it by D. That is, find* $D = X - A$.

Step 3. *Find the absolute value of the deviation of each observation from A ignoring ± signs and denote it by* $|D|$.

Step 4. *Find the sum of all absolute deviations obtained in Step 3 to get* $\sum |D|$.

Step 5. *Divide the sum obtained in Step 4 by the number of observations to get the required mean deviation about the average A.*

Computation of Mean Deviation - Discrete Series

In case of discrete series where the variable X takes the values $X_1, X_2, \ldots X_n$ with respective frequencies $f_1, f_2, \ldots, f_n$, the mean deviation about an average A is given by

$$\text{Mean Deviation about an average } A = \frac{\sum f|X - A|}{N} = \frac{\sum f|D|}{N},$$

where $N = \sum f$ is the total frequency and $D = X - A$

Procedure for Computing the Mean Deviation

Step 1. *Calculate the average A about which mean deviation is to be computed.*

Step 2. *Take the deviation of each observation from A and denote it by D. That is, find* $D = X - A$.

Step 3. *Find the absolute value of the deviation of each observation from A ignoring ± signs and denote it by* $|D|$.

Step 4. *Multiply each absolute deviation* $|D|$ *by the corresponding frequency f to get* $f|D|$.

Step 5. *Add all the products obtained in Step 4 to get* $\sum f|D|$.

Step 6. *Divide the sum obtained in Step 5 by N, the total frequency, to get the required mean deviation.*

Computation of Mean Deviation - Continuous Series

The computation of the mean deviation in the case of continuous series is exactly the same as discussed above for discrete series. The only difference is that here we have to

...the class marks (or mid-values) of the various classes and take absolute deviations ...se values from the average A. Thus, if $X_1, X_2, \ldots, X_n$ are the class marks (or mid-...) of a set of grouped data with corresponding class frequencies $f_1, f_2, \ldots, f_n$, then ...ean deviation about an average A is given by

$$\text{Mean Deviation (about an average } A) = \frac{\sum f_i |X_i - A|}{N} = \frac{\sum f_i |D_i|}{N}$$

where $N = \sum f_i$ is the total frequency.

...fficient of Mean Deviation

...measures of mean deviation as defined above are absolute measures depending on ... units of measurement. The relative measure corresponding to the mean deviation, ...d the **coefficient of mean deviation**, is given by

$$\text{Coefficient of } M.D. = \frac{\text{Mean Deviation}}{\text{Average about which it is calculated}}$$

$$\text{Coefficient of } M.D. \text{ about mean} = \frac{M.D}{\text{Mean}}$$

...d $\quad$ $\text{Coefficient of } M.D. \text{ about median} = \frac{M.D}{\text{Median}}$

...efficient of mean deviation is a pure number independent of the units of measurement and can be used to compare two distributions expressed in different units.

...AMPLE 9. *Find the mean deviation about the median for the following data:*

56 46 79 26 85 39 65 99 29 72

...nd also the coefficient of mean deviation.

SOLUTION. Arranging the data in ascending order of magnitude, we get

26 29 39 46 56 65 72 79 85 99

Here, $n = 10$

$$\frac{n+1}{2} = \frac{10+1}{2} = 5.5$$

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

$$= \text{size of (5.5)th item}$$

$$= 5\text{th item} + 0.5\,(6\text{th item} - 5\text{th item})$$

$$= 56 + 0.5\,(65 - 56) = 56 + 4.5 = 60.5$$

CALCULATION OF MEAN DEVIATION

26	-34.5	34.5
29	-31.5	31.5
39	-21.5	21.5
46	-14.5	14.5
56	-4.5	4.5
65	4.5	4.5
72	11.5	11.5
79	18.5	18.5
85	24.5	24.5
99	38.5	38.5
$n = 10$		$\Sigma\|D\| = 204$

$\therefore$ M.D. (about median) $= \frac{\Sigma|D|}{n} = \frac{204}{10} = 20.4.$

Coefficient of M.D. $= \frac{\text{Mean Deviation}}{\text{Median}} = \frac{20.4}{60.5} = 0.34.$

EXAMPLE 10. *Calculate mean deviation about the mean for the following data:*

X :	10	11	12	13	14	Total
f :	3	12	18	12	3	48

SOLUTION.

CALCULATIONS FOR MEAN DEVIATION

X	f	fX	$D = X - \bar{X}$	$\|D\|$	$f\|D\|$
10	3	30	-2	2	6
11	12	132	-1	1	12
12	18	216	0	0	0
13	12	156	1	1	12
14	3	42	2	2	6
	$N = \Sigma f = 48$	$\Sigma fX = 576$			$\Sigma f\|D\| = 36$

$$\text{Mean} : \bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{576}{48} = 12$$

$$\text{Mean Deviation about Mean} = \frac{\Sigma f|D|}{N} = \frac{36}{48} = 0.75.$$

EXAMPLE 11. *Calculate the mean deviation from the median for the following data:*

Marks :	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90
No. of Students :	2	6	12	18	25	20	10	7

Computation of Median :

$\frac{N}{2} = \frac{120}{2} = 60$; the c.f. just greater than or equal to 60 is 90.

Thus median lies in the class 35.5 - 40.5 and is given by

$$\text{Median } (Md) = l + \frac{\frac{N}{2} - C}{f} \times h = 35.5 + \frac{60 - 48}{42} \times 5 = 35.5 + 1.43 = 36.93$$

$$\therefore \quad M.D. \text{ (about Median)} = \frac{\Sigma f |D|}{N} = \frac{630.08}{120} = 5.26$$

$$\text{Coefficient of Mean Deviation (about median)} = \frac{M.D.}{\text{Median}} = 0.142.$$

4.10 MERITS AND DEMERITS OF MEAN DEVIATION

Merits. 1. *It is easy to understand and simple to calculate.*

2. *It is based on each and every item of the data.*

3. *It is rigidly defined.*

4. *As compared to standard deviation, it is less affected by extreme observations.*

5. *Since deviations are taken from a central value, comparison about formation of different distributions can easily be made.*

Demerits. 1. *The major drawback of mean deviation is that algebraic signs are ignored while taking the deviations of the items.*

2. *It is not suitable for further mathematical treatment.*

3. *It cannot be computed for distribution with open-end classes.*

4. *It is rarely used in sociological studies.*

EXERCISE 4.2

7. What do you understand by mean deviation?

What is coefficient of mean deviation? State the formula.

Calculate the mean deviation about the median in respect of the following data:

8. 15 53 49 19 62 7 15 93 [Calcutta Univ. B Com. 19

Find the mean deviation about the mean in respect of the following data:

Coefficient of *M.D.* = 0.206.

4.11 STANDARD DEVIATION

The concept of standard deviation was first introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of studying dispersion.

Definition. The **standard deviation**, abbreviated as *S.D.*, of a given set of observations is defined as the positive square root of the arithmetic mean of the squares of deviations of the observations from their arithmetic mean. It is denoted by the Greek small letter σ (read as sigma).

Thus standard deviation of a set of n observations $X_1, X_2, \ldots, X_n$ is given by

$$\sigma = \sqrt{\frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n}} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}},$$

where $\bar{X} = \frac{\Sigma X}{n}$ is the arithmetic mean of the given observations.

If $X_1, X_2, \ldots, X_n$ are the class marks of a set of grouped data with class frequencies $f_1, f_2, \ldots, f_n$ then the standard deviation is given by

$$\sigma = \sqrt{\frac{f_1(X_1-\bar{X})^2+f_2(X_2-\bar{X})^2+\ldots f_n(X_n-\bar{X})^2}{f_1+f_2+\ldots+f_n}} = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}},$$

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

$\bar{X} = \frac{\Sigma fX}{N}$ is the arithmetic mean of the distribution.

Definition. The variance of a given set of observations is defined as the square of standard deviation and is denoted by σ^2. Thus

EXAMPLE 14. Calculate standard deviation of the following marks obtained by 5 students in a tutorial group: [Delhi Univ. B.Com. 1997]

Marks obtained : 8 12 13 15 22

SOLUTION. **CALCULATION OF STANDARD DEVIATION**

X	$X-\bar{X}$	$(X-\bar{X})^2$
8	−6	36
12	−2	4
13	−1	1
15	1	1
22	8	64
$\Sigma X = 70$		$\Sigma(X-\bar{X})^2 = 106$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{70}{5} = 14$$

and $$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}} = \sqrt{\frac{106}{5}} = \sqrt{21.2} = 4.6 \text{ (app.)}.$$

EXAMPLE 15. *Initially there were 9 workers, all being paid a uniform wage. Later on, a 10th worker is added whose wage rate is Rs. 20 less than for the others. Compute :*

(i) the effect on the mean wage.

(ii) standard deviation of wages for the group of 10 workers.

[*Delhi Univ. B.A. (Econ. Hons.) 1998*]

SOLUTION. Let the uniform wage paid to each of the 9 workers be Rs. w. Then the wage paid to the 10th worker is Rs. $(w - 20)$.

(i) The mean wage of 10 workers $= \frac{w+w+...+w+(w-20)}{10} = \frac{10w-20}{10} = w-2$

Thus after the addition of 10th worker, the mean wage has reduced by Rs. 2.

(ii) **COMPUTATION OF S.D.**

S.No.	X	$X-\bar{X}$	$(X-\bar{X})^2$
1	w	2	4
2	w	2	4
.	.	.	.
.	.	.	.
.	.	.	.
9	w	2	4
10	$w-20$	−18	324

$$\sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n}} = \sqrt{\frac{(4+4+\ldots+4)+324}{10}} = \sqrt{\frac{4\times 9+324}{10}} = \sqrt{\frac{360}{10}} = \sqrt{36} = 6$$

Thus the standard deviation of wages for the group of 10 workers is Rs. 6.

EXAMPLE 16. *Calculate the standard deviation for the following data:*

X :	20	30	40	50	60	70
f :	8	12	20	10	6	4

SOLUTION.

CALCULATION OF STANDARD DEVIATION

X	f	fX	$X-\bar{X}$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
20	8	160	−21	441	3528
30	12	360	−11	121	1452
40	20	800	−1	1	20
50	10	500	9	81	810
60	6	360	19	361	2166
70	4	280	29	841	3364
	$N=60$	$\Sigma fX=2460$			$\Sigma f(X-\bar{X})^2=11{,}340$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{2460}{60} = 41$$

and

$$\sigma = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}} = \sqrt{\frac{11340}{60}} = \sqrt{189} = 13.75.$$

EXAMPLE 17. *Calculate the mean and standard deviation from the following data:*

Marks :	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of Students :	10	15	25	25	10	10	5

SOLUTION.

CALCULATION OF STANDARD DEVIATION

Marks	Mid value	No. of Students				
	X	f	fX	$X-\bar{X}$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
0 – 10	5	10	50	−26	676	6760
10 – 20	15	15	225	−16	256	3840
20 – 30	25	25	625	−6	36	900
30 – 40	35	25	875	4	16	400
40 – 50	45	10	450	14	196	1960
50 – 60	55	10	550	24	576	5760
60 – 70	65	5	325	34	1156	5780

$$\bar{X} = \frac{\sum fX}{N} = \frac{3100}{100} = 31$$

and

$$\sigma = \sqrt{\frac{\sum f(X-\bar{X})^2}{N}} = \sqrt{\frac{25400}{100}} = \sqrt{254} = 15.94.$$

Different Methods of Calculating Standard Deviation - Ungrouped Data

By definition the standard deviation of an ungrouped data consisting of n observations $X_1, X_2, \ldots, X_n$ is given by

$$\sigma = \sqrt{\frac{\sum(X-\bar{X})^2}{N}} \quad \text{where} \quad \bar{X} = \frac{\sum X}{n}$$

The computation of standard deviation by definition is very effective if $\bar{X}$ is an integer. However, if $\bar{X}$ comes out to be in fraction, its computation becomes very cumbersome and time-consuming. In that case we apply the following short-cut method which is very effective and reduces the numerical calculations to a great extent.

Formula 1 (Short-cut Method). According to this method, standard deviation is given by

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

Again if the values of X are large, Formula 1 can further be simplified by using what is known as an *assumed mean method*. When this method is used we take deviations of each item from an assumed mean, say A.

Formula 2 (Assumed Mean Method). According to this method, standard deviation is given by

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}, \quad \text{where } d = X - A$$

Finally, if c is the common factor of the given data, then we may apply the following method, called *step-deviation method.*

Formula 3 (Step-deviation Method). According to this method, standard deviation is given by

$$\sigma = \sqrt{\frac{\sum u^2}{n} - \left(\frac{\sum u}{n}\right)^2} \times c,$$

where $u = \frac{X-A}{c}$ and c = common factor.

REMARK. If we compare Formula 2 with Formula 1, we immediately conclude that standard deviation is independent of change of origin. However, if we compare Formula

3 with Formula 2, we find that standard deviation is not independent of change of scale. Thus the above two observations lead to the following important conclusion:

"Standard deviation of a distribution is independent of change of origin but not of scale."

An Important Result: $\sum X^2 = n(\sigma^2 + \bar{X}^2)$

The above result follows immediately from Formula 1. In fact, applying Formula 1, we get

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

$$\Rightarrow \quad \frac{1}{n}\sum X^2 = \sigma^2 + \bar{X}^2 \quad \text{or} \quad \sum X^2 = n(\sigma^2 + \bar{X}^2).$$

The above result is quite useful to find the correct value of standard deviation whenever one or more of the observations are wrongly copied down. For instance, see Example 27.

Different Methods of Calculating Standard Deviation - Grouped Data

If $X_1, X_2, \ldots, X_n$ are the class marks of a set of grouped data with corresponding class frequencies $f_1, f_2, \ldots, f_n$ then the standard deviation is given by

$$\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$$

where $N = \sum f$ is the total frequency and $\bar{X} = \frac{\sum fX}{N}$ is the arithmetic mean of the distribution.

All the methods discussed earlier for calculating standard deviation in the case of ungrouped data can also be used in the case of grouped data. However, in practice it is the step deviation method that is mostly used.

Short-cut Method. According to this method, standard deviation is given by

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

Assumed Mean Method. According to this method, standard deviation is given by

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}, \quad \text{where } d = X - A$$

$$\sigma = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h \quad \text{where } u = \frac{X - A}{h}$$

EXAMPLE 18. *From the following information, find the standard deviation of X and Y variable:*

$\Sigma X = 235$, $\Sigma Y = 250$, $\Sigma X^2 = 6750$, $\Sigma Y^2 = 6840$, $N = 10$

[Delhi Univ. B.Com. (H) 1997]

SOLUTION. *Standard Deviation of X variable.*

$$\sigma_X = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{6750}{10} - \left(\frac{235}{10}\right)^2}$$

$$= \sqrt{675 - (23.5)^2} = \sqrt{675 - 552.25} = \sqrt{122.75} = 11.08$$

Standard deviation of Y variable.

$$\sigma_Y = \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2} = \sqrt{\frac{6840}{10} - \left(\frac{250}{10}\right)^2} = \sqrt{684 - (25)^2}$$

$$= \sqrt{684 - 625} = \sqrt{59} = 7.68.$$

EXAMPLE 19. *Find the mean and standard deviation of first n natural numbers.*

SOLUTION. The first n natural numbers are 1, 2, 3, . . . , n. We know that

$$\Sigma X = 1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

and

$$\Sigma X^2 = 1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\therefore \quad \text{Mean} = \bar{X} = \frac{\Sigma X}{n} = \frac{1+2+3+\ldots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

and

$$S.D. = \sigma = \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} = \sqrt{\frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2}$$

$$= \sqrt{\frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}} = \sqrt{\frac{2(n+1)(2n+1) - 3(n+1)^2}{12}}$$

$$= \sqrt{\frac{(n+1)[4n+2-3n-3]}{12}} = \sqrt{\frac{(n+1)(n-1)}{12}} = \sqrt{\frac{n^2-1}{12}}$$

Thus, the mean of first n natural numbers is $\frac{n+1}{2}$ and their standard deviation is $\sqrt{\frac{n^2-1}{12}}$.

EXAMPLE 20. *Twenty passengers were found ticketless on a bus. The sum of squares and the standard deviation of the amount found in their pockets were Rs. 2000 and Rs. 6 respectively. If the total fine imposed on these passengers is equal to the total amount recovered from them and fine imposed is uniform, what is the amount each one of them has to pay as fine? What difficulties do you visualize if such a system of penalty were imposed?*

[Delhi Univ. B.A. (Eco. Hons.) 1993]

SOLUTION. Let X_i denote the amount (in Rs.) found in the pocket of ith passenger, $i = 1, 2, ..., 20$. Then we are given :

$$n = 20 \qquad \sum_{i=1}^{20} X_i^2 = 2000 \quad \text{and} \quad \sigma = 6$$

The total fine imposed on 20 ticketless passengers is given to be equal to the total amount recovered from them.

$\therefore$ Total fine imposed on 20 passengers $= \sum_{i=1}^{20} X_i$

$\Rightarrow$ Fine to be paid by each passenger $= \bar{X} = \frac{1}{20} \sum_{i=1}^{20} X_i$ (*$\because$ fine imposed is uniform*)

Now, $$\sigma^2 = \frac{\sum X_i^2}{20} - \bar{X}^2$$

$\Rightarrow$ $$\bar{X}^2 = \frac{\sum X_i^2}{20} - \sigma^2 = \frac{2000}{20} - 36 = 100 - 36 = 64$$

i.e., $$\bar{X}^2 = 64 \Rightarrow \bar{X} = 8.$$

Thus, the fine to be paid by each passenger is Rs. 8.

If among these ticketless passengers, we find a few rich persons with large amount of money in their pockets, then an obvious shortcoming of this system of imposing penalty is that it will give undue heavy penalty to the poor passengers (having less amount of money in their pocket).

EXAMPLE 21. *A charitable organisation decided to give old age pension to people over sixty years of age. The scales of pension were fixed as follows :*

Age Group (years)	Amount Per Month (Rs.)
60 – 65	20 per month
65 – 70	25 per month
70 – 75	30 per month
75 – 80	35 per month
80 – 85	40 per month

4.26

74	62	64	72	61	83	72	81	64	71	63	61
60	67	74	64	79	75	76	69	68	78	66	67

Calculate the monthly average pension payable per pension and the standard deviation.

[Delhi Univ. B.Com. (H) 2005 (C.C.)]

SOLUTION. **CALCULATION OF AVERAGE MONTHLY PENSION AND STANDARD DEVIATION**

Class		f	x	$d = x - 30$	fd	fd^2
60 – 65	IIII	7	20	– 10	– 70	700
65 – 70	IIII	5	25	– 5	– 25	125
70 – 75	IIII	6	30	0	0	0
75 – 80	IIII	4	35	5	20	100
80 – 85	III	3	40	10	30	300
		$N = \sum f = 25$			$\sum fd = -45$	$\sum fd^2 = 1225$

The average monthly pension is given by

$$\bar{X} = A + \frac{\sum fd}{N} = 30 + \frac{-45}{25} = 30 - 1.8 = \text{Rs. } 28.20.$$

Computation of Standard Deviation :

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{1225}{25} - \left(\frac{-45}{25}\right)^2}$$

$$= \sqrt{49 - 3.24} = \sqrt{45.76} = \text{Rs. } 6.76.$$

EXAMPLE 22. *Calculate the arithmetic mean and standard deviation from the following series:*

Class Interval	:	5 – 15	15 – 25	25 – 35	35 – 45	45 – 55
Frequency	:	8	12	15	9	6

[Delhi Univ. B.Com. 1973]

SOLUTION. **CALCULATION OF A.M. AND S.D.**

Class Interval	Mid value (x)	Frequency (f)	$u = \frac{x - A}{h}$ ($A = 30, h = 10$)	fu	fu^2
5 – 15	10	8	– 2	– 16	32
15 – 25	20	12	– 1	– 12	12
25 – 35	30	15	0	0	0
35 – 45	40	9	1	9	9
45 – 55	50	6	2	12	24
		$N = 50$		$\sum fu = 7$	$\sum fu^2 = 77$

Calculation of Arithmetic Mean. The A.M. is given by the formula

$$\bar{X} = A + \frac{\sum fu}{N} \times h = 20 + \frac{-7}{50} \times 10 = 30 - 1.4 = 28.6.$$

Calculation of Standard Deviation: The S.D. is given by the formula

$$\sigma = \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} \times h = \sqrt{\frac{77}{50} - \left(\frac{-7}{50}\right)^2} \times 10$$

$$= \sqrt{1.54 - 0.0196} \times 10 = 12.33.$$

EXAMPLE 23. *In a certain city, a survey was conducted in respect of profits or losses made by 100 firms. The following results were obtained :*

Profit or Loss (in crore Rs.)	No. of Firms	Profit or Loss (in crore Rs.)	No. of Firms
– 40 to – 30	7	10 to 20	5
– 30 to – 20	9	20 to 30	10
– 20 to – 10	8	30 to 40	30
– 10 to 0	6	40 to 50	12
0 to 10	5	50 to 60	8

Calculate the average profits and standard deviation of profits.

SOLUTION. **CALCULATION FOR MEAN AND STANDARD DEVIATION**

Profit or Loss (in crore Rs.)	Mid-value X	No. of Firms f	$u = \frac{X - 15}{10}$ $(A = 15, h = 10)$	fu	fu^2
– 40 to – 30	– 35	7	– 5	– 35	175
– 30 to – 20	– 25	9	– 4	– 36	144
– 20 to – 10	– 15	8	– 3	– 24	72
– 10 to 0	– 5	6	– 2	– 12	24
0 to 10	5	5	– 1	– 5	5
10 to 20	15	5	0	0	0
20 to 30	25	10	1	10	10
30 to 40	35	30	2	60	120
40 to 50	45	12	3	36	108
50 to 60	55	8	4	32	128
		$N = \sum f = 100$		$\sum fu = 26$	$\sum fu^2 = 786$

Calculation of arithmetic mean :

Calculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h = \sqrt{\frac{786}{100} - \left(\frac{26}{100}\right)^2} \times 10$$

$$= \sqrt{7.86 - 0.0676} \times 10 = \sqrt{7.7924} \times 10$$

$$= 2.791 \times 10 = \text{Rs. } 27.91 \text{ crores.}$$

EXAMPLE 24. *Calculate the arithmetic mean and standard deviation from the following data:*

Values (more than):	800	700	600	500	400	300	200	100
Frequency	14	44	96	175	381	527	615	660

[*Kerala Univ. B.Com. 2003*]

SOLUTION. Since data is given in the form of cumulative frequency distribution, it has to be arranged in a frequency distribution as shown in the following table :

CALCULATION OF *A.M.* AND *S.D.*

Values (more than)	Class	Cumulative Frequency	Frequency	Mid-value	$d = \frac{x - 450}{100}$ ($A = 450$, $h = 100$)	fu	fu^2
800	800 – 900	14	14	850	+4	56	224
700	700 – 800	44	44 – 14 = 30	750	+3	90	270
600	600 – 700	96	96 – 44 = 52	650	+2	104	208
500	500 – 600	175	175 – 96 = 79	550	+1	79	79
400	400 – 500	381	381 – 175 = 206	450	0	0	0
300	300 – 400	527	527 – 381 = 146	350	–1	–146	146
200	200 – 300	615	615 – 527 = 88	250	–2	–176	352
100	100 – 200	660	660 – 615 = 45	∴ 150	–3	–135	405
			$N = 660$			$\Sigma fu = -128$	$\Sigma fu^2 = 1684$

Computation of arithmetic mean :

$$\bar{X} = A + \frac{\Sigma fu}{N} \times h = 450 + \frac{-128}{660} \times 100 = 450 - 19.39 = 430.61.$$

Calculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h = \sqrt{\frac{1684}{660} - \left(\frac{-128}{660}\right)^2} \times 100.$$

$$= \sqrt{2.5515 - 0.0376} \times 100 = \sqrt{2.5139} \times 100 = 1.5855 \times 100 = 158.55.$$

EXAMPLE 25. *Mean and Standard deviation of the following continuous series are 31 and 15.9 respectively. The distribution after taking step deviation is as follows :*

d	–3	–2	–1	0	1	2	3
f	10	15	25	25	10	10	5

Determine the actual class intervals. [Delhi Univ. B.Com. (H) 1990]

SOLUTION. Let A be the assumed mean, h the width of each class interval so that $d = \frac{X-A}{h}$.

COMPUTATION FOR DETERMINING CLASS INTERVALS

d	f	fd	fd^2
−3	10	−30	90
−2	15	−30	60
−1	25	−25	25
0	25	0	0
1	10	10	10
2	10	20	40
3	5	15	45
	$N = \Sigma f = 100$	$\Sigma fd = -40$	$\Sigma fd^2 = 270$

The *S.D.* is given by the formula :

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

$$\Rightarrow \quad 15.9 = \sqrt{\frac{270}{100} - \left(\frac{-40}{100}\right)^2} \times h$$

$$\Rightarrow \quad 15.9 = \sqrt{2.70 - 0.16} \times h \quad \text{or} \quad 15.9 = \sqrt{2.54}\, h \; 1.59h$$

$$\Rightarrow \quad 1.59h = 15.9 \quad \text{or} \quad h = \frac{15.9}{1.59} = 10.$$

Thus the width of each class interval is 10.

The mean is given by the formula

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

$$\Rightarrow \quad 31 = A + \frac{-40}{100} \times 10 \quad \text{or} \quad 31 = A - 4 \quad \Rightarrow \quad A = 35$$

Hence assumed mean from which deviations have been taken is $A = 35$. Using the formula $d = \frac{X-A}{h}$ or $X = A + dh$, the various class marks are :

5 15 25 35 45 55 65

Class Interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Frequency	10	15	25	25	10	10	5

EXAMPLE 26. *The mean of 5 observations is 4.4 and the variance is 8.24. If three of the five observations are 1, 2 and 6, find the values of the other two.* [Delhi Univ. B.Com. (H) 1995]

SOLUTION. We are given : $n = 5$, $\bar{X} = 4.4$ and $\sigma^2 = 8.24$

Now $\Sigma X = n\bar{X} = 5 \times 4.4 = 22$

and $\Sigma X^2 = n(\sigma^2 + \bar{X}^2) = 5\,(8.24 + 19.36) = 5 \times 27.60 = 138$

Let the two missing observations be X_1 and X_2. Then

$$\Sigma X = 22 \quad \Rightarrow \quad 1 + 2 + 6 + X_1 + X_2 = 22$$

$$\Rightarrow \quad X_1 + X_2 = 22 - 9 = 13 \qquad \ldots (1)$$

and $$\Sigma X^2 = 138 \quad \Rightarrow \quad 1^2 + 2^2 + 6^2 + X_1^2 + X_2^2 = 138$$

$$\Rightarrow \quad X_1^2 + X_2^2 = 138 - 41 = 97 \qquad \ldots (2)$$

Equation (1) gives $X_2 = 13 - X_1$. Substituting this value of X_2 in Equation (2), we obtain

$$X_1^2 + (13 - X_1)^2 = 97$$

$$\Rightarrow \quad X_1^2 + 169 + X_1^2 - 26X_1 = 97$$

or, $$2X_1^2 - 26X_1 + 72 = 0 \quad \text{or} \quad X_1^2 - 13X_1 + 36 = 0$$

$$\Rightarrow \quad X_1 = \frac{13 \pm \sqrt{169 - 144}}{2} = \frac{13 \pm 5}{2} = 9 \quad \text{or} \quad 4$$

If $X_1 = 9$, then $X_2 = 4$, and if $X_1 = 4$, then $X_2 = 9$.

Thus the values of the other two observations are 9 and 4.

Correcting Incorrect values of Mean and Standard Deviation

Quite often while tabulating data some values are wrongly copied down. For example, it may happen that an observation, say 36, was wrongly copied down as 63 and we might have used this wrong value in the calculation of Mean and Standard Deviation. When such an error is found out there are two options – either to make all calculations a fresh, which is a very tedious task or to correct the values of mean and standard deviation by adjustment and readjustment of wrong and right figures. This second option is much easier and consumes much less time. To illustrate, let us consider the following example.

EXAMPLE 27. *The mean and standard deviation of 100 items are found to be 40 and 10. If at the time of calculations, two items are wrongly taken as 30 and 70 instead of 3 and 27, find the corrected mean and corrected standard deviation.*

[Delhi Univ. B.Com. 1986, 2001]

SOLUTION. In usual notations, we are given : $n = 100$, $\bar{X} = 40$ and $\sigma = 10$

Now, $\Sigma X = n\bar{X} = 100 \times 40 = 4000$

and $\Sigma X^2 = n(\sigma^2 + \bar{X}^2) = 100\,(100 + 1600) = 100\,(1700) = 1,70,000$

If the wrong observations 30 and 70 are replaced by the correct observations 3 and 27, then

$$\text{Corrected } \Sigma X = 4000 - 30 - 70 + 3 + 27 = 3930$$

$$\text{Corrected } \Sigma X^2 = 1,70,000 - (30)^2 - (70)^2 + (3)^2 + (27)^2$$

$$= 1,70,000 - 900 - 4900 + 9 + 729 = 1,64,938$$

$$\therefore \quad \text{Corrected } \bar{X} = \frac{\text{Corrected } \Sigma X}{n} = \frac{3930}{100} = 39.30$$

$$\text{Corrected } \sigma = \sqrt{\frac{\text{Corrected } \Sigma X^2}{n} - (\text{Corrected } \bar{X})^2} = \sqrt{\frac{164938}{100} - (39.30)^2}$$

$$= \sqrt{1649.38 - 1544.49} = \sqrt{104.89} = 10.24.$$

4.12 COMBINED STANDARD DEVIATION

If two sets of data contain n_1 and n_2 observations having means $\bar{X}_1$ and $\bar{X}_2$, and standard deviations σ_1 and σ_2 respectively, then the standard deviation, σ, of the combined data with $n_1 + n_2$ observations is given by

$$\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}},$$

where $d_1 = \bar{X}_1 - \bar{X}$ $d_2 = \bar{X}_2 - \bar{X}$

and $\bar{X} = \dfrac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$ is the combined mean

The result can be generalised to more than two sets of data. For example, if $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ be the means, $\sigma_1, \sigma_2, \ldots, \sigma_k$ be the standard deviations and $n_1, n_2, \ldots, n_k$ be the number of observations in each set, then the standard deviation of the combined data with $n_1 + n_2 + \ldots + n_k$ observations is given by

$$\sqrt{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_k(\sigma_k^2 + d_k^2)}$$

where $d_1 = \bar{X}_1 - \bar{X}$, $d_2 = \bar{X}_2 - \bar{X}$, ..., $d_k = \bar{X}_k - \bar{X}$

and $\bar{X} = \dfrac{n_1\bar{X}_1 + n_2\bar{X}_2 + ... + n_k\bar{X}_k}{n_1 + n_2 + ... + n_k}$ is the combined mean.

EXAMPLE 28. *From the following information, calculate the combined standard deviation:*

$n_1 = 90 \quad \bar{X}_1 = 20 \quad \sigma_1 = 8$

$n_2 = 60 \quad \bar{X}_2 = 15 \quad \sigma_2 = 6$ **[Delhi Univ. B.Com. 2004]**

SOLUTION. Let $\bar{X}$ denote the combined mean and σ denote the combined standard deviation. Then

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{90\times20 + 60\times15}{90+60} = \frac{1800+900}{150} = \frac{2700}{150} = 18$$

and

$$\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}},$$

where

$d_1 = \bar{X}_1 - \bar{X} = 20 - 18 = 2 \Rightarrow d_1^2 = 4$

$d_2 = \bar{X}_2 - \bar{X} = 15 - 18 = -3 \Rightarrow d_2^2 = 9$

$$\sigma = \sqrt{\frac{90(64+4) + 60(36+9)}{90+60}} = \sqrt{\frac{90\times68 + 60\times45}{150}}$$

$$= \sqrt{\frac{6120+2700}{150}} = \sqrt{\frac{8820}{150}} = \sqrt{58.8} = 7.67 \text{ (app.)}$$

EXAMPLE 29. *Find the missing information from the following :*

	Group I	Group II	Group III	Combined
Number	50	?	90	200
Standard Deviation	6	?	?	7.746
Mean	113	?	115	116

[Delhi Univ. B.Com. (H) 1992]

SOLUTION. Let n_1, n_2, n_3 denote the number of observations in Group *I*, Group *II*, Group *III* respectively. Then we are given

$$n_1 + n_2 + n_3 = 200$$

But $n_1 = 50$ and $n_3 = 90 \quad \therefore \quad 50 + n_2 + 90 = 200 \Rightarrow n_2 = 60$

Let $\bar{X}_1, \bar{X}_2, \bar{X}_3$ denote the means of Group *I*, Group *II*, Group *III* respectively. Then the combined mean, $\bar{X}$, is given by

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3}$$

$$\Rightarrow \quad 116 = \frac{50\times113 + 60\times\bar{X}_2 + 90\times115}{200}$$

$$\Rightarrow \quad 23200 = 5650 + 60\,\bar{X}_2 + 10350$$

$$\Rightarrow \quad 60\,\bar{X}_2 = 23200 - 5650 - 10350 = 7200$$

$$\Rightarrow \quad \bar{X}_2 = \frac{7200}{60} = 120$$

Let $\sigma_1, \sigma_2, \sigma_3$ denote the standard deviations of Group *I*, Group *II*, Group *III* respectively. Then the combined standard deviation, σ, is given by

$$\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)}{n_1 + n_2 + n_3}}$$

$$\Rightarrow \quad (n_1 + n_2 + n_3)\sigma^2 = n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)$$

where $d_1 = \bar{X}_1 - \bar{X} = 113 - 116 = -3$

$d_2 = \bar{X}_2 - \bar{X} = 120 - 116 = 4$

$d_3 = \bar{X}_3 - \bar{X} = 115 - 116 = -1$

$$\therefore \quad 200\,(7.746)^2 = 50\,(36 + 9) + 60\,(49 + 16) + 90\,(\sigma_3^2 + 1)$$

$$\Rightarrow \quad 200\,(60.000516) = 2250 + 3900 + 90 + 90\,\sigma_3^2$$

$$\Rightarrow \quad 12000 = 6240 + 90\,\sigma_3^2$$

$$\Rightarrow \quad 90\,\sigma_3^2 = 12000 - 6240 = 5760$$

$$\Rightarrow \quad \sigma_3^2 = \frac{5760}{90} = 64$$

$$\Rightarrow \quad \sigma_3 = 8 \qquad \text{(rejecting - ve value)}$$

Thus the missing figures are: $n_2 = 60$, $\bar{X}_2 = 120$ and $\sigma_3 = 8$.

EXAMPLE 30. *The first of two sub-groups has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$, find the standard deviation of the second sub-group.*

No. of items	$n_1 = 100$	$n_2 = 150$	$n_1 + n_2 = 250$
Mean	$\bar{X}_1 = 15$	$\bar{X}_2 = ?$	$\bar{X} = 15.6$
Standard deviation	$\sigma_1 = 3$	$\sigma_2 = ?$	$\sigma = \sqrt{13.44}$

Applying the formula for the combined mean :

$$\therefore \quad \bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

we get $$15.6 = \frac{100 \times 15 + 150\bar{X}_2}{250}$$

$$\Rightarrow \quad 3900 = 1500 + 150\,\bar{X}_2 \quad \Rightarrow \quad 150\,\bar{X}_2 = 2400 \quad \Rightarrow \quad \bar{X}_2 = 16$$

Also, $$\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

where $$d_1 = \bar{X}_1 - \bar{X} = 15 - 15.6 = -0.6$$

$$d_2 = \bar{X}_2 - \bar{X} = 16 - 15.6 = 0.4$$

$$13.44 = \frac{100(9 + 0.36) + 150\left(\sigma_2^2 + 0.16\right)}{250}$$

$$\Rightarrow \quad 13.44 \times 250 = 936 + 24 + 150\,\sigma_2^2$$

or, $$3360 = 960 + 150\,\sigma_2^2 \quad \Rightarrow \quad 150\,\sigma_2^2 = 2400$$

$$\Rightarrow \quad \sigma_2^2 = 16 \quad \text{or} \quad \sigma_2 = 4$$

Thus the mean and standard deviation of the second sub-group are 16 and 4 respectively.

EXAMPLE 31. *For a group containing 100 items, the arithmetic mean and standard deviation are 8 and $\sqrt{10.5}$. For 50 observations selected from these 100 observations, the mean and standard deviation are 10 and 2 respectively. Find the mean and standard deviation of the remaining 50 observations.* **[Delhi Univ. B.Com (H) 1994]**

SOLUTION. The given information can be put in the tabular form as follows :

	Group I	Group II	Combined
Number	$n_1 = 50$	$n_2 = 50$	$n_1 + n_2 = 100$
Mean	$\bar{X}_1 = 10$	$\bar{X}_2 = ?$	$\bar{X} = 8$
Standard deviation	$\sigma_1 = 2$	$\sigma_2 = ?$	$\sigma = \sqrt{10.5}$

We have $\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} \Rightarrow 8 = \frac{50\times 10 + 50\bar{X}_2}{100}$

$\Rightarrow 800 = 500 + 50\bar{X}_2 \Rightarrow 50\bar{X}_2 = 300 \Rightarrow \bar{X}_2 = 6$

Also $\sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$

$\Rightarrow (n_1 + n_2)\sigma^2 = n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)$,

where $d_1 = \bar{X}_1 - \bar{X} = 10 - 8 = 2$

$d_2 = \bar{X}_2 - \bar{X} = 6 - 8 = -2$

$\therefore 100\,(10.5) = 50\,(4 + 4) + 50\,(\sigma_2^2 + 4)$

$\Rightarrow 1050 = 400 + 200 + 50\sigma_2^2$

$\Rightarrow 50\sigma_2^2 = 1050 - 600 = 450$

$\Rightarrow \sigma_2^2 = \frac{450}{50} = 9$ or $\sigma_2 = 3$

Thus the mean and standard deviation of the remaining 50 items are 6 and 3 respectively.

EXAMPLE 32. *For two groups of observations the following results were available :*

Group I	*Group II*
$\Sigma(X-5) = 8$	$\Sigma(X-8) = -10$
$\Sigma(X-5)^2 = 40$	$\Sigma(X-8)^2 = 70$
$n_1 = 20$	$n_2 = 25$

Find the mean and standard deviation of both the groups taken together.

[Delhi Univ B.Com (H) 2006]

SOLUTION. In terms of usual notations, we are given the following information :

	Group I	*Group II*
Σd	8	-10
Σd^2	40	70
n	20	25
A	-5	8

COMPUTATION OF MEAN AND STANDARD DEVIATION

$\bar{X}_1 = A + \frac{\sum d}{n} = 5 + \frac{8}{20} = 5 + 0.4 = 5.4$	$\bar{X}_2 = A + \frac{\sum d}{n} = 8 + \frac{-10}{25} = 8 - 0.4 = 7.6$
$\sigma_1 = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{40}{20} - \left(\frac{8}{20}\right)^2}$	$\sigma_2 = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{70}{25} - \left(\frac{-10}{25}\right)^2}$
$= \sqrt{2 - 0.16} = \sqrt{1.84} = 1.36$	$= \sqrt{2.8 - 0.16} = \sqrt{2.64} = 1.62$

Computation of Combined Mean and Combined Standard Deviation

Combined Mean: $\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{20 \times 5.4 + 25 \times 7.6}{20 + 25}$

$= \frac{108 + 190}{45} = \frac{298}{45} = 6.62.$

Combined S.D.: $\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$,

where $d_1 = \bar{X}_1 - \bar{X} = 5.4 - 6.62 = -1.22 \Rightarrow d_1^2 = 1.48$ (app.)

$d_2 = \bar{X}_2 - \bar{X} = 7.6 - 6.62 = 0.98 \Rightarrow d_2^2 = 0.96$ (app.)

$\therefore \sigma = \sqrt{\frac{20(1.84 + 1.48) + 25(2.62 + 0.96)}{20 + 25}} = \sqrt{\frac{20 \times 3.32 + 25 \times 3.58}{45}}$

$= \sqrt{\frac{66.4 + 89.5}{45}} = \sqrt{\frac{155.9}{45}} = \sqrt{3.46} = 1.86$ (app.)

4.13 COEFFICIENT OF VARIATION

The standard deviation is an absolute measure of dispersion, depending upon the units of measurement. It does not tell us much about the variability of a single set of data. The *coefficient of standard deviation*, based on standard deviation, is a relative measure of dispersion and is given by

$$\text{Coefficient of Standard Deviation} = \frac{\text{S.D.}}{\text{Mean}} = \frac{\sigma}{\bar{X}}$$

This is a pure number independent of the units of measurement and hence can be used to compare the variability of two distributions expressed in different units. Perhaps a more appropriate measure is the *coefficient of variation* (C.V.), defined by

$$\text{Coefficient of Variation} = \frac{S.D.}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{X}} \times 100$$

which expresses the standard deviation as a percentage of the mean. Since C.V. is a measure of *relative variation* expressed as a per cent, the coefficient of variation can be used to compare the variability of two or more sets of data even when the observations are expressed in different units of measurement.

A distribution for which the coefficient of variation is smaller is said to be less variable or more consistent, more uniform, more stable or more homogeneous. On the other hand, the distribution for which the coefficient of variation is greater is said to be more variable or less consistent, less uniform, less stable or less homogeneous.

EXAMPLE 33. *Arithmetic mean and standard deviation of monthly profits of two companies X and Y for a year are given below:*

	X	*Y*
Mean	*100*	90
Standard Deviation	*25*	*18*

Comment on the consistency of these companies with respect to their profits.

[*Delhi Univ. B.Com. 1990*]

SOLUTION. *Company X*

$$C.V. = \frac{\sigma_X}{\bar{X}} \times 100$$

$$= \frac{25}{100} \times 100 = 25\%$$

Company Y

$$C.V. = \frac{\sigma_Y}{\bar{Y}} \times 100$$

$$= \frac{18}{90} \times 100 = 20\%$$

Since *C.V.* for company *Y* is less than *C.V.* for company *X*, company *Y* is considered to be more consistent.

EXAMPLE 34. *The means and standard deviation of two brands of light bulbs are given below:*

	Brand I	*Brand II*
Mean	*800 hours*	*770 hours*
Standard deviation	*100 hours*	*60 hours*

Calculate a measure of relative dispersion for the two brands and interpret the result.

[*Delhi Univ B.Com. (H) 2000*]

SOLUTION. Coefficient of variation for Brand *I* $= \frac{\sigma}{\bar{X}} \times 100 = \frac{100}{800} \times 100 = 12.5\%$

Coefficient of variation for Brand *II* $= \frac{\sigma}{\bar{X}} \times 100 = \frac{60}{770} \times 100 = 7.8\%$

Since C.V. for Brand *I*, Brand *II* bulbs are considered to be

EXAMPLE 35. *After settlement the average weekly wage in a factory increased from Rs. 8000 to Rs. 12000 and standard deviation had increased from Rs. 100 to Rs. 150. After settlement the wage has become higher and more uniform." Do you agree?*

[Delhi Univ B.Com. (H) 2005]

SOLUTION. Since the average weekly wage has increased from Rs. 8000 to Rs. 12000, therefore weekly wages have become higher.

$$C.V.\text{ (before settlement)} = \frac{100}{8000} \times 100 = 1.25\%$$

$$C.V.\text{ (after settlement)} = \frac{150}{12000} \times 100 = 1.25\%$$

Since there is no change in *C.V.*, there is no improvement in uniformity.

EXAMPLE 36. *Verify the correctness of the following statement:*

A batsmen scored at an average of 60 runs an inning against Pakistan. The standard deviation of the runs scored by him was 12. A year later against Australia, his average came down to 50 runs an inning and the standard deviation of the runs scored fell down to 9. Therefore, it is correct to say that his performance was worse against Australia and that there was lesser consistency in his batting against Australia.

[Delhi Univ. B.Com (H) 1986]

SOLUTION. The average and standard deviation of runs scored by the batsman against Pakistan and Australia are shown below :

	Pakistan	*Australia*
Average	60	50
Standard deviation	12	9

Since the average runs scored by the batsman are less against Australia, it is correct to say that his performance was worse against Australia. To check consistency, we compute coefficient of variation :

$$C.V.\text{ (Pakistan)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{12}{60} \times 100 = 20\%$$

$$C.V.\text{ (Australia)} = \frac{\sigma}{\bar{X}} \times 100 = \frac{9}{50} \times 100 = 18\%.$$

Since coefficient of variation is less against Australia, there was greater consistency in his batting against Australia. Hence the given statement about consistency is not correct.

EXAMPLE 37. *The mean and coefficient of standard deviation of 100 items are found to be 50 and 0.1. If at the time of calculations, two items are wrongly taken as 40 and 50 instead of 60 and 30, find the correct mean and correct standard deviation.*

[Delhi Univ. B.Com. (H) 1996]

SOLUTION. In usual notations, we are given :

equal distribution than the curve *APB*. Thus variation in group *B* is greater than the variation in group *A*.

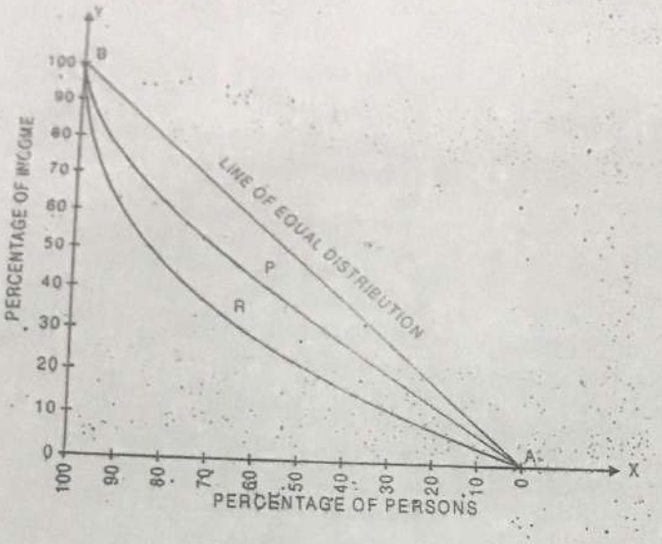


Fig. 4.3. Lorenz Curves

REMARK. The Lorenz curve has a great drawback. It does not give us numerical value of the measure of distribution. It merely gives us a picture of the extent to which a series is pulled away from the line of equal distribution. Accordingly it should be used along with some numerical measure of dispersion.

LIST OF FORMULAE USED

Measure	*Ungrouped Data (Individual Series)*	*Grouped Data (Discrete and Continuous Series)*
Range	Range $= L - S$	Range $= L - S$
	Coefficient of Range $= \dfrac{L - S}{L + S}$	Coefficient of Range $= \dfrac{L - S}{L + S}$
	L = Largest value	L = Upper limit of the highest class
	S = Smallest value	S = lower limit of the smallest class
Quartile Deviation	$Q.D. = \dfrac{Q_3 - Q_1}{2}$	$Q.D. = \dfrac{Q_3 - Q_1}{2}$
	Coefficient of $Q.D. = \dfrac{Q_3 - Q_1}{\quad}$	Coefficient of $Q.D. = \dfrac{Q_3 - Q_1}{\quad}$

Mathematical Property of Median. The sum of the deviations of a given set of observations taken from median, ignoring ± signs, is the least.

For example, the median of 2, 4, 6, 8, 10 is 6. The deviations from 6, ignoring signs, are 4, 2, 0, 2 and 4 and the sum is 12. This sum is smaller than the one obtained if deviations are taken from any other value. Thus if deviations are taken from 7, the values ignoring signs would be 5, 3, 1, 1 and 3 and the sum is 13.

3.11 OTHER PARTITION VALUES : QUARTILES, DECILES AND PERCENTILES

For a given set of data arranged in order of magnitude, we have seen that the median divides the data into two equal parts. There are as many numbers below the median as there are above it. There are several other values which divide the given data into a number of equal parts. These values, often referred to as *partition values*, are values below which a specific fraction or percentage of the observations in a given set must fall. Of special interest are those partition values which are commonly referred to as **quartiles, deciles and percentiles.** Let us first define a quartile.

Quartile. The values which divide the given data into four equal parts are called **quartiles.** Obviously, there are three quartiles which are labelled as Q_1, Q_2 and Q_3. The first quartile Q_1, also called **lower quartile,** is such that 25% of the data falls below it. Similarly, the second and third quartiles are such that 50% of the data falls below Q_2 and 75% falls below Q_3. Obviously Q_2 is the median. The third quartile, Q_3, is also called **upper quartile.**

Calculation of Quartiles - Individual Observations

For ungrouped data consisting of n observations (not necessarily all distinct), the calculation of kth quartile Q_k $(k = 1, 2, 3)$ involves the following steps :

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

Step 2. The value of kth quartile Q_k is given by

$$Q_k = \text{size of } \frac{k(n+1)}{4} \text{ th observation}$$

Thus,

$$Q_1 = \text{size of } \frac{(n+1)}{4} \text{ th observation,}$$

$$Q_2 = \text{size of } \frac{2(n+1)}{4} \text{ th observation}$$

and

$$Q_3 = \text{size of } \frac{3(n+1)}{4} \text{ th observation.}$$

Calculation of Quartiles - Discrete Series

In case of discrete frequency distribution where the variable X takes the values $X_1, X_2, \ldots, X_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ with $\Sigma f = N$, the calculation of kth quartile Q_k $(k = 1, 2, 3)$ involves the following steps :

Step 1. *Prepare the 'less than' cumulative frequency distribution.*

Step 2. *Find* $\frac{k(N+1)}{4}$

Step 3. *See the c.f. just greater than or equal to* $\frac{k(N+1)}{4}$ . . .

Step 4. *The value of X corresponding to the c.f. obtained in Step 3 gives the required value of* Q_k

Calculation of Quartiles - Continuous Series

In case of continuous frequency distribution, the calculation of Q_k ($k = 1, 2, 3$) involves the following steps :

Step 1. *Prepare the 'less than' cumulative frequency distribution.*

Step 2. *Find* $\frac{kN}{4}$, *where* $N = \Sigma f$ *is the total frequency.*

Step 3. *See the c.f. just greater than or equal to* $\frac{kN}{4}$.

Step 4. *Find the* Q_k *class, the class corresponding to c.f. obtained in Step 3.*

Step 5. *The value of* Q_k *is then obtained by using the following interpolation formula :*

$$Q_k = l + \frac{\frac{kN}{4} - C}{f} \times h,$$

where l = lower limit of Q_k class C = *c.f.* of the class preceding the Q_k class

f = frequency of the Q_k class, and h = size or width of Q_k class.

Deciles. The values which divide the given data into 10 equal parts are called **deciles.** Obviously, there are nine deciles which are labelled as $D_1, D_2, \ldots, D_9$. These values are such that 10% of the data falls below D_1, 20% falls below D_2, . . ., and 90% falls below D_9. Thus D_5 is the median.

Calculation of Deciles - Individual Observations

For ungrouped data consisting of n observations (not necessarily all distinct), the calculation of k*th* decile D_k ($k = 1, 2, \ldots, 9$) involves the following steps :

Step 1. *Arrange the given data in an ascending or descending order of magnitude.*

Step 2. *The value of kth decide* D_k *is given by*

$$D_k = \text{size of } \frac{k(n+1)}{10} th \text{ observation}$$

Thus,

$$D_1 = \text{size of } \frac{(n+1)}{10} th \text{ observation}$$

$$\ldots\ldots\ldots\ldots D_2 = \text{size of } \frac{2(n+1)}{10} th \text{ observation and so on.}$$

Calculation of Deciles - Discrete Series

In case of discrete frequency distribution where the variable X takes $X_1, X_2, \ldots, X_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ with $\Sigma f = N$, the calculation of k*th* decile involves the following steps :

Step 1. *Prepare the 'less than' cumulative frequency distribution.*

Step 2. *Find* $\frac{k(N+1)}{10}$

Step 3. *See the c.f. just greater than or equal to* $\frac{k(N+1)}{10}$

Step 4. *The value of X corresponding to the c.f. obtained in Step 3 gives the required value of* D_k

Calculation of Deciles - Continuous Series

In case of continuous frequency distribution, the calculation of D_k $(k = 1, 2, \ldots, 9)$ involves the following steps :

Step 1. *Prepare the 'less than' cumulative frequency distribution.*

Step 2. *Find* $\frac{kN}{10}$, *where* $N = \Sigma f$ *is the total frequency.*

Step 3. *See the c.f. just greater than or equal to* $\frac{kN}{10}$.

Step 4. *Find the* D_k *class, the class corresponding to c.f. obtained in Step 3.*

Step 5. *The value of* D_k *is then obtained by using the following interpolation formula :*

$$D_k = l + \frac{\frac{kN}{10} - C}{f} \times h,$$

where l = lower limit of D_k class C = c.f. of the class preceding the D_k class
f = frequency of the D_k class, and h = size or width of D_k class.

Percentiles. The values which divide the given data into 100 equal parts are called **percentiles.** Obviously, there are 99 percentiles which are labelled as $P_1, P_2, \ldots, P_{99}$. These values are such that 1% of the data falls below P_1, 2% falls below P_2, ..., and 99% falls below P_{99}. Thus P_{50} is the median.

Calculation of Percentiles - Individual Observations

For ungrouped data consisting of n observations (not necessarily all distinct), the calculation of k*th* percentile P_k $(k = 1, 2, \ldots, 99)$ involves the following steps :

Step 1. *Arrange the given data in an ascending or descending order of magnitude.*

Step 2. *The value of kth percentile* P_k *is given by*

$$P_k = \text{size of } \frac{k(n+1)}{100} \text{ } th \text{ observation}$$

In particular, P_1 = size of $\frac{(n+1)}{100}$ *th* observation

P_2 = size of $\frac{2(n+1)}{100}$ *th* observation and so on.

Calculation of Percentiles – Discrete Series

In case of discrete frequency distribution where the variable X takes the values $X_1, X_2, \ldots, X_n$ with respective frequencies $f_1, f_2, \ldots, f_n$ with $\Sigma f = N$, the calculation of k*th* percentiles involves the following steps :

Step 1. *Prepare the 'less than' cumulative frequency distribution.*

Step 2. *Find* $\frac{k(N+1)}{100}$, *where* $N = \Sigma f$ *is the total frequency.*

Step 3. *See the c.f. just greater than or equal to* $\frac{k(N+1)}{100}$.

Step 4. *The value of X corresponding to the c.f. obtained in Step 3 gives the required value of* P_k.

Calculation of Percentiles – Continuous Series

In case of continuous frequency distribution, the calculation of P_k $(k = 1, 2, \ldots, 99)$ involves the following steps :

Step 1. *Prepare the 'less than' cumulative frequency distribution.*

Step 2. *Find* $\frac{kN}{100}$ *where* $N = \Sigma f$ *is the total frequency.*

Step 3. *See the c.f. just greater than or equal to* $\frac{kN}{100}$.

Step 4. *Find the* P_k *class, the class corresponding to c.f. obtained in Step 3.*

Step 5. *The value of* P_k *is then obtained by using the following interpolation formula :*

$$P_k = l + \frac{\frac{kN}{100} - C}{f} \times h,$$

where l = lower limit of P_k class $\quad C$ = c.f. of the class preceding the P_k class

f = frequency of the P_k class, and $\quad h$ = size or width of P_k class.

EXAMPLE 54. *The marks obtained by 9 students in a test are 25, 20, 15; 45, 18, 7, 10, 38 and 12. Find the values of* $Q_1, Q_3, D_2, D_8, P_{45}$, *and* P_{70}, *for the data.*

SOLUTION. Arranging the marks in an increasing order of magnitude, we get

7 10 12 15 18 20 25 38 45

Here, $n = 9$

Computation of Q_1 : We have $\frac{n+1}{4} = \frac{10}{4} = 2.5$

Q_1 = size of $\left(\frac{n+1}{4}\right)$ *th* item = size of (2.5)*th* item

= 2nd item + 0.5 (3rd item – 2nd item) = 10 + 0.5 (12 – 10) = 11

Computation of Q_3. We have $\frac{3(n+1)}{4} = 7.5$

Q_3 = size of $\frac{3(n+1)}{4}$ *th* item = size of (7.5)*th* item

= 7th item + 0.5 (8th item – 7th item)

= 25 + 0.5 (38 – 25) = 25 + 6.5 = 31.5

Computation of D_2: We have $\frac{2(n+1)}{10} = 2$

$\therefore$ D_2 = size of $\frac{2(n+1)}{10}$ *th* item = size of 2nd item = 10

Computation of D_8: We have $\frac{8(n+1)}{10} = 8$

$\therefore$ D_8 = size of $\frac{8(n+1)}{10}$ *th* item = size of 8th item = 38

Computation of P_{45}: We have $\frac{45(n+1)}{100} = \frac{45\times10}{100} = 4.5$

$\therefore$ P_{45} = size of $\frac{45(n+1)}{100}$ *th* item = size of 4.5th item

= 4th item + 0.5 (5th item – 4th item) = 15 + 0.5 (18 – 15) = 16.5

Computation of P_{70}: We have $\frac{70(n+1)}{100} = \frac{70\times10}{100} = 7$

P_{70} = size of 7*th* item = 25

EXAMPLE 55. *Calculate the values of* Q_1, Q_3, D_4 *and* P_{65} *from the following data :*

X:	10	5	7	11	8
f:	15	20	15	18	12

SOLUTION. We prepare the following table.

CALCULATION OF MEDIAN

X		
5	20	20
7	15	35
8	12	47
10	15	62
11	18	80
	$N = \sum f = 80$	

Computation of Q_1: We have $\frac{N+1}{4} = 20.25$; and the *c.f.* just greater than or equal to 20.25 is 35. The corresponding value of X is 7.

$$\therefore \quad Q_1 = 7$$

Computation of Q_3: We have $\frac{3(N+1)}{4} = 60.75$; and the *c.f.* just greater than or equal to 60.75 is 62. The corresponding value of X is 10.

$$\therefore \quad Q_3 = 10$$

Computation of D_4: We have $\frac{4(N+1)}{10} = 32.4$; and the *c.f.* just greater than or equal to 32.4 is 35. The corresponding value of X is 7.

$$\therefore \quad D_4 = 7$$

Computation of P_{65}: We have $\frac{65(N+1)}{100} = 52.65$; and the *c.f.* just greater than or equal to 52.65 is 62. The corresponding value of X is 10.

$$\therefore \quad P_{65} = 10.$$

EXAMPLE 56. *From the following data, calculate the median and the first and third quartile wages :*

Daily wages (Rs.)	No. of workers	Daily wages (Rs.)	No. of workers
30 – 32	2	40 – 42	62
32 – 34	9	42 – 44	39
34 – 36	25	44 – 46	20
36 – 38	30	46 – 48	11
38 – 40	49	48 – 50	3

SOLUTION.

CALCULATION OF QUARTILES

Daily wages (Rs.)	No. of workers (f)	Less than c.f.
30 – 32	2	2
32 – 34	9	11
34 – 36	25	36
36 – 38	30	66
38 – 40	49	115
40 – 42	62	177
42 – 44	39	216
44 – 46	20	236
46 – 48	11	247
48 – 50	3	250 = N

Calculation of Median : Median is the size of $\left(\frac{250}{2}\right)^{th}$ or 125th item which lies in the class 40 – 42. It is given by

$$Md = l + \frac{\frac{N}{2} - C}{f} \times h = 40 + \frac{125 - 175}{62} \times 2 = 40 + 0.32 = 40.32$$

Calculation of Lower Quartile : Lower quartile is the size of $\left(\frac{250}{4}\right)^{th}$ or 62.5th item which lies in the class 36 – 38. It is given by

$$Q_1 = l + \frac{\frac{N}{4} - C}{f} \times h = 36 + \frac{62.5 - 36}{30} \times 2 = 36 + 1.77 = 37.77$$

Calculation of Upper Quartile : Upper quartile is the size of $\left(\frac{3N}{4}\right)^{th}$ item or 187.5th item which lies in the class 42 – 44. It is given by

$$Q_3 = l + \frac{\frac{3N}{4} - C}{f} \times h = 42 + \frac{187.5 - 177}{39} \times 2 = 42 + 0.54 = 42.54$$

EXAMPLE 57. *Find the 45th and 57th percentiles for the following data on marks obtained by 100 students*

Marks :	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
No. of Students :	10	20	20	15	15	20

[C.A. Foundation, May 1996]

SOLUTION. **CALCULATION OF PARTITION VALUES**

Marks	*No. of Students (f)*	*Less than c.f.*
20 – 25	10	10
25 – 30	20	30
30 – 35	20	50
35 – 40	15	65
40 – 45	15	80
45 – 50	20	100
	$N = \Sigma f = 100$	

Calculation of P_{45} : We have $\frac{45N}{100} = 45$. The *c.f.* just greater than or equal to 45 is 50.

Hence the corresponding class 30 – 35 contains P_{45} which is given by

$$P_{45} = l + \frac{\frac{45N}{100} - C}{f} \times h$$

where $l = 30$, $C = 30$, $f = 20$ and $h = 5$.

$$P_{45} = 30 + \frac{45-30}{20} \times 5 = 30 + 3.75 = 33.75$$

Calculation of P_{57} : We have $\frac{57N}{100} = 57$. The *c.f.* just greater than or equal to 57 is 65.

Hence the corresponding class 35 – 40 contains P_{57} which is given by

$$P_{57} = l + \frac{\frac{57N}{100} - C}{f} \times h$$

where $l = 35$, $C = 50$, $f = 15$ and $h = 5$.

$$P_{57} = 35 + \frac{57-50}{15} \times 5 = 35 + 2.34 = 37.34.$$

EXAMPLE 58. *From the following data, calculate the 7th decile and 60th percentile :*

Wages	*30 – 40*	*40 – 50*	*50 – 60*	*60 – 70*	*70 – 80*	*80 – 90*	*90 – 100*
No. of Workers:	*1*	*3*	*11*	*21*	*43*	*32*	*9*

SOLUTION.

CALCULATION OF D_7 AND P_{60}

Wages	*No. of Workers (f)*	*Less than c.f.*
30 – 40	1	1
40 – 50	3	4
50 – 60	11	15
60 – 70	21	36
70 – 80	43	79
80 – 90	32	111
90 – 100	9	120
	$N = \Sigma f = 120$	

Calculation of D_7 : We have $\frac{7N}{10} = 84$. The *c.f.* just greater than or equal to 84 is 111.

Hence the corresponding class 80 – 90 contains D_7 which is given by

$$D_7 = l + \frac{\frac{7N}{10} - C}{f} \times h,$$

where $l = 80$, $C = 79$, $f = 32$ and $h = 10$.

$$\therefore \quad D_7 = 80 + \frac{84-79}{32} \times 10 = 80 + 1.56 = \text{Rs. } 81.56.$$

Calculation of P_{60} : We have $\frac{60N}{100} = 72$. The *c.f.* just greater than or equal to 72 is 79.

Hence the corresponding class 70 – 80 contains P_{60} which is given by

SOLUTION

CALCULATION OF VARIOUS PARTITION VALUES

Marks	No. of Students (f)	c.f. (less than)
below 10	8	8
10 – 20	10	18
20 – 40	22	40
40 – 60	25	65
60 – 80	10	75
above 80	5	80
	$N = \Sigma f = 80$	

Calculation of Q_1 : We have $\frac{N}{4} = 20$. The *c.f.* just greater than or equal to 20 is 40.

Hence the corresponding class 20 – 40 contains Q_1 which is given by

$$Q_1 = l + \frac{\frac{N}{4} - C}{f} \times h,$$

where $l = 20$, $C = 18$, $f = 22$ and $h = 20$.

$$Q_1 = 20 + \frac{20 - 18}{22} \times 20 = 20 + 1.82 = 21.82.$$

Calculation of Q_3 : We have $\frac{3N}{4} = 60$. The *c.f.* just greater than or equal to 60 is 65.

Hence the corresponding class 40 – 60 contains Q_3 which is given by

$$Q_3 = l + \frac{\frac{3N}{4} - C}{f} \times h,$$

where $l = 40$, $C = 40$, $f = 25$ and $h = 20$.

$$Q_3 = 40 + \frac{60 - 40}{25} \times 20 = 40 + 16 = 56.$$

Calculation of D_2 : We have $\frac{2N}{10} = 16$. The *c.f.* just greater than or equal to 16 is 18.

Hence the corresponding class 10 – 20 contains D_2 which is given by

Unit 5
Regression and correlation

CHAPTER 7

Regression Analysis

OBJECTIVES

After studying the material in this chapter, you should be able to:

- *Understand the concepts of linear regression analysis.*
- *Compute regression coefficients and regression lines.*
- *Know the properties of regression coefficients.*
- *Apply the regression analysis to estimate or predict the values of a dependent variable from known values of an independent variable.*
- *Determine standard error of estimate.*
- *Distinguish between correlation and regression.*

7.1 INTRODUCTION

From our discussion in the previous chapter, we were able to find a way of determining whether or not a relationship existed between two variables. If such a relationship can be expressed by a mathematical formula, we will then be able to use it for the purpose of making predictions. The reliability of any prediction will, of course, depend on the strength of the relationship between the variables included in the formula. *Regression analysis* attempts to establish the "nature of relationship" between variables that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting. A mathematical equation that allows us to predict value of one variable from known values of one or more other variables is called a *regression equation.* The variable whose value is to be predicted is called the *dependent variable* or *explained variable.* The variables which are used to predict the values of a dependent variable are called *independent variables* or *explanatory variables.* The regression analysis confined to the study of only two variables, a dependent variable and an independent variable, is called *simple regression analysis.* When the relationship between the dependent variable and the independent variable is linear, the technique for prediction is called *simple linear regression.*

7.2

7.2 MEANING OF REGRESSION

The literal dictionary meaning of the word 'regression' is 'stepping back' or 'moving backward' or 'returning to average value. It was Sir Francis Galton who first used the term regression as a statistical concept in 1877. He studied the relationship between the heights of fathers and their sons, and arrived at some interesting conclusions, which are described below :

(i) Tall fathers have tall sons and short fathers have short sons.

(ii) The average height of the sons of short fathers is more than the average height of their fathers.

(iii) The average height of the sons of tall fathers is less than the average heights of their fathers.

In other words, Galton's studies revealed that the off springs of abnormally tall or short parents tend to revert or step back to the average height of the population, a phenomenon which Galton described as 'regression to mediocrity'. Regression thus implies going back or returning towards the average. Galton used the term regression as a statistical technique to predict one variable (the height of children) from another variable (the height of parents).

But today the word regression as used in statistics has a much wider perspective without any reference to biometry. Regression analysis, in the general sense, means the estimation or prediction of the unknown value of one variable from the known values of one or more other variables. The variable whose value is to be predicted is called the *dependent variable* or *explained variable*. The variables which are used to predict the values of a dependent variable are called *independent variables* or *explanatory variables*. In the study done by Galton, mentioned above, the height of the parents was the independent variable and the height of children was the dependent variable.

In the following we shall give some important definitions of the term regression.

1. *Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.* — M.M. Blair

2. *One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis.* — Taro Yamane

3. *The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process.* — Morris Hamburg

4. *Regression analysis attempts to establish the nature of the relationship between variables — that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting.* — Ya Lun Chou

7.3 LINES OF REGRESSION – THE LEAST SQUARES APPROACH

In this section we consider the problem of estimating or predicting the values of a

dependent variable from known values of an independent variable. To make such a prediction, suppose that we have a bivariate data that consists of n pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ on two quantitative variables X and Y. We assume that X and Y are approximately linearly related so that the data points follow closely a straight line on a scatter diagram (See Fig. 7.1). On this basis we may fit "by eye" a straight line that approximates the given data. This line can then be used to predict a value of Y for a given value of X. Unfortunately, determining a line "by eye" is not very objective, because so many lines exist for a given data (unless the correlation is equal to plus or minus 1). We therefore need to establish a criterion for selecting a line of "best fit". A frequently used criterion is the **least squares criterion.** *According to the least squares criterion, the line of "best fit" is the one that minimizes the sum of the squares of the vertical distances from the observed points to the line.*

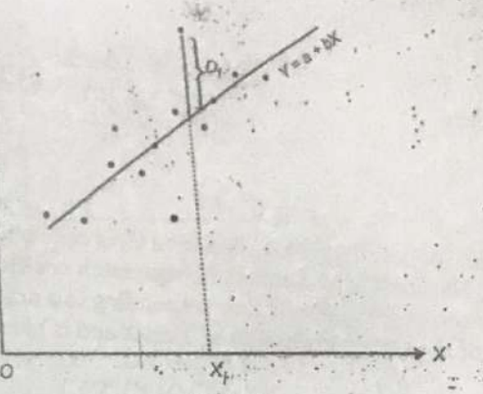


Fig. 7.1

Thus if a line of best fit approximating the given data has the equation

$$Y = a + bX,$$

then the method of least squares requires that we must determine constants a and b so as to minimize

$$S = (Y_1 - a - bX_1)^2 + (Y_2 - a - bX_2)^2 + \ldots + (Y_n - a - bX_n)^2$$
$$= D_1^2 + D_1^2 + \ldots + D_n^2$$

where $D_i = Y_i - a - bX_i$ represents the vertical deviation of the ith observed point from the line of best fit, as indicated in Fig. 7.1. The determination of a and b so as to minimize S can be accomplished by means of differential calculus. We omit the proof and state the final two equations which are used to determine the values of a and b. These equations, known as the normal equations for estimating a and b, are given by

$$\sum Y = na + b\sum X \quad \ldots(1)$$

$$\sum XY = a\sum X + b\sum X^2 \quad \ldots(2)$$

Solving Eqs. (1) and (2) simultaneously for a and b, we obtain

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

143

2.4

and $a = \bar{Y} - b\bar{X}$

Hence the line of best fit approximating the n pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots$ (X_n, Y_n) is

$$Y = a + bX \quad \ldots (3)$$

where

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} \quad \ldots (4)$$

$$a = \bar{Y} - b\bar{X} \quad \ldots (5)$$

The line of best fit given by Equation (3) is called the **least squares line of regression of Y on X**. The constant b is called the **regression coefficient of Y on X** and is denoted by b_{YX}. It measures the change in Y corresponding to a unit change in X. Thus b_{YX} represents the slope of the line of regression of Y on X and is given by

$$b_{YX} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} \quad \ldots (6)$$

The formula (5) estimating the value of "a" clearly shows that the line of regression of Y on X passes through the point $(\bar{X}, \bar{Y})$ and hence the equation of the line of regression of Y on X can also be written as

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \quad \ldots (7)$$

This equation is then used to estimate a value of Y for a given value of X. On the other hand, if we wish to estimate a value of X for a given value of Y, we have to obtain the **regression line of X on Y**:

$$X = c + dY$$

where the constants c and d are determined according to the least squares criterion. The two normal equations for estimating c and d are given by

$$\sum X = nc + d\sum Y$$

$$\sum XY = c\sum Y + d\sum Y^2$$

Solving these normal equations simultaneously for c and d, we obtain

$$d = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} \quad \ldots (8)$$

$$c = \bar{X} - d\bar{Y} \qquad \ldots(9)$$

The constant d is called the **regression coefficient of X on Y** and is denoted by b_{XY}. It measures the change in X corresponding to a unit change in Y.

Clearly, $\frac{1}{b_{XY}}$ represents the slope of the regression line of X on Y. Further, the Formula (9) estimating the value of "c" clearly shows that the line of regression of X on Y passes through the point $(\bar{X}, \bar{Y})$ and hence the equation of the line of regression of X on Y can be written as

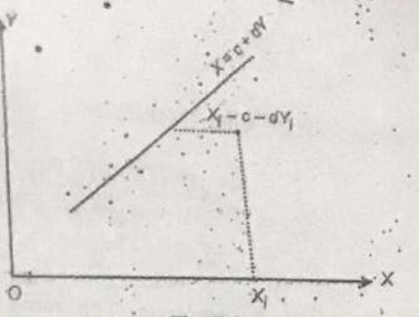


Fig. 7.2

$$Y - \bar{Y} = \frac{1}{b_{XY}}(X - \bar{X})$$

or,
$$X - \bar{X} = b_{XY}(Y - \bar{Y}), \qquad \ldots(10)$$

where
$$b_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} \qquad \ldots(11)$$

REMARK 1. It may be remarked that there are always two lines of regression, one of Y on X and the other X on Y. The regression line of Y on X is used to estimate or predict the value of Y from known values of X, *i.e.*, Y is a dependent variable and X is an independent variable. According to the least squares criterion, the line of regression of Y on X is the line of "best fit" in the sense that it minimizes the sum of the squares of the vertical distances from the observed points to the line (see Fig. 7.1). However, if we want to estimate or predict the value of X from known values of Y, we will use regression line of X on Y which is the line of "best fit" in the sense that it minimizes the sum of the squares of the horizontal distances from the observed points to the line (see Fig. 7.2). The two regression equations are not reversible because of the simple reason that the basis and assumptions for deriving these equations are quite different. However, it may be remarked that in case of perfect correlation (positive or negative), the two regression lines would coincide.

REMARK 2. Since the two lines of regression pass through the point $(\bar{X}, \bar{Y})$, the mean values $(\bar{X}, \bar{Y})$ can be obtained as the point of intersection of the two regression lines.

EXAMPLE 1. *Calculate the regression coefficients from the following information:*

$\sum X = 50, \quad \sum Y = 30, \quad \sum XY = 1000, \quad \sum X^2 = 3000, \quad \sum Y^2 = 1800, \quad N = 10$

[C.A. Foundation, Nov. 1995]

143

Regress...
Obtain
SOLUTION.

Regression coefficient of X on Y:

$$b_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}} = \frac{1000 - \frac{(50)(30)}{10}}{1800 - \frac{(30)^2}{10}} = \frac{1000-150}{1800-90} = 0.497$$

Regression coefficient of Y and X:

$$b_{YX} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}} = \frac{1000 - \frac{(50)(30)}{10}}{3000 - \frac{(50)^2}{10}} = \frac{850}{2750} = 0.309.$$

EXAMPLE 2. *Find both the regression equations from the following:*

$\sum X = 60, \ \sum Y = 40, \ \sum X^2 = 4160, \ \sum Y^2 = 1720, \ \sum XY = 1150 \ \ N = 10$

[C.A. Foundation, May 1998]

Regre

SOLUTION. We have $b_{YX} = \dfrac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \dfrac{10\times1150 - 60\times40}{10\times4160 - (60)^2}$

$$= \frac{115-24}{416-36} = \frac{91}{380} = 0.239$$

or,
i.e.,
⇒

and $b_{XY} = \dfrac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \dfrac{10\times1150 - 60\times40}{10\times1720 - (40)^2}$

$$= \frac{115-24}{172-16} = \frac{91}{156} = 0.583$$

$$\bar{X} = \frac{\sum X}{N} = \frac{60}{10} = 6 \quad \text{and} \quad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{10} = 4.$$

EXAMPLE
SOLUTI

Regression Equation of Y on X is:

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 4 = 0.239(X - 6)$

i.e., $Y - 4 = 0.239X - 1.434$

⇒ $Y = 0.239X + 2.566$

Regression Equation of X on Y is:

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 6 = 0.583(Y - 4)$

i.e., $X - 6 = 0.583Y - 2.332$

⇒ $X = 0.583Y + 3.668$

EXAMPLE 3. *In the estimation of regression equation of two variables X and Y, the following results were obtained:*

$\sum X = 900, \ \sum Y = 700, \ \sum X^2 = 6360, \ \sum Y^2 = 2860, \ \sum XY = 3900, \ N = 10$

Obtain two regression equations.

[Delhi Univ. B.Com. (H) 2008]

SOLUTION. We have $b_{YX} = \dfrac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} = \dfrac{(10\times 3900) - (900\times 700)}{10\times 6360 - (900)^2}$

$$= \frac{39000 - 630000}{63600 - 810000} = \frac{-591000}{-746400} = 0.792$$

$$b_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2} = \frac{(10\times 3900) - (900\times 700)}{10\times 2860 - (700)^2}$$

$$= \frac{39000 - 630000}{28600 - 490000} = \frac{-591000}{-461400} = 1.28$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{900}{10} = 90 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{700}{10} = 70$$

Regression Equation of Y on X is :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 70 = 0.792\,(X - 90)$

i.e., $Y - 70 = 0.792X - 71.28$

$\Rightarrow$ $Y = 0.792X + 1.28$

Regression Equation of X on Y is :

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 90 = 1.28\,(Y - 70)$

i.e., $X - 90 = 1.28Y - 89.6$

$\Rightarrow$ $X = 1.28Y + 0.4$

EXAMPLE 4. *Find the two lines of regression on the basis of the following data :*

X :	1	2	3	4	5	6	7
Y :	2	4	7	6	5	6	5

SOLUTION.

CALCULATION FOR REGRESSION LINES

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	4	4	16	8
3	7	9	49	21
4	6	16	36	24
5	5	25	25	25
6	6	36	36	36
7	5	49	25	35
$\Sigma X = 28$	$\Sigma Y = 35$	$\Sigma X^2 = 140$	$\Sigma Y^2 = 191$	$\Sigma XY = 151$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{28}{7} = 4 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{35}{7} = 5$$

$$b_{YX} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{7\times151 - 28\times35}{7\times140 - (28)^2} = \frac{151-140}{140-112} = \frac{11}{28} = 0.39$$

$$b_{XY} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} = \frac{7\times151 - 28\times35}{7\times191 - (35)^2}$$

$$= \frac{151-140}{191-175} = \frac{11}{16} = 0.69$$

Regression Equation of Y on X is :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 5 = 0.39(X - 4)$

i.e., $Y = 3.44 + 0.39X$

Regression Equation of X on Y is :

$$\therefore \quad X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 4 = 0.69(Y - 5)$

i.e., $X = 0.55 + 0.69Y$

EXAMPLE 5. *The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age :*

Age of Cars (in years) :	2	4	6	8
Maintenance cost (in Rs. hundred) :	10	20	25	30

Also estimate the annual maintenance cost for a ten year old car.

SOLUTION. Let X denote the age of car and Y denote its annual maintenance cost. Then it is required to find the regression equation of Y on X.

CALCULATION FOR REGRESSION EQUATION OF Y ON X

X	Y	X^2	Y^2	XY
2	10	4	100	20
4	20	16	40	80
6	25	36	625	150
8	30	64	900	240
$\sum X = 20$	$\sum Y = 85$	$\sum X^2 = 120$	$\sum Y^2 = 2025$	$\sum XY = 490$

$$\bar{X} = \frac{\sum X}{n} = \frac{20}{4} = 5 \quad \text{and} \quad \bar{Y} = \frac{\sum Y}{n} = \frac{85}{4} = 21.25$$

$$b_{YX} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{4\times490 - (20)(85)}{4\times120 - (20)^2}$$

$$= \frac{1960 - 1700}{480 - 400} = \frac{260}{80} = 3.25$$

Regression Equation of Y on X is given by

$$Y-\bar{Y} = b_{YX}(X-\bar{X}) \quad \text{or,} \quad Y-21.25 = 3.25\,(X-5)$$

i.e., $Y-21.25 = 3.25X-16.25 \Rightarrow Y = 5+3.25X$ ...(1)

Substituting $X = 10$ in (1), the estimated annual maintenance cost for a ten-year old car is :

$$Y = 5+3.25\times 10 = 37.5 \text{ Rs. hundred or Rs. } 3750.$$

EXAMPLE 6. *A furniture retailer in a certain locality is interested in studying the relationship that exists between the number of building permits issued in that locality in the past years and the volume of sales in those years. The data for sales (Y, in thousand rupees) and the number of building permits issued (X, in hundreds) for the last 10 years are collected to give the following results :*

$$\Sigma X = 200,\ \Sigma Y = 2200,\ \Sigma XY = 45800,\ \Sigma X^2 = 4600,\ \Sigma Y^2 = 4,90,400$$

Find :

(i) the level of sales expected in a year when 2000 building permits are to be issued, and

(ii) the change in sales expected for every increase of 100 building permits.

[Delhi Univ. B.A. (Econ. Hons.) 2001]

SOLUTION. Since sales (in thousand rupees) are represented by Y and the number of building permits issued (in hundreds) by X, we need the regression equation of Y on X to estimate the level of sales, which is given by

$$Y-\bar{Y} = b_{YX}(X-\bar{X}),$$

where $$\bar{X} = \frac{\Sigma X}{n} = \frac{200}{10} = 20, \qquad \bar{Y} = \frac{\Sigma Y}{n} = \frac{2200}{10} = 220,$$

and $$b_{YX} = \frac{n\Sigma XY-(\Sigma X)(\Sigma Y)}{n\Sigma X^2-(\Sigma X)^2} = \frac{10\times 45800-(200)(2200)}{10\times 4600-(200)^2}$$

$$= \frac{458000-440000}{46000-40000} = \frac{18000}{6000} = 3$$

$\therefore$ Regression equation of Y on X is : $Y-220 = 3(X-20)$ or $Y = 3X+160$ ...(1)

(i) Since building permits have been expressed in hundreds, 2000 building permits will mean $X = 20$. Thus the level of sales that can be expected in a year when 2000 building permits are to be issued is given by substituting $X = 20$ in Eq. (1) :

$Y = 3\times 20+160 = 220$ thousand rupees $=$ Rs. 2,20,000.

(ii) For every increase of 100 building permits, the expected change in sales is given by the slope $b_{YX} = 3$ thousand rupees $=$ Rs. 3000.

7.4 REGRESSION COEFFICIENTS – SOME FORMULAS

In this section we shall derive some more formulas for regression coefficients.

I. Formulas for Regression Coefficients in terms of Covariance and Variances

By definition, the regression coefficient of Y on X is given by

$$b_{YX} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \quad \ldots (1)$$

Similarly, the regression coefficient of X on Y is given by

$$b_{XY} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}} \quad \ldots (2)$$

The reader may also recall that the covariance between X and Y is given by

$$\text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \left(\frac{\Sigma X}{n}\right)\left(\frac{\Sigma Y}{n}\right) \quad \ldots (3)$$

Further, the variances of X- and Y-values are respectively given by

$$\sigma_X^2 = \text{Var}(X) = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2 \quad \ldots (4)$$

$$\sigma_Y^2 = \text{Var}(Y) = \frac{\Sigma Y^2}{n} - \left(\frac{\Sigma Y}{n}\right)^2 \quad \ldots (5)$$

From (1), (3) and (4), we find that

$$b_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} \quad \ldots (6)$$

Similarly, from (2), (3) and (5), we find that

$$b_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_Y^2} \quad \ldots (7)$$

Thus we have the following formulas :

Formulas 1 : $b_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X^2}$ and $b_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_Y^2}$

II. Formulas for Regression Coefficients in terms of Deviations of X- and Y-values from their respective means

By definition, the covariance between X and Y is given by

$$\text{Cov}(X, Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n}$$

Further, the variances of X- and Y-values are respectively given by

$$\sigma_X^2 = \frac{\Sigma(X-\bar{X})^2}{n} \quad \text{and} \quad \sigma_Y^2 = \frac{\Sigma(Y-\bar{Y})^2}{n}$$

Thus using Formulas *I*, we obtain

$$b_{YX} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$$

and

$$b_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_Y^2} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2}$$

If we let x and y denote the deviations of X- and Y- values from their respective means, *i.e.*, $x = X - \bar{X}$ and $y = Y - \bar{Y}$, then the above formulas for regression coefficients can be put in the following forms :

Formulas 2 : $b_{YX} = \frac{\Sigma xy}{\Sigma x^2}$ and $b_{XY} = \frac{\Sigma xy}{\Sigma y^2}$, where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

III. Formulas for Regression Coefficients in terms of r, σ_X and σ_Y

We know that the coefficient of correlation r is given by

$$r = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

$$\therefore \quad b_{YX} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} = \frac{\sigma_Y}{\sigma_X} \cdot \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sigma_Y}{\sigma_X} \cdot r$$

i.e.,

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

Similarly,

$$b_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_Y^2} = \frac{\sigma_X}{\sigma_Y} \cdot \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sigma_X}{\sigma_Y} \cdot r$$

151

i.e., $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

Thus we have the following formulas :

Formulas 3 : $b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ and $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

REMARK : The two regression equations can be put in the following forms :

Y on X: $$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \quad \ldots(1)$$

X on Y: $$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \quad \ldots(2)$$

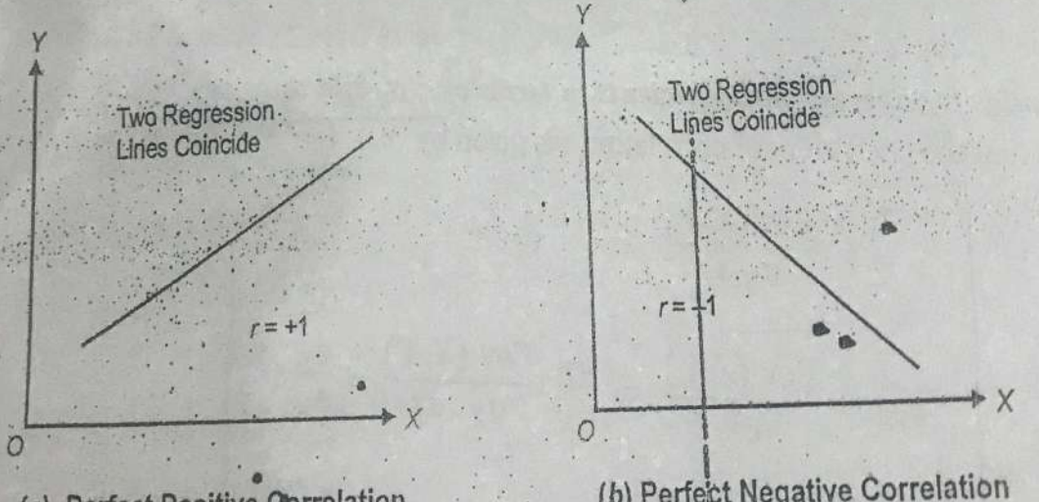
(*i*) If there is a perfect correlation between the two variables, *i.e.*, $r = \pm 1$, the regression equation of Y on X becomes

$$Y - \bar{Y} = \pm\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \quad \text{or,} \quad \frac{Y - \bar{Y}}{\sigma_Y} = \pm\left(\frac{X - \bar{X}}{\sigma_X}\right) \quad \ldots(3)$$

Similarly, the regression equation of X on Y becomes

$$X - \bar{X} = \pm\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \quad \text{or,} \quad \frac{Y - \bar{Y}}{\sigma_Y} = \pm\left(\frac{X - \bar{X}}{\sigma_X}\right) \quad \ldots(4)$$

which is same as (3). Hence in the case of perfect correlation, the two regression lines coincide (see **Fig. 7.3**).



(*a*) Perfect Positive Correlation (*b*) Perfect Negative Correlation

Fig. 7.3.

(*ii*) If $r = 0$, *i.e.*, if X and Y are uncorrelated, the two regression equations reduce to $Y = \bar{Y}$ and $X = \bar{X}$, and hence they are perpendicular to each other (see **Fig. 7.4**).

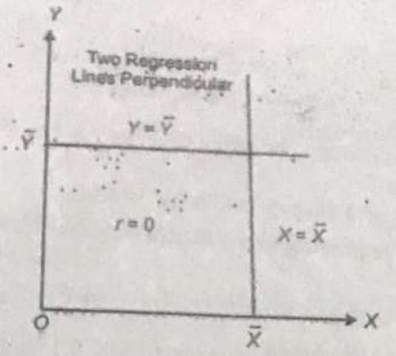


Fig. 7.4. No Correlation

EXAMPLE 7. *Find the regression coefficients b_{YX} and b_{XY} of Y on X and X on Y respectively, if standard deviations of X and Y are 4 and 3 respectively, and coefficient of correlation between X and Y is 0.8.*

SOLUTION. We are given the following:

$$\sigma_X = 4, \quad \sigma_Y = 3 \quad \text{and} \quad r = 0.8$$

To find b_{YX} and b_{XY}, we make use of the following formulas :

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad \text{and} \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

Substituting the given values, we obtain

$$b_{YX} = 0.8\times\frac{3}{4} = 0.6 \quad \text{and} \quad b_{XY} = 0.8\times\frac{4}{3} = 1.07.$$

EXAMPLE 8. *Compute the two lines of regression on the basis of the following information :*

	X	Y
Mean	40	45
Standard Deviation	10	9

Karl Pearson's coefficient correlation between X and Y = 0.50. Also estimate the value of Y for X = 48 using the appropriate regression equation. [C.A. Foundation, Dec. 1993]

SOLUTION. $\bar{X} = 40, \quad \bar{Y} = 45, \quad \sigma_X = 10, \quad \sigma_Y = 9$ and $r = 0.5$

$$\therefore \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.5\times\frac{9}{10} = 0.45$$

and

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.5\times\frac{10}{9} = 0.556$$

Regression line of Y on X :

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \quad \text{or} \quad Y - 45 = 0.45\,(X - 40)$$

$\Rightarrow \quad Y = 0.45X + 27$

Regression line of X on Y :

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad \text{or} \quad X - 40 = 0.556\,(Y - 45)$$

$\Rightarrow \quad X = 0.556Y + 14.98$

The appropriate regression line for estimating the value of Y for a given value of X is $Y = 0.45X + 27$, the regression line of Y on X.

Hence if $X = 48$, the estimated value of $Y = 0.45 \times 48 + 27 = 48.6$.

EXAMPLE 9. *The following data are given regarding expenditure on advertising and sales of a particular firm :*

	Adv. Expenditure (X) (Rs. lakhs)	Sales (Y) (Rs. lakhs)
Mean	10	90
Standard deviation	3	12
Correlation coefficient	0.8	

(i) Calculate the regression equation of Y on X

(ii) Estimate the advertisement expenditure required to attain a sales target of Rs. 120 lakhs. **[C.A. Foundation, Nov. 1999]**

SOLUTION. We are given the following information :

$$\bar{X} = 10, \quad \bar{Y} = 90, \quad \sigma_X = 3, \quad \sigma_Y = 12 \quad \text{and} \quad r = 0.8.$$

$$\therefore \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.8 \times \frac{12}{3} = 3.2 \quad \text{and} \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.8 \times \frac{3}{12} = 0.2$$

Regression line of Y on X :

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \quad \text{or,} \quad Y - 90 = 3.2\,(X - 10) \Rightarrow Y = 3.2X + 58$$

Regression line of X on Y :

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad \text{or,} \quad X - 10 = 0.2\,(Y - 90) \Rightarrow X = 0.2Y - 8$$

$\therefore$ The advertisement expenditure required to attain a sales target of Rs. 120 lakhs is given by :

$$X = 0.2 \times 120 - 8 = \text{Rs. } 16 \text{ lakhs.}$$

EXAMPLE 10. *Given :*

	X series	Y series
Mean	18	100
Standard Deviation	14	20

Coefficient of correlation between X and Y series = 0.8. Find the most probable value of Y if X is 70, and most probable value of X if Y is 90. **[Delhi Univ. B.Com. (H) 2005 (C.C)]**

SOLUTION. We are given the following information :

$$\bar{X} = 18, \quad \bar{Y} = 100, \quad \sigma_X = 14, \quad \sigma_Y = 20 \quad \text{and} \quad r = 0.8$$

$$\therefore \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.8\times\frac{20}{14} = 1.14 \quad \text{and} \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.8\times\frac{14}{20} = 0.56$$

Regression Equations

Regression Equation of Y on X is :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 100 = 1.14\,(X - 18)$

i.e., $Y - 100 = 1.14X - 20.52$

$\Rightarrow \quad Y = 1.14X + 79.48 \quad \ldots (1)$

Regression Equation of X on Y is :

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 18 = 0.56\,(Y - 100)$

i.e., $X - 18 = 0.56Y - 56$

$\Rightarrow \quad X = 0.56Y - 38 \quad \ldots (2)$

To find the probable value of Y if X is 70, we use regression line of Y on X. Thus substituting $X = 70$ in (1), we get

$$Y = 1.14\,(70) + 79.48 = 159.28$$

To find the probable value of X if Y is 90, we use regression line of X on Y. Thus substituting $Y = 90$ in (2), we get

$$X = 0.56\,(90) - 38 = 12.4.$$

EXAMPLE 11. *The following data relate to marks in Advanced Accounts and Business Statistics in B.Com. (Hons.) 1st year Examination of a particular year in Delhi University:*

Mean Marks in Advanced Accounts	= 30
Mean Marks in Business Statistics	= 35
Standard Deviation of Marks in Advanced Accounts	= 10
Standard Deviation of Marks in Business Statistics	= 7
Coefficient of correlation between the Marks of Advanced Accounts and Business Statistics	= 0.8

Form the two regression lines and calculate the expected marks in Advanced Accounts if the marks secured by a student in Business Statistics are 40.

[Delhi Univ. B.Com. (H) 1987]

SOLUTION. Let X denote the marks in Advanced Accounts and Y denote the marks in Business Statistics. Then we are given:

$$\bar{X} = 30, \quad \bar{Y} = 35, \quad \sigma_X = 10, \quad \sigma_Y = 7 \quad \text{and} \quad r = 0.8$$

$$\therefore \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.8\times\frac{10}{7} = 1.143$$

and

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = 0.8\times\frac{7}{10} = 0.56$$

Regression Equations

Regression Equation of X on Y is :

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 30 = 1.143\,(Y - 35)$

i.e., $X - 30 = 1.143Y - 40$

$\Rightarrow$ $X = 1.143Y - 10$ ... (1)

Regression Equation of Y on X is :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 35 = 0.56\,(X - 30)$

i.e., $Y - 35 = 0.56X - 16.8$

$\Rightarrow$ $Y = 0.56X + 18.2$ ... (2)

To find the expected marks in Advanced Accounts (X) if the marks in Business Statistics (Y) are 40, we use regression line of X on Y. Thus substituting $Y = 40$ in (1), we get

$$X = 1.143 \times 40 - 10 = 35.72.$$

EXAMPLE 12. *You are given the data relating to purchases and sales. Obtain the two regression equations and estimate the likely sales when the purchases equal 100.*

Sales	:	91	97	108	121	67	124	51	73	111	57
Purchases	:	71	75	69	97	70	91	39	61	80	47

SOLUTION.

CALCULATIONS FOR REGRESSION LINES

Sales X	Purchases Y	$X - \bar{X}$ x	$Y - \bar{Y}$ y	x^2	y^2	xy
91	71	1	1	1	1	1
97	75	7	5	49	25	35
108	69	18	−1	324	1	−18
121	97	31	27	961	729	837
67	70	−23	0	529	0	0
124	91	34	21	1156	441	714
51	39	−39	−31	1521	961	1209
73	61	−17	−9	289	81	153
111	80	21	10	441	100	210
57	47	−33	−23	1089	529	759
ΣX = 900	ΣY = 700			Σx^2 = 6360	Σy^2 = 2868	Σxy = 3900

$$\bar{X} = \frac{\Sigma X}{n} = \frac{900}{10} = 90 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{700}{10} = 70$$

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{3900}{6360} = 0.61 \quad \text{and} \quad b_{XY} = \frac{\Sigma xy}{\Sigma y^2} = \frac{3900}{2868} = 1.36$$

Regression Equations

Regression Equation of Y on X is :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 70 = 0.61\,(X - 90)$

i.e., $Y - 70 = 0.61X - 54.9$

$\Rightarrow$ $Y = 0.61X + 15.1$ ... (1)

Regression Equation of X on Y is :

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 90 = 1.36\,(Y - 70)$

i.e., $X - 90 = 1.36Y - 95.2$

$\Rightarrow$ $X = 1.36Y - 5.2$ ... (2)

To estimate the likely sales when the purchases equal 100, we use the regression line of X on Y. Thus substituting $Y = 100$ in (2), we get $X = 130.8$.

EXAMPLE 13. *The following data give the aptitude test scores and productivity indices of 10 workers selected at random :*

Aptitude scores (X)	:	60	62	65	70	72	48	53	73	65	82
Productivity index (Y)	:	68	60	62	80	85	40	52	62	60	81

Calculate the two regression equations and estimate the productivity index of a worker whose test score is 92.

SOLUTION. **CALCULATIONS FOR REGRESSION EQUATIONS**

Aptitude Score X	Productivity Index Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
60	68	−5	3	25	9	−15
62	60	−3	−5	9	25	15
65	62	0	−3	0	9	0
70	80	5	15	25	225	75
72	85	7	20	49	400	140
48	40	−17	−25	289	625	425
53	52	−12	−13	144	169	156
73	62	8	−3	64	9	−24
65	60	0	−5	0	25	0
82	81	17	16	289	256	272
$\Sigma X = 650$	$\Sigma Y = 650$			$\Sigma x^2 = 894$	$\Sigma y^2 = 1752$	$\Sigma xy = 1044$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{650}{10} = 65 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{650}{10} = 65$$

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{1044}{894} = 1.168 \quad \text{and} \quad b_{XY} = \frac{\Sigma xy}{\Sigma y^2} = \frac{1044}{1752} = 0.596$$

Regression Equations

Regression Equation of Y on X is :

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

or, $Y - 65 = 1.168(X - 65)$

i.e., $Y - 65 = 1.168X - 75.92$

$\Rightarrow$ $Y = 1.168X - 10.92$ ... (1)

Regression Equation of X on Y is :

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

or, $X - 65 = 0.596(Y - 65)$

i.e., $X - 65 = 0.596Y - 38.74$

$\Rightarrow$ $X = 0.596Y + 26.26$ ... (2)

To estimate the productivity index of a worker whose test score is 92, we use the regression equation of Y on X. Thus substituting $X = 92$ in (1), we get $Y = 1.168 \times 92 - 10.92 = 96.536$

7.5 PROPERTIES OF REGRESSION COEFFICIENTS

In this section we shall derive some important properties of regression coefficients.

Property 1. The coefficient of correlation and the two regression coefficients have the same signs.

Proof. We know that the two regression coefficients are given by

$$b_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} \quad \text{and} \quad b_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_Y^2}$$

Since σ_X^2 and σ_Y^2 are always positive, whereas Cov (X, Y) may be positive or negative, therefore we conclude from the above formulas that the regression coefficients b_{YX} and b_{XY} have the same signs as that of Cov (X, Y). Further, we know that the coefficient of correlation r has the same sign as that of Cov (X, Y). Thus the coefficient of correlation and the two regression coefficients have the same signs.

Property 2. The coefficient of correlation is the geometric mean between the regression coefficients.

Proof. We have

$$b_{YX} \cdot b_{XY} = r\frac{\sigma_Y}{\sigma_X} \cdot r\frac{\sigma_X}{\sigma_Y} = r^2 \quad \text{i.e.,} \quad r^2 = b_{XY} \cdot b_{YX}$$

which shows that the coefficient of correlation is the geometric mean between the regression coefficients.

Property 3. If one of the regression coefficients is greater than unity, the other must be less than unity.

Proof. We know that the correlation coefficient r ranges from -1 to $+1$. And, therefore, $r^2 \leq 1$. Hence

$$b_{YX} \cdot b_{XY} = r^2 \leq 1 \qquad (\because b_{YX} \cdot b_{XY} = r^2)$$

and from the last inequality, the result follows immediately.

We now state (without proof) yet another property of regression coefficients.

Property 4. The regression coefficients are independent of change of origin but not of scale.

Infact, if we define $U = \frac{X-A}{h}$ and $V = \frac{Y-B}{k}$

where A, B, $h\,(>0)$ and $k\,(>0)$ are constants, then

$$b_{XY} = \frac{h}{k}\, b_{UV} \quad \text{and} \quad b_{YX} = \frac{k}{h}\, b_{VU}.$$

In particular, if we take $h = k = 1$, *i.e.*, we transform the variables X and Y to U and V by the relation :

$$U = X - A \quad \text{and} \quad V = Y - B,$$

then $b_{XY} = b_{UV}$ and $b_{YX} = b_{VU}$

EXAMPLE 14. *For some bivariate data, the following results were obtained :*

Mean value of variable $X = 53.2$ and of $Y = 39.5$. Regression coefficient of Y on $X = -1.5$ and of X on $Y = -0.38$. What should be the most likely value of X when $Y = 50$? Also find the coefficient of correlation between two variables. **[Delhi Univ. B.Com (H) 2005]**

SOLUTION. In terms of usual notations, we are given :

$$\bar{X} = 53.2, \quad \bar{Y} = 39.5, \quad b_{YX} = -1.5 \quad \text{and} \quad b_{XY} = -0.38$$

To find the most likely value of X when $Y = 50$, we use regression line of X on Y given by

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad \Rightarrow \quad X - 53.2 = -0.38\,(Y - 39.5)$$

$$\Rightarrow \quad X - 53.2 = -0.38Y + 15.01 \quad \Rightarrow \quad X = -0.38Y + 68.21 \qquad \ldots(1)$$

Thus, substituting $Y = 50$ in Eq. (1), we get

$$X = -0.38 \times 50 + 68.21 = -19 + 68.21 = 49.21$$

To find the coefficient of correlation, we use

$$r^2 = b_{YX} \cdot b_{XY} = (-1.5)(-0.38) = 0.57$$

$$\Rightarrow \quad r = \pm\sqrt{0.57} = \pm 0.7549$$

As both the regression coefficients are negative, $r = -0.7549$.

EXAMPLE 15. *Given below is the information relating to a bivariate distribution :*

Regression equation of Y on X : $Y = 20 + 0.4X$.

Mean of $X = 30$, Correlation coefficient between X and $Y = 0.8$.

Find regression equation of X on Y. **[Delhi Univ. B.Com (H) 2005, 2010]**

SOLUTION. Regression equation of Y on X is : $Y = 20 + 0.4X$. Since the line of regression of Y on X passes through the point $(\bar{X}, \bar{Y})$,

$$\bar{Y} = 20 + 0.4\bar{X}$$

Substituting $\bar{X} = 30$, we get $\bar{Y} = 20 + 0.4 \times 30 = 32$

Also, b_{YX} = regression coefficient of Y on $X = 0.4$

However, $b_{YX} = r\frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{\sigma_X}{\sigma_Y} = \frac{r}{b_{YX}} = \frac{0.8}{0.4} = 2$ $(\because r = 0.8)$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = 0.8 \times 2 = 1.6$$

Thus, regression equation of X on Y is :

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad \text{i.e.,} \quad X - 30 = 1.6(Y - 32)$$

$$\Rightarrow \quad X = 1.6Y - 21.2$$

EXAMPLE 16. *Compute the two regression coefficients from the data given below and find the value of 'r' (the correlation coefficient) using the same:*

X :	7	4	8	6	5
Y :	6	5	9	8	2

[*C.A. Foundation, May 1996, Nov. 1998*]

SOLUTION. **COMPUTATION OF REGRESSION COEFFICIENTS**

Y	X	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
7	6	1	0	1	0	0
4	5	-2	-1	4	1	2
8	9	2	3	4	9	6
6	8	0	2	0	4	0
5	2	-1	-4	1	16	4
$\sum X$ = 30	$\sum Y$ = 30			$\sum x^2$ = 10	$\sum y^2$ = 30	$\sum xy$ = 12

$$\bar{X} = \frac{\sum X}{n} = \frac{30}{5} = 6 \quad \text{and} \quad \bar{Y} = \frac{\sum Y}{n} = \frac{30}{5} = 6$$

Regression coefficient of X on Y :

$$b_{XY} = \frac{\sum xy}{\sum y^2} = \frac{12}{30} = 0.4$$

Regression coefficient of Y on X :

$$b_{YX} = \frac{\sum xy}{\sum x^2} = \frac{12}{10} = 1.2$$

To find the coefficient of correlation, we use $r^2 = b_{XY} \cdot b_{YX}$

CHAPTER 6

Correlation Analysis

OBJECTIVES

After studying the material in this chapter, you should be able to :

- *Understand the concept of correlation between two variables.*
- *Identify different types of correlation.*
- *Understand the notion of correlation coefficient.*
- *Interpret the value of coefficient of correlation.*
- *Discuss various methods of computing the correlation coefficient.*
- *Compute correlation coefficient for bivariate frequency distribution.*
- *Appreciate properties of correlation coefficient.*
- *Know the merits and demerits of different methods of studying correlation.*
- *Calculate probable error and interpret its value.*
- *Understand the concept of coefficient of determination.*

6.1 INTRODUCTION

Thus far we have examined numerical methods used to describe various characteristics of a univariate data, *i.e.*, the data involving only one variable. The reader may recall that in univariate data only one variable is associated with each unit of observation. However, we may have data in which more than one variable can be associated with each unit of observation. For example, for providing information about the marks obtained by the students of a class in two subjects, say Statistics and Economics, we can associate two variables, one representing the marks in Statistics and the other marks in Economics to each unit of observation, namely, a student in the class. When we have two variables for which values are being observed for each unit of observation, we say that we have *bivariate* data. In general, the study of those data which involve more than two variables are termed *multivariate data.*

Two variables are said to be *correlated* if the change in one variable is accompanied by a

small values of Y and small values of X to correspond to large values of Y. Hence we can say that price and demand of a product are correlated. *Correlation analysis* is a statistical procedure by which we determine the degree of association or relationship between two or more variables. That is, in correlation analysis, the purpose is to measure the strength or closeness of the relationship between the variables. For example, we might find a high degree of relationship between the price of a product and consumer demand for that product. Correlation is said to be *linear* if the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable.

In this chapter we shall consider the problem of measuring the linear relationship involving two variables only. The study of such a problem is called the *simple linear correlation.*

6.2 CORRELATION : SOME DEFINITIONS

In the following we shall give some important definitions of correlation.

1. *If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the other (s), then they are said to be correlated.* — **L.R. Connor**
2. *Correlation analysis attempts to determine the degree of relationship between variables.* — **Ya-Lun Chou**
3. *Correlation is an analysis of the covariation between two or more variables.* — **A.M. Tuttle**
4. *Correlation analysis deals with the association between two or more variables.* — **Simpson and Kafka**
5. *When the relationship is of quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.* — **Croxton and Cowden**
6. *Correlation means that between two series or group of data there exists some casual connection.* — **W.I. King**
7. *When a group of items are recorded with respect to the values of two distinct variables and it is found that pairs of values tend to be associated, the two variables are said to be correlated.* — **Wessel and Willet**

6.3 TYPES OF CORRELATION

The following are different types of correlation :

(i) *Positive and Negative Correlation.*

(ii) *Simple, Partial and Multiple Correlation.*

(iii) *Linear and Non-linear Correlation.*

(i) Positive and Negative Correlation

The correlation between two variables is said to be **positive or direct** if an increase (or a decrease) in one variable corresponds to an increase (or a decrease) in the other

...example, if X represents the amount of money spent annually on advertising by a ... and Y represents the total annual sales, then we might expect an increase (or a ...crease) in the advertising budget to be accompanied by an increase (or decrease) in ... annual sales. Thus we can say that the correlation between the advertising budget ... the total sales is positive.

...correlation between two variables is said to be **negative or inverse** if an increase (or ...decrease) in one variable corresponds to a decrease (or an increase) in the other. ...example, if X represents the price of a product and Y represents the demand for that ...duct, then we would expect large values of X to correspond to small values of Y and ...all values of X to correspond to large values of Y. Hence we can say that the correlation ...ween the price and demand of a product is negative.

Simple, Partial and Multiple Correlation

...he study of simple, partial and multiple correlation is based upon the number of variables ...olved.

...imple Correlation : It involves the study of only two variables. That is, in simple ...relation, we measure the degree of association or relationship between two variables ...nly. For example, when we study the correlation between the price and demand of a ...duct, it is a problem of simple correlation.

...artial Correlation : It involves the study of three or more variables, but consider only ...wo variables to be influencing each other, the effect of other influencing variables being ...pt constant. Thus in partial correlation we measure the degree of relationship between ...e variable Y and one of the variables $X_1, X_2, \ldots, X_n$ with the effect of all the other ...riables removed. For example, if we consider three variables, namely yield of wheat, amount of rainfall and amount of fertilizers and limit our correlation analysis to yield ...d rainfall, with the effect of fertilizers removed, it becomes a problem relating to partial correlation only.

...ultiple Correlation : It involves the study of three or more variables simultaneously. Thus in multiple correlation we measure the degree of relationship between the variable ... and all the variables $X_1, X_2, \ldots, X_n$ taken together. For example, if we study the relationship between the yield of wheat per acre and both amount of rainfall and the ...ount of fertilizers used, it becomes a problem relating to multiple correlation.

...i) Linear and Non-linear Correlation

Linear Correlation : The correlation between two variables is said to be *linear* if the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable. For example, consider the following data:

X	:	10	20	30	40	50
Y	:	40	80	120	160	200

We observe that the ratio of changes between the two variables is same and hence the correlation between X and Y is linear. It may be remarked that if the values of two ... which are linearly correlated are plotted on a graph paper all the plotted points

Non-linear (or Curvilinear) : The correlation between two variables is said to be *non-linear or curvilinear* if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. For example, if the amount of fertilizer used is doubled, the production of wheat may not necessarily be doubled. Thus we can say that the correlation between the amount of fertilizer and the production of wheat is non-linear.

6.4 CORRELATION AND CAUSATION

Although correlation analysis helps us in determining the degree of association or relationship between two or more variables, yet it does not tell us anything about cause and effect relationship that exist among the variables. Even a high degree of correlation between two variables does not necessarily imply a cause and effect relationship between them. The correlation between two variables may be due to one or more of the following reasons :

1. ***Mutual Dependence :*** Sometimes there is a high degree of correlation between two variables but it may be difficult to pinpoint the cause and effect variable. Such situations are usually observed in data relating to economic and business problems. For instance, economic theory tells us that price of a product may exert an influence on the demand of that product. Hence we can say that price is the cause and demand the effect. However, it is also possible that demand of the product due to growth of population or due to fashion or other factors such as changes in the tastes and habits of people may exert an influence on the price of that product. Hence we can say that demand is the cause and price the effect. Thus at times it becomes difficult to find out from two correlated variables which is the cause and which is the effect.

2. ***Influence of Third Variable :*** The correlation between two variables may be due to effect or influence of a third variable or a number of other variables. For example, we may find a high degree of correlation between the yield per acre of wheat and tea. But in reality this may be due to the effect of number of other factors such as fertilisers used, rainfall, irrigation facilities, etc. They have acted upon both the variables, causing them to respond together. However, neither of the two is the cause of the other.

3. ***Pure Chance :*** The correlation between two variables may due to chance particularly when the data pertain to a very small sample. For example, a small sample chosen from a large universe may show the relationship but such a relationship may not exist in the universe.

4. ***Non-sense or Spurious Correlation*** : There might be a situation of non-sense or spurious correlation between two variables under study. For example, we may observe a high degree of correlation between the incomes and weights of a group of persons even though there does not seem to be a common link between them.

6.5 THE CORRELATION COEFFICIENT

126

Correlation analysis attempts to measure the strength or closeness of linear relationships between two variables by means of a single number called a *correlation coefficient*.

nition : The quantitative measure of strength in the linear relationship between two riables is called the **correlation coefficient.** It is denoted by r.

the correlation coefficient r measures the extent to which the points cluster about traight line. The correlation coefficient ranges from +1 to − 1. If two variables have no r relationship, the correlation between them is zero. Consequently, the more relation differs from zero, the stronger the linear relationship between the two ables.

following table shows degrees of correlation according to various values of r:

TABLE 6.1

Degree of Correlation	Positive	Negative
Perfect correlation	+ 1	− 1
Very high degree of correlation	+ 0.9 to + 1	− 0.9 to − 1
Fairly high degree of correlation	+ 0.75 to + 0.9	− 0.75 to − 0.9
Moderate degree of correlation	+ 0.50 to + 0.75	− 0.50 to − 0.75
Low degree of correlation	+ 0.25 to + 0.50	− 0.25 to − 0.50
Very low degree of correlation	0 to + 0.25	− 0.25 to 0
No correlation	0	0

6 METHODS OF STUDYING CORRELATION

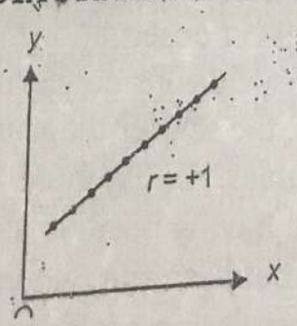
e shall discuss the following methods of measuring the linear relationship between wo variables:

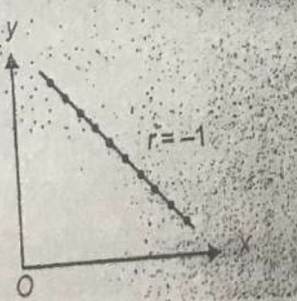
(i) Scatter Diagram Method,
(ii) Karl Pearson's Coefficient of Correlation,
(iii) Rank Correlation Method, and
(iv) Concurrent Deviation Method.

7 SCATTER DIAGRAM METHOD

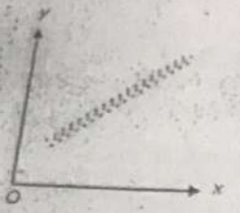
scatter diagram is a graphical presentation of bivariate data $\{(X_i, Y_i) : i = 1, 2, ..., n\}$ on two quantitative variables X and Y that allows us to show two variables together, one on ach axis, each pair being represented by a point on the graph as in coordinate geometry.

SCATTER DIAGRAMS SHOWING VARIOUS DEGREES OF CORRELATION

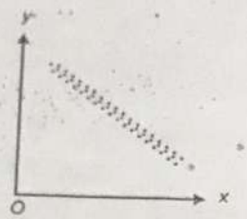




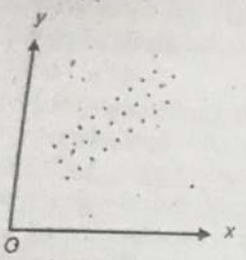
183



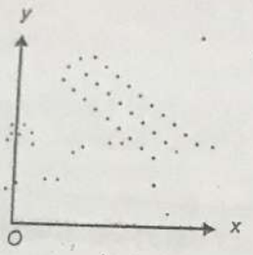
(c) High degree of positive correlation



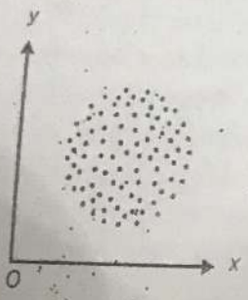
(d) High degree of negative correlation



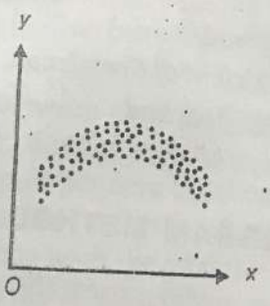
(e) Low degree of positive correlation



(f) Low degree of negative correlation



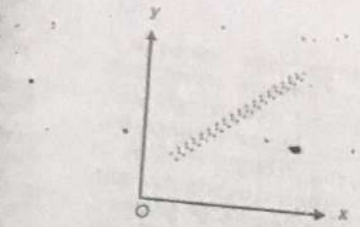
(g) No correlation



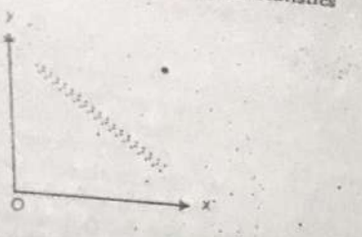
(h) No correlation

Fig. 6.1

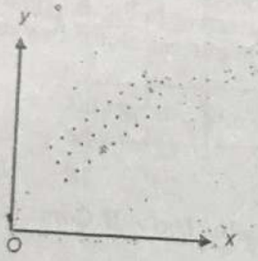
The scatter diagram is the simplest method of measuring the linear relationship between two variables. By constructing a scatter diagram for the n pairs of observations (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ on two variables, we can draw certain conclusions concerning the extent to which the points cluster about a straight line. Thus, the scatter diagram helps us to see if there is a useful relationship between the two variables. For example, if all the plotted points representing a given data lie on a straight line having positive slope [*see* Fig. 6.1 (*a*)], we say that there is a **perfect positive correlation** between the two variables. If two variables have a perfect positive correlation, then the correlation coefficient would be equal to + 1. On the other hand, if all the points lie on a straight line having negative slope [*see* Fig. (6.1(*b*)], we say that there is a **perfect negative correlation** between the two variables. If two variables have a perfect negative correlation, then the correlation



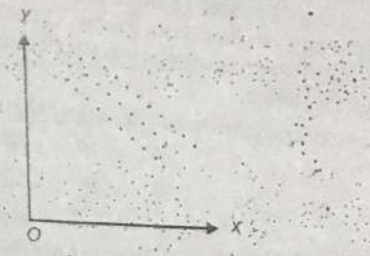
(c) High degree of positive correlation



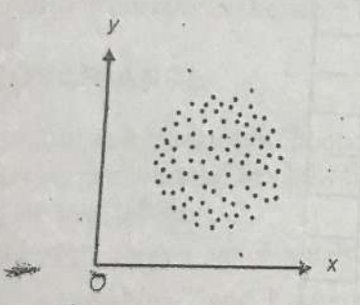
(d) High degree of negative correlation



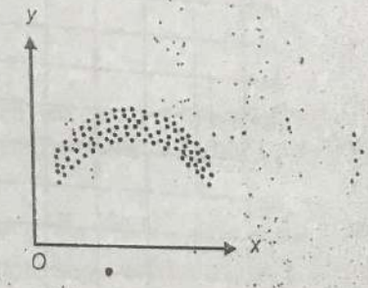
(e) Low degree of positive correlation



(f) Low degree of negative correlation



(g) No correlation



(h) No correlation

Fig. 6.1

The scatter diagram is the simplest method of measuring the linear relationship between two variables. By constructing a scatter diagram for the n pairs of observations (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ on two variables, we can draw certain conclusions concerning the extent to which the points cluster about a straight line. Thus, the scatter diagram helps us to see if there is a useful relationship between the two variables. For example, if all the plotted points representing a given data lie on a straight line having positive slope [*see* Fig. 6.1 (*a*)], we say that there is a **perfect positive correlation** between the two variables. If two variables have a perfect positive correlation, then the correlation coefficient would be equal to + 1. On the other hand, if all the points lie on a straight line having negative slope [*see* Fig. (6.1(*b*)], we say that there is a **perfect negative correlation** between the two

coefficient would be equal to −1. If the plotted points are all close to a straight line having positive slope [*see* Fig. 6.1(*c*)], we say that there is a **high degree of positive correlation** between the two variables. Similarly, if the plotted points are all close to a straight line having negative slope [*see* Fig. 6.1 (*d*)], we say that there is a **high degree of negative correlation.** In fact, the closer the correlation coefficient is to ±1, the stronger the linear relationship between the variables. If the plotted points are widely scattered over a straight line having positive slope [*see* Fig. 6.1 (*e*)], we say that there a low degree of positive correlation between the two variables. Similarly, if the plotted points are widely scattered over a straight line have negative slope [*see* Fig. 6.1 (*f*)], we say that there is a low degree of negative correlation. If the points follow a strictly random pattern as in Fig. 6.1 (*g*) and Fig. 6.1 (*h*), we have a zero correlation and conclude that no linear relationship exists between the two variables.

EXAMPLE 1. *Draw a scatter diagram to represent the following data and interpret it.*

X	4	5	6	7	8	9	10	11	12	13	14	15
Y	78	72	66	60	54	48	42	36	30	24	18	12

[*Delhi Univ. B.Com. 1978*]

SOLUTION. The scatter diagram is shown in Fig. 6.2.

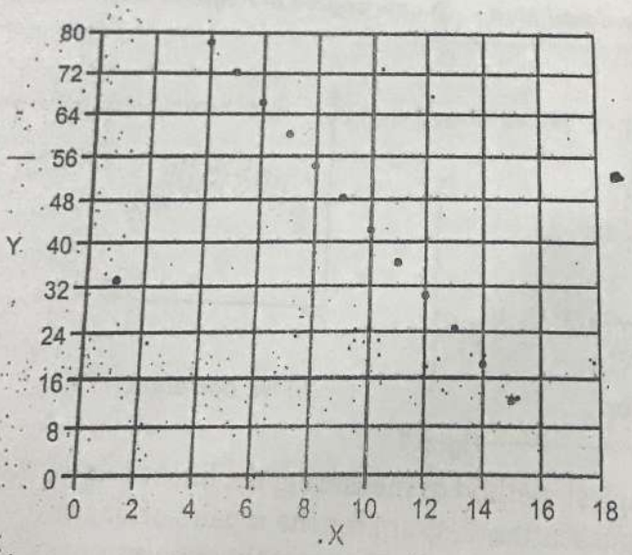


Fig. 6.2. Scatter Diagram

From the above scatter diagram, we find that the plotted points representing the given data lie on a straight line having negative slope. Hence we can conclude that there is a *perfect negative correlation* between the two variables.

EXAMPLE 2. *Draw a scatter diagram for the following data:*

X	8	10	12	11	9	7	13	14	15	17	16
Y	5	7	9	8	6	4	10	11	12	14	13

Also describe the relationship between X and Y. [*Delhi Univ. B.Com. 1979*]

SOLUTION. The scatter diagram is shown in Fig. 6.3.

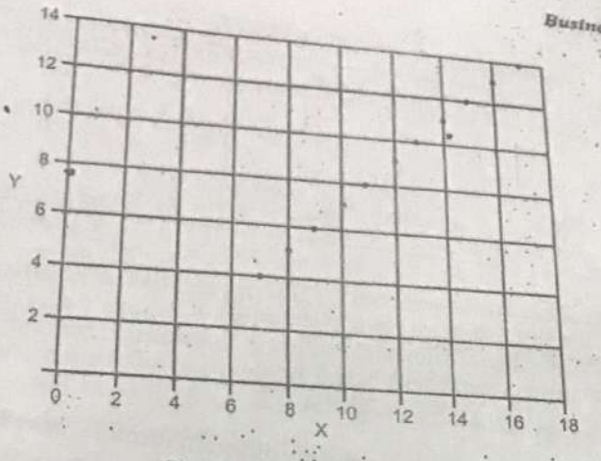


Fig. 5.3. Scatter Diagram

om the above scatter diagram, we find that the plotted points representing the given data lie on a straight line having positive slope. Hence we can conclude that there is a *rfect positive correlation* between the two variables.

8 COVARIANCE

this section we introduce the concept of covariance between two quantitative variables. In the next section we shall see how this concept is used to measure the linear relationship tween two variables.

Definition : Consider a set of n pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$ on two quantitative variables X and Y, where $X_1, X_2, \ldots$ denote observed values of the variable X, nd $Y_1, Y_2, \ldots$ those of Y. The **covariance** between X and Y, denoted by Cov (X, Y), is defined as

$$\text{Cov}(X, Y) = \frac{(X_1 - \bar{X})(Y_1 - \bar{Y}) + (X_2 - \bar{X})(Y_2 - \bar{Y}) + \ldots + (X_n - \bar{X})(Y_n - \bar{Y})}{n}$$

$$= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} = \frac{\sum xy}{n}$$

where $\bar{X} = \frac{\sum X}{n}$, $\bar{Y} = \frac{\sum X}{n}$, $x = X - \bar{X}$ and $y = Y - \bar{Y}$

EXAMPLE 3. *Find Cov (X, Y) between X and Y if*

X	:	3	4	5	6	7
Y	:	8	7	6	5	4

SOLUTION. **CALCULATION OF COVARIANCE**

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	$xy = (X - \bar{X})(Y - \bar{Y})$
3	8	−2	2	−4
4	7	−1	1	−1
5	6	0	0	0
6	5	1	−1	−1
7	4	2	−2	−4
$\Sigma X = 25$	$\Sigma Y = 30$			$\Sigma xy = -10$

Here $n = 5$, $\bar{X} = \frac{\Sigma X}{n} = \frac{25}{5} = 5$ and $\bar{Y} = \frac{\Sigma Y}{n} = \frac{30}{5} = 6$

$$\text{Cov}(X, Y) = \frac{\Sigma xy}{n} = \frac{-10}{5} = -2.$$

Another Formula for Cov (X, Y): We now give a slightly different formula (proof omitted) for calculating the covariance. This formula is particularly useful when $\bar{X}$ or $\bar{Y}$ is not an integer. The formula is :

$$\text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \left(\frac{\Sigma X}{n}\right)\left(\frac{\Sigma Y}{n}\right) \quad \ldots (1)$$

EXAMPLE 4. *Calculate the covariance between X and Y for the following data :*

X	1	2	3	4	5	6	7	8	9	10
Y	6	9	6	7	8	5	12	3	17	1

SOLUTION. **CALCULATION OF COVARIANCE**

X	Y	XY
1	6	6
2	9	18
3	6	18
4	7	28
5	8	40
6	5	30
7	12	84
8	3	24
9	17	153
10	1	10
$\Sigma X = 55$	$\Sigma Y = 74$	$\Sigma XY = 411$

We have $n = 10$, $\Sigma X = 55$, $\Sigma Y = 74$ and $\Sigma XY = 411$

$$\text{Cov}(X, Y) = \frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right) = \frac{411}{10} - \frac{55}{10} \times \frac{74}{10}$$
$$= 41.1 - 5.5 \times 7.4 = 41.1 - 40.7 = 0.4.$$

EXAMPLE 5. *Find the covariance between X and Y, given that* $\sum X = 60$, $\sum Y = 90$, $\sum XY =$ *74, and* $n = 10$.

SOLUTION. We have $$\text{Cov}(X, Y) = \frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right) = \frac{574}{10} - \left(\frac{60}{10}\right)\left(\frac{90}{10}\right)$$
$$= 57.4 - 6 \times 9 = 57.4 - 54 = 3.4.$$

REMARK. It may be remarked that formula (1) for computing covariance is effective if the values of X or/and Y are small. However, if the values of X or/and Y are large, the calculation of covariance by means or Formula (1) is quite tedious and time consuming. In such a case, we use the following method, called *step deviation method.*

Let $u = X - A$ and $v = Y - B$

where A and B are arbitrary constants. Then,

$$\text{Cov}(X, Y) = \frac{\sum uv}{n} - \left(\frac{\sum u}{n}\right)\left(\frac{\sum v}{n}\right) \quad \ldots (2)$$

This formula, in fact, shows that covariance is independent of change of origin.

EXAMPLE 6. *Find the covariance between X and Y for the following data :*

X :	66	67	68	69	70	71	72
Y :	68	65	70	70	69	70	69

SOLUTION. We shall use the following formula :

$$\text{Cov}(X, Y) = \frac{\sum uv}{n} - \left(\frac{\sum u}{n}\right)\left(\frac{\sum v}{n}\right)$$

where $u = X - A$ and $v = Y - B$. Taking $A = 69$ and $B = 70$, we prepare the following table :

TABLE 6.2. CALCULATION OF COVARIANCE

X	$u = X - 69$	Y	$v = Y - 70$	uv
66	−3	68	−2	6
67	−2	65	−5	10
68	−1	70	0	0
69	0	70	0	0
70	1	69	−1	−1
71	2	70	0	0
72	3	69	−1	−3
	$\sum u = 0$		$\sum v = -9$	$\sum uv = 12$

(ii) Coefficient of correlation between the two variables must be in the same units as the original data. [Delhi. Univ. B.Com. 1984]

[Delhi. Univ. B.Com. 1983]

ANSWERS

8. Positive	9. Positive	10. Positive	11. 4
12. 3.57	13. – 2.45	14. 10.2	15. 0.4
16. 3	17. 3.5	18. 4.93	
19. (*i*) False	(*ii*) False		

6.9 KARL PEARSON'S COEFFICIENT OF CORRELATION

The **Karl Pearson's coefficient of correlation**, also called the *Pearson's product - moment correlation coefficient*, is the most widely used method of measuring the linear correlation between two variables.

Definition. The ***Karl Pearson's coefficient of correlation*** between two variables X and Y, denoted by $\rho(X, Y)$ or r, is defined by

$$\rho(X, Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var } X}.\sqrt{\text{Var } Y}} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y},$$

where Var X and Var Y are the variances of the values of X and Y respectively, while σ_X and σ_Y are their standard deviations.

REMARK : It may be remarked that the correlation coefficient r is a measure of the linear relationship between the two variables X and Y. That is, r measures the extent to which the points of the scatter diagram cluster about a straight line. Hence a value of $r = 0$ implies a lack of linearity between the two variables and not a lack of association.

Characteristics of the Correlation Coefficient

the variables X and Y are replaced by the variables $U = aX + b$ and $V = cY + d$, then we have

$$\rho(U, V) = \rho(aX + b, cY + d) = \rho(X, Y),$$

where a and c are positive constants. However, if a and c are arbitrary constants, then

$$\rho(aX + b, cY + d) = \frac{a \times c}{|a| \times |c|} \rho(X, Y)$$

2. The value of r ranges from -1 to $+1$. That is, $-1 \leq r \leq +1$.
3. If $r = 1$, then all the points of the scatter diagram lie on a straight line having positive slope and we say that a perfect positive linear relationship exists between the two variables. Similarly, if $r = -1$, then all the points of the scatter diagram lie on a straight line having negative slope and we say that a perfect negative linear relationship exists between the two variables.
4. If r is close to $+1$, then all the points of the scatter diagram follow closely a straight line having positive slope and we say that a high positive correlation exists between the two variables. Similarly, if r is close to -1, then all the points of the scatter diagram follow closely a straight line having negative slope and we say that a high negative correlation exists between the two variables.
5. If r is close to 0, the linear relationship between the two variables is weak or perhaps non-existent.

EXAMPLE 7. *(a) "If the correlation coefficient between two variables X and Y is positive, then the coefficient of correlation between $-X$ and $-Y$ is also positive". Comment.*

[Delhi Univ. B.A. (Econ. Hons) 1996]

(b) The correlation coefficient between two variables X and Y is found to be 0.4. What is the correlation coefficient between 2X and $(-Y)$?

[Delhi Univ. B.A. (Econ. Hons) 1997]

(c) "If the coefficient of correlation between two variables X and Y is 0.8, then the coefficient of correlation between $-X$ and $-Y$ is -0.8" Comment.

[Delhi Univ. B.Com. (H) 1997, 2010]

SOLUTION. We know that

$$\rho(aX, cY) = \frac{a \times c}{|a| \times |c|} \rho(X, Y) \quad \ldots(1)$$

(a) We are given : $\rho(X, Y) > 0$. Using (1), we get

$$\rho(-X, -Y) = \frac{(-1) \times (-1)}{|-1| \times |-1|} \rho(X, Y) = \rho(X, Y)$$

Thus, if $\rho(X, Y)$ is positive, then so is $\rho(-X, -Y)$.

(b) We are given : $\rho(X, Y) = 0.4$. Using (1), we get

$$\rho(2X, -Y) = \frac{2\times(-1)}{|2|\times|-1|}\rho(X, Y) = \frac{-2}{2\times1}\rho(X, Y)$$

$$= -\rho(X, Y) = -0.4.$$

Thus, the correlation coefficient between $2X$ and $-Y$ is -0.4.

(c) We are given : $\rho(X, Y) = 0.8$. Using (1), we get

$$\rho(-X, -Y) = \frac{(-1)\times(-1)}{|-1|\times|-1|}\rho(X, Y) = \rho(X, Y) = 0.8$$

Thus, the coefficient of correlation between $-X$ and $-Y$ is also 0.8.

EXAMPLE 8. *The covariance between the length and weight of five items is 6 and their standard deviations are 2.45 and 2.61 respectively. Find the coefficient of correlation between length and weight.* **[Delhi Univ. B.Com. (H) 2000]**

SOLUTION. If we let X denote the length and Y denote the weight, then we are given :

$$\text{Cov}(X, Y) = 6, \quad \sigma_X = 2.45 \quad \text{and} \quad \sigma_Y = 2.61$$

The coefficient of correlation, r, between X and Y is given by

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{6}{2.45\times2.61} = \frac{6}{6.3945} = 0.9383.$$

EXAMPLE 9. *If covariance of 10 pairs of items is 7, variance of X is 36, $\sum Y - \bar{Y})^2 = 90$. find out*

[Delhi Univ. B.Com. 2004]

SOLUTION. We are given : $\text{Cov}(X, Y) = 7$

$$\text{Var}(X) = 36 \Rightarrow \sigma_X = \sqrt{\text{Var}\,X} = 6$$

$$\sum(Y - \bar{Y})^2 = 90 \quad \text{and} \quad n = 10$$

$$\sigma_Y = \sqrt{\frac{\sum(Y - \bar{Y})^2}{n}} = \sqrt{\frac{90}{10}} = 3$$

$$\therefore \quad r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{7}{6\times3} = \frac{7}{18} = 0.39 \text{ (app.)}$$

EXAMPLE 10. *The coefficient of correlation between two variables X and Y is 0.3 and their covariance is 9. If the variance of X series is 16, find the standard deviation of Y series.*

SOLUTION. The coefficient of correlation between X and Y is given by

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Given $r = 0.3$, $\text{Cov}(X, Y) = 9$, $\text{Var}(X) = 16 \Rightarrow \sigma_X = \sqrt{\text{Var}\,X} = 4$

9 9 193

SOLUTION. CALCULATION OF r

−4	3	16	9	−12
−3	−3	9	9	9
−2	−4	4	16	8
−1	0	1	0	0
0	4	0	16	0
1	1	1	1	1
2	2	4	4	4
3	−2	9	4	−6
4	−1	16	1	−4
		$\sum x^2 = 60$	$\sum y^2 = 60$	$\sum xy = 0$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{0}{\sqrt{60 \times 60}} = 0.$$

EXAMPLE 18. *From the following data, calculate Karl Pearson's coefficient of correlation:*

Height of Fathers (in inches) :	66	68	69	72	65	59	62	67	61	71
Height of Sons (in inches) :	65	64	67	69	64	60	59	68	60	64

[Delhi Univ. B.Com. 2005]

SOLUTION. CALCULATION OF COEFFICIENT OF CORRELATION

Height of Father X	$X - A$	x^2	Height of Son Y	$Y - Y$	y^2	xy
66	0	0	65	1	1	0
68	2	4	64	0	0	0
69	3	9	67	3	9	9
72	6	36	69	5	25	30
65	−1	1	64	0	0	0
59	−7	49	60	−4	16	28
62	−4	16	59	−5	25	20
67	1	1	68	4	16	4
61	−5	25	60	−4	16	20
71	5	25	64	0	0	0
$\sum X = 660$		$\sum x^2 = 166$	$\sum Y = 640$		$\sum y^2 = 108$	$\sum xy = 111$

$n = 10$, $\bar{X} = \frac{\sum X}{n} = \frac{660}{10} = 66$ and $\bar{Y} = \frac{\sum Y}{n} = \frac{640}{10} = 64$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{111}{\sqrt{166 \times 108}} = \frac{111}{133.89} = 0.829.$$

LE 19. *Making use of the data summarized below calculate the coefficient of correlation.*

Case	X_1	X_2	Case	X_1	X_2
A	10	9	E	12	11
B	6	4	F	13	13
C	9	6	G	11	8
D	10	9	H	9	4

[Delhi Univ. B.Com, 1982]

ITION.

CALCULATION OF COEFFICIENT OF CORRELATION

Case	X_1	$X_1-\bar{X}_1$ x_1	x_1^2	X_2	$X_2-\bar{X}_2$ x_2	x_2^2	x_1x_2
	10	0	0	9	1	1	0
	6	−4	16	4	−4	16	16
	9	−1	1	6	−2	4	2
	10	0	0	9	1	1	0
	12	2	4	11	3	9	6
	13	3	9	13	5	25	15
	11	1	1	8	0	0	0
	9	−1	1	4	−4	16	4
	$\sum X_1 = 80$	—	$\sum x_1^2 = 32$	$\sum X_2 = 64$		$\sum x_2^2 = 72$	$\sum x_1x_2 = 43$

e $n=8$, $\bar{X}_1 = \frac{\sum X_1}{n} = \frac{80}{8} = 10$ and $\bar{X}_2 = \frac{\sum X_2}{n} = \frac{64}{8} = 8$

$$r_{12} = \frac{\sum x_1x_2}{\sqrt{\sum x_1^2 \times \sum x_2^2}} = \frac{43}{\sqrt{32\times72}} = \frac{43}{48} = 0.896.$$

E 20. *Calculate the correlation coefficient from the following data :*

:	12	9	8	10	11	13	7
:	14	8	6	9	11	12	3

each value of X be multiplied by 2 and then 6 be added to it. Similarly, multiply lue of Y be 3 and subtract 2 from it. What will be the correlation coefficient between series of X and Y? **[C.A. (Foundation), May 1997]**

SOLUTION. CALCULATION OF COEFFICIENT OF CORRELATION

12	2	4	14	5	25	10
9	−1	1	8	−1	1	1
8	−2	4	6	−3	9	6
10	0	0	9	0	0	0
11	1	1	11	2	4	2
13	3	9	12	3	9	−9
7	−3	9	3	−6	36	18
$\Sigma X = 70$		$\Sigma x^2 = 28$	$\Sigma Y = 63$		$\Sigma y^2 = 84$	$\Sigma xy = 46$

Here $n = 7$, $\bar{X} = \frac{\Sigma X}{n} = \frac{70}{7} = 10$ and $\bar{Y} = \frac{\Sigma Y}{n} = \frac{63}{7} = 9$

$$\therefore \quad r(X, Y) = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{46}{\sqrt{28 \times 84}} = \frac{46}{48.5} = 0.948$$

We know that if the variables X and Y are replaced by the variables $U = aX + b$ and $V = cY + d$, where a, b, c and d are constants, $a > 0$, $c > 0$, then $r(U, V) = r(X, Y)$.

Using this fact, we find that the coefficient of correlation between the new series of X and Y is also 0.948.

EXAMPLE 21. *From the following data, calculate Karl Pearson's coefficient of correlation :*

X :	6	2	10	4	8
Y :	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

SOLUTION. First we find the missing value of Y series.

Given : $\bar{Y} = 8 \Rightarrow \frac{\Sigma Y}{5} = 8 \Rightarrow \Sigma Y = 8 \times 5 = 40$

$\therefore$ missing value $= \Sigma Y - (9 + 11 + 8 + 7) = 40 - 35 = 5$

CALCULATION OF COEFFICIENT OF CORRELATION

X	$X - \bar{X}$ x	x^2	Y	$Y - \bar{Y}$ y	y^2	xy
6	0	0	9	+1	1	0
2	−4	16	11	+3	9	−12
10	+4	16	5	−3	9	−12
4	−2	4	8	0	0	0
8	+2	4	7	−1	1	−2
		$\Sigma x^2 = 40$			$\Sigma y^2 = 20$	$\Sigma xy = -26$

25	5	25	10	−10	100	−50
40	20	400	25	5	25	100
30	10	100	40	20	400	200
25	5	25	15	−5	25	−25
10	−10	100	20	0	0	0
5	−15	225	40	20	400	−300
10	−10	100	28	8	64	−80
15	−5	25	22	2	4	−10
30	10	100	15	−5	25	−50
20	0	0	5	−15	225	0
	$\Sigma U = 10$	$\Sigma U^2 = 1100$		$\Sigma V = 20$	$\Sigma V^2 = 1268$	$\Sigma UV = -215$

$$r = \frac{n\Sigma UV - \Sigma U \cdot \Sigma V}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\sqrt{n\Sigma V^2 - (\Sigma V)^2}}$$

$$= \frac{10(-215) - 10 \times 20}{\sqrt{10 \times 1100 - 100}\cdot\sqrt{10 \times 1268 - 400}}.$$

$$= \frac{-2350}{\sqrt{10900}\sqrt{12280}} = \frac{-2350}{11569.44} = -0.203 \text{ (app.)},$$

which is same as before.

EXAMPLE 23. *The total of the multiplication of deviation of X and Y = 3044. No of pairs of observations = 10. Total of deviations of X = − 170. Total of deviations of Y = − 20. Total of the squares of deviations of X = 8288. Total of the squares of deviations of Y = 2264. Find out the coefficient of correlation when the arbitrary means of X and Y are 82 and 68 respectively.* **[Delhi Univ. B.Com. 2001]**

SOLUTION. In terms of usual notations, we are given :

$$n = 10; \quad \Sigma UV = 3044, \quad \Sigma U = -170, \quad \Sigma V = -20, \quad \Sigma U^2 = 8288, \quad \Sigma V^2 = 2264.$$

Applying the formula : $r = \dfrac{n\Sigma UV - (\Sigma U).(\Sigma V)}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\cdot\sqrt{n\Sigma V^2 - (\Sigma V)^2}}$, we get

$$r = \frac{10 \times 3044 - (-170)(-20)}{\sqrt{10 \times 8288 - (-170)^2}\cdot\sqrt{10 \times 2264 - (-20)^2}}$$

$$r = \frac{30440 - 3400}{\sqrt{82880 - 28900}\ \sqrt{22640 - 400}}$$

$$= \frac{27040}{\sqrt{53980}\ \sqrt{22240}} = +0.78$$

MPLE 24. *Calculate coefficient of correlation from the following data :*

X	10,000	20,000	30,000	40,000	50,000	60,000	70,000
Y	0.3	0.5	0.6	0.8	1.0	1.1	1.3

UTION. Since coefficient of correlation is independent of change of origin and scale, erefore if we let

$$U = X/10{,}000 \quad \text{and} \quad V = 10Y$$

en

$$r = \rho(X, Y) = \rho(U, V)$$

aking use of the above fact, we prepare the following table :

CALCULATION OF COEFFICIENT OF CORRELATION

X	$U = X/10{,}000$	U^2	Y	$V = 10Y$	V^2	UV
10,000	1	1	0.3	3	9	3
20,000	2	4	0.5	5	25	10
30,000	3	9	0.6	6	36	18
40,000	4	16	0.8	8	64	32
50,000	5	25	1.0	10	100	50
60,000	6	36	1.1	11	121	66
70,000	7	49	1.3	13	169	91
	$\sum U = 28$	$\sum U^2 = 140$		$\sum V = 56$	$\sum V^2 = 524$	$\sum UV = 270$

$$r = \frac{n\sum UV - (\sum U).(\sum V)}{\sqrt{n\sum U^2 - (\sum U)^2}\ .\ \sqrt{n\sum V^2 - (\sum V)^2}}$$

$$= \frac{7 \times 270 - 28 \times 56}{\sqrt{7 \times 140 - 784}\ .\ \sqrt{7 \times 524 - 3136}}$$

$$= \frac{1890 - 1568}{\sqrt{196} \times \sqrt{532}} = \frac{322}{322.91} = +0.997.$$

MPLE 25. *Find Karl Pearson's coefficient of correlation between the age and the playing bits of the people from the following information :*

$$= \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}}$$

$$= \frac{(100\times34) - (-25)\times(76)}{\sqrt{(100\times157) - (-25)^2}\cdot\sqrt{100\times26 - (76)^2}}$$

$$= \frac{3400 + 1900}{\sqrt{15700 - 625}\cdot\sqrt{12600 - 5776}}$$

$$= \frac{5300}{\sqrt{15075}\times\sqrt{6824}} = \frac{5300}{10142.57} = +0.523.$$

6.12 PROBABLE ERROR OF COEFFICIENT OF CORRELATION

After calculating the coefficient of correlation, the next step is to find the extent to which it is dependable. For this purpose, the probable error of the coefficient of correlation is calculated. With the help of probable error, it is possible to determine the reliablility of an observed value of correlation coefficient in so far as it depends on the conditions of random sampling. The probable error of the coefficient of correlation is given by

$$P.E.(r) = 0.6745\frac{1-r^2}{\sqrt{n}},$$

where r is the coefficient of correlation and n is the number of pairs of observations.

Uses of Probable Error

1. *If the numerical value of r is less than the probable error (i.e., $|r| < P.E.(r)$), there is no evidence of correlation, i.e., the value of r is not at all significant.*
2. *If the numerical value of r is more than six times the probable error (i.e., $|r| > 6\,P.E.(r)$), the value of r is significant.*
3. *If the value of probable error is added to and subtracted from the coefficient of correlation, we obtain respectively the upper and lower limits within which the population coefficient of correlation may be expected to lie. Symbolically, limits for population correlation coefficient are*

$$r \pm P.E.(r) \qquad \ldots(1)$$

This means that if we take another sample of size n from the same population, then its correlation coefficient can be expected to lie within the limits given by (1). For example, f a coefficient of correlation of + 0.9 is observed by a study of 16 pairs of observations, ne probable error would be :

$$P.E.(r) = 0.6745\frac{1-(0.9)^2}{\sqrt{16}} = 0.032$$

AMPLE 39. *Find the coefficient of correlation between price and sales from the following ta and interpret its value through probable error*

e (Rs.)	103	98	85	92	90	84	88	90	94	99
les (units)	500	610	700	630	670	800	800	570	700	680

LUTION. [*Delhi Univ. B.Com. (H) 2006*]

COMPUTATION OF COEFFICIENT OF CORRELATION

Price (Rs.) X	*Sales (units) Y*	$U = X - 90$	$V = (Y - 700)/10$	U^2	V^2	UV
103	500	13	−20	169	400	−260
98	610	8	−9	64	81	−72
85	700	−5	0	25	0	0
92	630	2	−7	4	49	−14
90	670	0	−3	0	9	0
84	800	−6	10	36	100	−60
88	800	−2	10	4	100	−20
90	570	0	−13	0	169	0
94	700	4	0	16	0	0
95	680	5	−2	25	4	−10
		$\Sigma U = 19$	$\Sigma V = -34$	$\Sigma U^2 = 343$	$\Sigma V^2 = 912$	$\Sigma UV = -436$

efficient of correlation between price and sales is given by

$$r = \frac{n\Sigma UV - (\Sigma U)(\Sigma V)}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\sqrt{n\Sigma V^2 - (\Sigma V)^2}}$$

$$= \frac{10\times(-436) - (19)\times(-34)}{\sqrt{(10\times343) - (19)^2}\sqrt{(10\times912) - (-34)^2}}$$

$$\therefore = \frac{-4360 + 646}{\sqrt{3069}\cdot\sqrt{7964}} = \frac{-3714}{4943.84} = -0.751.$$

he probable error (*P.E.*) of the coefficient of correlation is given by :

$$P.E. = 0.6745\,\frac{1 - r^2}{\sqrt{n}},$$

where n is the number of pairs of observations. Substituting $r = -0.751$ and $n = 10$, get

$$P.E. = 0.6745\left[\frac{1 - (-0.751)^2}{\sqrt{10}}\right] = 0.6745\left[\frac{1 - 0.564}{3.162}\right]$$

207

Thus the limits within which r lies for another sample from the same universe would be :

$$r \pm P.E.(r) = 0.7 \pm 0.0687 \text{ or } 0.6313 - 0.7687.$$

EXAMPLE 37. *Test the significance of correlation for the following values based on the number of observations (i) 10 and (ii) 100, $r = +0.4$ and $+0.9$.*

SOLUTION. The coefficient of correlation is definitely significant if it is more than 6 times the probable error, i.e.,

$$r > 6\,P.E.(r) \quad \text{or} \quad \frac{r}{P.E.(r)} > 6.$$

The significance of correlation for various values has been shown in the following table :

No. of observations	r	$P.E.(r)$	$\frac{r}{P.E.(r)}$	Conclusion
10	0.4	$0.6745\,\frac{1-(0.4)^2}{\sqrt{10}} = 0.18$	$\frac{0.4}{0.18} = 2.22 < 6$	*not significant*
100	0.4	$0.6745\,\frac{1-(0.4)^2}{\sqrt{100}} = 0.06$	$\frac{0.4}{0.06} = 6.6 > 6$	*significant*
10	0.9	$0.6745\,\frac{1-(0.9)^2}{\sqrt{10}} = 0.04$	$\frac{0.9}{0.4} = 22.5 > 6$	*highly significant*
100	0.9	$0.6745\,\frac{1-(0.9)^2}{\sqrt{100}} = 0.0128$	$\frac{0.9}{0.0128} = 70.3 > 6$	*very highly significant*

EXAMPLE 38. *The coefficient of correlation and its probable error for n pairs of observations are found to be 0.917 and 0.034 respectively. Find the value of n.*

SOLUTION. We are given : $r = 0.917$ and $P.E.(r) = 0.034$.

$$P.E.(r) = 0.6745\,\frac{1-r^2}{\sqrt{n}}$$

$$\therefore \quad 0.034 = 0.6745\,\frac{1-(0.917)^2}{\sqrt{n}} = 0.6745\,\frac{1-0.841}{\sqrt{n}} = \frac{0.6745 \times 0.159}{\sqrt{n}}$$

$$\Rightarrow \quad 0.034\sqrt{n} = 0.1072 \qquad \Rightarrow \quad \sqrt{n} = \frac{0.1072}{0.034} = 3.153.$$

$$\Rightarrow \quad n = (3.153)^2 = 9.941 \simeq 10$$

Thus the value of n is 10.

...between age and playing habits.

Probable Error. Probable error, P.E. (r), is given by

$$P.E.(r) = 0.6745\,\frac{1-r^2}{\sqrt{n}}$$

$$= 0.6745\,\frac{1-(-0.9912)^2}{\sqrt{6}} = \frac{0.6745\times 0.0175}{2.45}$$

$$= \frac{0.0118}{2.45} = 0.0048$$

$$\Rightarrow \quad \frac{|r|}{P.E.(r)} = \frac{0.9912}{0.0048} = 206.5 > 6$$

Since the numerical value of r is more than 6 times the probable error, it is considered very highly significant.

3.13 COEFFICIENT OF DETERMINATION

Another very important and useful method of interpreting the value of coefficient of correlation is the **coefficient of determination** which is defined as the square of coefficient of correlation. Thus

$$\textit{coefficient of determination} = (\textit{coefficient of correlation})^2$$

The coefficient of determination enables us to find the effect of the independent variable on the dependable variable. For example, if the coefficient of correlation between price and supply is $r = +0.9$, it does not mean that 90 per cent of the variation in supply is due to change in price. However, if we calculate the coefficient of determination, r^2, which in this case would be 0.81, we can say that 81 per cent of the variation in supply is due to the change in price. In other words, it would mean that 81 per cent of the variation in supply has been explained by the variation in price. Thus we can say that

$$\textit{coefficient of determination} = \frac{\textit{Explained Variation}}{\textit{Total Variation}}$$

29

In the above example, the total variation is unity and the explained variation is 0.81. The ratio of unexplained variation to total variation is called the coefficient of non-determination, denoted by k^2. Thus

$$k^2 = \text{coefficient of non-determination} = \frac{\text{Unexplained variation}}{\text{Total variation}} = 1 - r^2$$

The square root of the coefficient of non-determination is called the coefficient of alienation, denoted by k. Thus,

$$k = \text{coefficient of non-alienation} = \sqrt{k^2} = \sqrt{1 - r^2}$$

In the above example, the unexplained variation is $1 - 0.81 = 0.19$, indicating that 19 per cent of the variation in supply is due to other factors.

The relationship between r and r^2 should be carefully noted. As the value of r decreases from its maximum value of 1, the value of r^2 decreases much more rapidly, as is clear from the following table:

r	r^2	r	r^2
1.0	1.0	0.5	0.25
0.9	0.81	0.4	0.16
0.8	0.64	0.3	0.09
0.7	0.49	0.2	0.04
0.6	0.36	0.1	0.01

The above table shows that r^2 is always less than r unless r happens to be unity in which case r and r^2 are equal. Further, it should be clearly understood that if the coefficient of correlation between two variables has a value of + 0.4 and the coefficient of correlation between the other two variables has a value of + 0.8, it does not mean that correlation between two variables in the second set is twice as strong as the correlation between two variables in the first set. In fact, the relationship between these two sets can be better understood by computing the value of r^2. When $r = 0.4$, $r^2 = 0.16$ and when $r = 0.8$, $r^2 = 0.64$. This means that the relationship in the second set is four times as strong as it is in the first set. In the first set only 16% of the total variation is explained while in the second set 64% of the total variation is explained.

EXAMPLE 41. *A correlation coefficient of 0.5 does not mean that 50% of the data are explained. Comment.* **[Delhi Univ. B.Com. (H) 1998]**

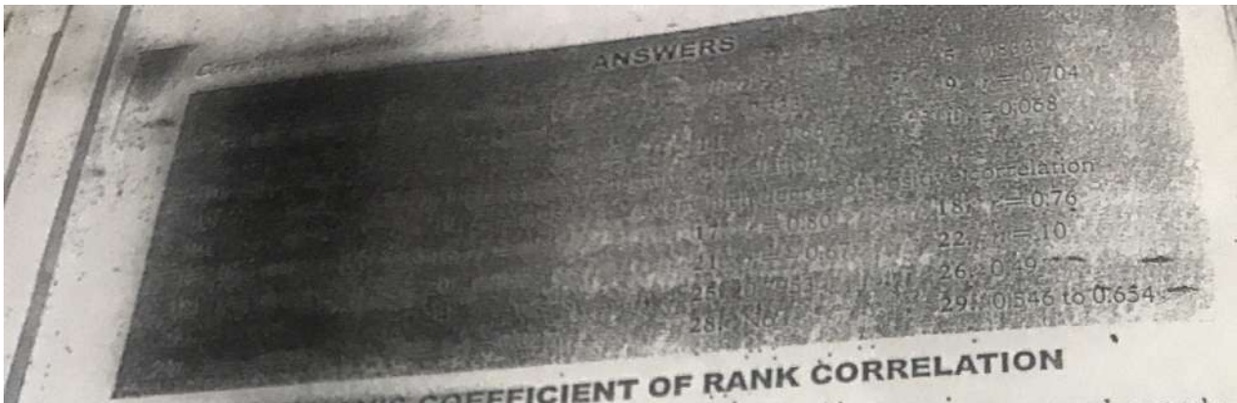
SOLUTION. The given statement is correct. In fact, to find the percentage of the data that are explained, we need to find the coefficient of determination, r^2, which in this case is 0.25. Thus we can say that 25% of the data are explained.

EXAMPLE 42. *Calculate correlation coefficient from the following data :*

$$N = 10, \quad \Sigma X = 140, \quad \Sigma Y = 150, \quad \Sigma (X - 10)^2 = 180,$$

$$\Sigma (Y - 15)^2 = 215, \quad \Sigma (X - 10)(Y - 15) = 60.$$

Also calculate coefficient of determination. **[Delhi Univ. B.Com. (H) 2007 (C.C.)]**



6.14 SPEARMAN'S COEFFICIENT OF RANK CORRELATION

Sometimes, we are given a series of items where no numerical measure can be made, but where best and worst or most favoured and least favoured can be identified. Rankings are often applied in these situations to put the series into an order. For example, the characteristics like beauty, intelligence, leadership ability, honesty, etc. cannot be measured numerically, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating its rank in the group.

If we have a group of individuals ranked according to two different qualities, it is natural to ask the following question:

"Is there an association between the rankings?"

To answer this question, we need to use a formula known as **Spearman's coefficient of rank correlation:**

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

where D is the difference between the two ranks given to each individual and n is the number of observations.

The Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks and is interpreted in much the same way. As before, the value of ρ will range from -1 to $+1$. A value of $+1$ indicates perfect association for identical rankings and a value of -1 indicates perfect association for reverse rankings. This will be clear from the following illustration:

Rank R_1	Rank R_2	D $R_1 - R_2$	D^2	Rank R_1	Rank R_2	D $R_1 - R_2$	D^2
1	1	0	0	1	4	−3	9
2	2	0	0	2	3	−1	1
3	3	0	0	3	2	1	1
4	4	0	0	4	1	3	9
			$\sum D^2 = 0$				$\sum D^2 = 20$

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)} \qquad \rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

$$= 1 - 0 = 1 \qquad = 1 - \frac{6 \times 20}{4(4^2-1)} = 1 - \frac{120}{60} = 1 - 2 = -1.$$

15 COMPUTING THE RANK CORRELATION COEFFICIENT

shall consider the following three cases to compute the rank correlation coefficient.

se I : When Actual Ranks Are Given

n this case the following steps are involved:

(i) Compute D, the difference between the two ranks given to each individual.

(ii) Compute D^2 and obtain the sum $\sum D^2$.

(iii) Apply the formula: $\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$, where n is the number of observations.

e following examples will illustrate.

KAMPLE 43. *Ten competitors in a beauty contest are ranked by two judges in the following der:*

I Judge	:	1	6	5	10	3	2	4	9	7	8
II Judge	:	6	4	9	8	1	2	3	10	5	7

lculate the Spearman's rank correlation coefficient. Is there an association between the ankings?

LUTION. COMPUTATION OF RANK CORRELATION COEFFICIENT

Rank by I Judge R_1	Rank by II Judge R_2	D $(R_1 - R_2)$	D^2
1	6	−5	25
6	4	2	4
5	9	−4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	−1	1
7	5	2	4
8	7	1	1
			$\sum D^2 = 60$

Rank correlation between the judgment of the two judges is given by

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 60}{10(10^2-1)} = 1 - \frac{36}{99} = 1 - \frac{4}{11} = \frac{7}{11} = 0.636.$$

This answer suggests that there is some degree of association between the rankings given by the two judges.

EXAMPLE 44. *Ten competitors in a beauty contest are ranked by three judges in the following order :*

I Judge	1	4	8	9	6	10	7	3	2	5
II Judge	4	8	7	5	9	6	10	2	3	1
III Judge	6	7	1	8	10	5	9	2	3	4

Use the rank correlation method to determine which pair of judges has the nearest approach to common taste in beauty.

SOLUTION. In order to find out which pair of judges has the nearest approach to common taste in beauty, we compare the Rank Correlation between the judgments of :

(i) *I* Judge and *II* Judge

(ii) *I* Judge and *III* Judge

(iii) *II* Judge and *III* Judge

COMPUTATION OF RANK CORRELATION

Rank by I Judge R_1	Rank by II Judge R_2	Rank by III Judge R_3	D_{12} (R_1-R_2)	D_{13} (R_1-R_3)	D_{23} (R_2-R_3)	D_{12}^2	D_{13}^2	D_{23}^2
1	4	6	−3	−5	−2	9	25	4
4	8	7	−4	−3	+1	16	9	1
8	7	1	+1	+7	+6	1	49	36
9	5	8	+4	+1	−3	16	1	9
6	9	10	−3	−4	−1	9	16	1
10	6	5	+4	+5	+1	16	25	1
7	10	9	−3	−2	+1	9	4	1
3	2	2	+1	+1	0	1	1	0
2	3	3	−1	−1	0	1	1	0
5	1	4	+4	+1	−3	16	1	9
						$\sum D_{12}^2 = 94$	$\sum D_{13}^2 = 132$	$\sum D_{23}^2 = 62$

The rank correlation coefficient is given by :

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

158 214

ank correlation between the judgments of first and second judges :

$$\rho_{12} = 1 - \frac{6 \times 94}{10(10^2 - 1)} = 1 - \frac{564}{990} = 0.43$$

k correlation between the judgments of first and third judges :

$$\rho_{13} = 1 - \frac{6 \times 132}{10(10^2 - 1)} = 1 - \frac{792}{990} = 0.2$$

nk correlation between the judgments of second and third judges :

$$\rho_{23} = 1 - \frac{6 \times 62}{10(10^2 - 1)} = 1 - \frac{372}{990} = 0.62$$

nce rank correlation coefficient is maximum in the judgment of the second and third ges, we conclude that they have the nearest approach to common tastes in beauty.

MPLE 45. *Rankings of 10 trainees at the beginning and at the end of a certain course are ven below:*

nees	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
nk at the beginning	1	6	3	9	5	2	7	10	8	4
nk at the end	6	8	3	7	2	1	5	9	4	10

culate spearman's rank correlation coefficient.

[*I.C.W.A. (Intermediate), June.1995*]

TION. **COMPUTATION OF RANK CORRELATION COEFFICIENT**

inees	*Rank at the beginning* R_1	*Rank at the end* R_2	D $R_1 - R_2$	D^2
A	1	6	−5	25
B	6	8	−2	4
C	3	3	0	0
D	9	7	2	4
E	5	2	3	9
F	2	1	1	1
G	7	5	2	4
H	10	9	1	1
I	8	4	4	16
J	4	10	−6	36
				$\sum D^2 = 100$

rank correlation coefficient is given by :

$$\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)},$$

215

$$\rho = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{n(n^2-1)}$$

where m_1 represents the number of times first item is repeated, m_2 represents the number of times second item is repeated and so on.

EXAMPLE 52. *Calculate Spearman's coefficient of rank correlation from the following data:*

X :	57	16	24	65	16	16	9	40	33	48
Y :	19	6	9	20	4	15	6	24	13	13

[*Delhi Univ. B.Com. 1997*]

SOLUTION. The following table shows the ranking from the highest value in both the series. Moreover, certain items in both the series are repeated, so ranking is done in accordance with suitable average.

CALCULATION OF RANK CORRELATION COEFFICIENT

X	Rank assigned R_1	Y	Rank assigned R_2	D $R_1 - R_2$	D^2
57	2	19	3	−1	1
16	8	6	8.5	−0.5	0.25
24	6	9	7	−1	1
65	1	20	2	−1	1
16	8	4	10	−2	4
16	8	15	4	4	16
9	10	6	8.5	1.5	2.25
40	4	24	1	3	9
33	5	13	5.5	−0.5	0.25
48	3	13	5.5	−2.5	6.25
$n=10$		$n=10$			$\sum D^2 = 41$

Note that in series X, the item 16 is repeated 3 times (*i.e.*, $m_1 = 3$). In series Y, the item 13 is repeated twice (*i.e.*, $m_2 = 2$) and 6 is also repeated twice (*i.e.*, $m_3 = 2$).

Thus, Spearman's coefficient of rank correlation is given by

$$\rho = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[41 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{10(10^2 - 1)}$$

EXAMPLE 5A. Find the Rank Correlation Coefficient from the following marks awarded by the examiners in Statistics :

Roll Nos.	1	2	3	4	5	6	7	8	9	10	11
Marks awarded by Examiner A :	24	29	19	14	30	19	27	30	20	28	11
Marks awarded by Examiner B :	37	35	16	26	23	27	19	20	16	11	21
Marks awarded by Examiner C :	30	28	20	25	25	30	20	24	22	29	15

[*Delhi Univ. B.Com. (H) 2005*]

SOLUTION. The following table shows the ranking from the highest value in each series. Moreover, certain values in both the series are repeated, so ranking is done in accordance with suitable average.

CALCULATION OF RANK CORRELATION COEFFICIENT

Roll No. of Examinee	Marks awarded by Examiner A	Rank R_1	Marks awarded by Examiner B	Rank R_2	Marks awarded by Examiner C	Rank R_3	D_{AB}^2 $(R_1-R_2)^2$	D_{AC}^2 $(R_1-R_3)^2$	D_{BC}^2 $(R_2-R_3)^2$
1	24	6	37	1	30	1.5	25	20.25	0.25
2	29	3	35	2	28	4	1	1	4
3	19	8.5	16	9.5	20	9.5	1	1	0
4	14	10	26	4	25	5.5	36	20.25	2.25
5	30	1.5	23	5	25	5.5	12.25	16	0.25
6	19	8.5	27	3	30	1.5	30.25	49	2.25
7	27	5	19	8	20	9.5	9	20.25	2.25
8	30	1.5	20	7	24	7	30.25	30.25	0
9	20	7	16	9.5	22	8	6.25	1	2.25
10	28	4	11	11	29	3	49	1	64
11	11	11	21	6	15	11	25	0	25
							$\sum D_{AB}^2$ = 225	$\sum D_{AC}^2$ = 160	$\sum D_{BC}^2$ = 102.5

$$R_{AB} = 1-\frac{6\left[\sum D_{AB}^2+\frac{\sum(m^3-m)}{12}\right]}{n(n^2-1)}$$

$$= 1-\frac{6\left[225+\frac{1}{12}(2^3-2)+\frac{1}{12}(2^3-2)\right]}{11(11^2-1)} = 1-\frac{6[225+0.5+0.5]}{1320}$$

$$= 1-\frac{6\times226}{1320} = 1-\frac{226}{220} = -\frac{6}{220} = -0.027$$

$$R_{AC} = 1 - \frac{6\left[\sum D_{AC}^2 + \frac{\sum(m^3 - m)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[160 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right]}{11(11^2-1)}$$

$$= 1 - \frac{6[160 + 0.5 \times 5]}{1320} = 1 - \frac{6(162.5)}{1320} = 1 - \frac{162.5}{220} = \frac{57.5}{220} = 0.26136$$

$$R_{BC} = 1 - \frac{6\left[\sum D_{BC}^2 + \frac{\sum(m^3 - m)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[102.5 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right]}{11(11^2-1)}$$

$$= 1 - \frac{6[102.5 + 2]}{1320} = 1 - \frac{6 \times 104.5}{1320} = 1 - \frac{104.5}{220} = \frac{115.5}{220} = 0.525$$

EMARK. Sometimes we are given data in the form of ranks but the highest rank in the ...es exceeds the number of pairs of observations. In such situations ranks are treated ...values and then fresh ranks are determined. This is illustrated in the following example.

AMPLE 55. *Calculate the coefficient of correlation from the following data by the method of ...k differences :*

ank of X	:	10	4	2	5	8	5	6	9
ank of Y	:	10	6	2	5	8	4	5	9

[C.A. Foundation, May 1994]

ILUTION. Though the data is given in the form of ranks but it cannot be used as ranks as highest rank exceeds the number of pairs of observations. Treating the ranks as ...alues, we assign the fresh ranks. Moreover, certain items in both the series are repeated. ...ranking is done in accordance with suitable average.

Correla

X
10
4
2
5
8
5
6
9
n = 8

Note th... repeate... Thus, ...

6.16 I...

Merits...
1. It...
2. T...
3.

Den...
1.

CALCULATION OF RANK CORRELATION COEFFICIENT

		Rank assigned R_2	D $R_1 - R_2$	D^2
10	1	10	1	0	0
4	7	6	4	3	9
2	8	2	8	0	0
5	5.5	5	5.5	0	0
8	3	8	3	0	0
3	6.5	4	7	-1.5	2.25
6	4	5	5.5	-1.5	2.25
9	2	9	2	0	0
$n=8$		$n=8$			$\sum D^2 = 13.50$

Note that in series X the item 5 is repeated twice (*i.e.*, $m_1 = 2$). In series Y, the item 5 is repeated twice (*i.e.*, $m_2 = 2$).

Thus, Spearman's coefficient of rank correlation is given by

$$\rho = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[13.5 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2)\right]}{8(8^2-1)} = 1 - \frac{6\times 14.5}{8\times 63}$$

$$= 1 - 0.173 = 0.827.$$

6.16 MERITS AND DEMERITS OF SPEARMAN'S RANK CORRELATION METHOD

Merits. The rank correlation method has the following merits:

1. *It is easy to understand and simple to apply.*
2. *The Spearman's rank correlation method is the only method that can be used to find correlation coefficient if we are dealing with data of qualitative characteristics like beauty, intelligence, honesty, etc.*
3. *This is the only method that can be used where we are given the ranks and not the actual bivariate data on two variables.*

Demerits. The rank correlation method has the following limitations.

1. *This method cannot be used for finding correlation in the case of bivariate frequency distribution.*
2. *This method is very difficult to apply when the number of items is more than 30.*

List Of Formulae for Chapter 3 & 4 & 5.

Arithmetic Mean (Denoted by $\bar{X}$)

Individual Series	Discrete Series	Continuous Series
Direct Method	Direct Method	Direct Method
$\bar{X} = \frac{\Sigma X}{N}$	$\bar{X} = \frac{\Sigma fx}{N}$	$\bar{X} = \frac{\Sigma fm}{N}$
Shortcut Method	Shortcut Method	Shortcut method
$\bar{X} = A + \frac{\Sigma d}{N}$	$\bar{X} = A + \frac{\Sigma fd}{N}$	$\bar{X} = A + \frac{\Sigma df}{N}$
Stepdeviation Method	Stepdeviation Method	Stepdeviation
$\bar{X} = A + \frac{\Sigma d}{N} \times i$	$\bar{X} = A + \frac{\Sigma fd}{N} \times i$	$\bar{X} = A + \frac{\Sigma fd}{N} \times i$

Median. (Denoted by M)

Individual Series	Discrete Series	Continuous Series
Size of $\frac{N+1}{2}$ th item.	Size of $\frac{N+1}{2}$ th item.	Size of $\frac{N+1}{2}$ th item. $M = l_1 + \frac{\frac{N}{2} - c.f.}{f} \times i$

Empirical Mode : Mode = 3 Median - 2 Mean

Individual Series	Discrete Series	Continuous Series
Either by inspection or the value that occurs largest Number of times.	Grouping method determines that value around which most of the frequencies are concentrated.	$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$ l_1 = lower limit of mode class f_1 = frequency of mode class. f_0 = frequency of preceding mode class f_2 = frequency of succeding i = size of class interval.

Measures of Dispersion :

1. Range (Denoted by R)

Individual Series	Discrete Series	Continuous Series
Range = Largest - Smallest $R = L - S$	$R = L - S$	$\therefore R = L - S$
Coefficient of Range $= \frac{L-S}{L+S}$	Coefficient of Range $= \frac{L-S}{L+S}$	Coff of Range = $\frac{L-S}{L+S}$

Mean Deviation (Denoted by MD)

Individual Series	Discrete series	continuous series
Mean deviation using Mean MD$\bar{x}$	Mean deviation using Mean MD$\bar{x}$	Mean deviation using mean MD$\bar{x}$
$MD_{\bar{x}} = \frac{\Sigma \lvert x-\bar{x} \rvert}{N}$	$MD_{\bar{x}} = \frac{\Sigma f \lvert x-\bar{x} \rvert}{N}$	$MD_{\bar{x}} = \frac{\Sigma f \lvert m-\bar{x} \rvert}{N}$
x = series N = No. of terms $\bar{x}$ = Mean.	f = frequency $\bar{x}$ = Mean N = No. of terms (Σf)	f = frequency $m = \frac{L_1+L_2}{2}$ $\bar{x}$ = mean $N = \Sigma f$ sum of freq
Coefficient of Mean deviation MD$\bar{x}$	Coefficient of Mean deviation MD$\bar{x}$	Coefficient of mean deviation
coff of $MD\bar{x} = \frac{MD\bar{x}}{\bar{x}}$	coff of $MD\bar{x} = \frac{MD\bar{x}}{\bar{x}}$	coff of $MD\bar{x} = \frac{MD\bar{x}}{\bar{x}}$
$\bar{x}$ = Mean	$\bar{x}$ = Mean	$\bar{x}$ = Mean
Mean deviation using Median MDM	Mean deviation using Median MDM	MD using Median MDM
$MD_M = \frac{\Sigma \lvert x-M \rvert}{N}$	$MD_M = \frac{\Sigma f \lvert x-M \rvert}{N}$	$MD_M = \frac{\Sigma f \lvert m-M \rvert}{N}$
M = Median N = No. of terms.	f = frequency M = Median N = Sum of freq	f = frequency $m = \frac{L_1+L_2}{2}$ M = Median N = sum of freq
coff of $MD_{\bar{M}}$	coff of $MD_{\bar{M}}$	coff of $MD_{\bar{M}}$
coff. of $MD_{\bar{M}} = \frac{MD_M}{M}$	coff of $MD_{\bar{M}} = \frac{MD_M}{M}$	coff of $MD_{\bar{M}} = \frac{MD_M}{M}$

Individual Series	Discrete series	Continuous series
Actual Mean Method	**Actual Mean Method**	**Actual Mean Method**
$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$	$\sigma = \sqrt{\frac{fx^2}{N}}$	$\sigma = \sqrt{\frac{fx^2}{N}}$
$x = X - \bar{X}$	$x = X - \bar{X}$	$x = m - \bar{X}$
$\bar{X}$ = mean	$\bar{X}$ = mean	$m = \frac{l_1 + l_2}{2}$
N = No. of terms.	N = sum of frequency	$\bar{X}$ = mean
		N = sum of freq.
Assumed Mean	**Assumed Mean**	**Assumed Mean**
$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$	$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$	$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$
$d = X - A$	$d = X - A$	$d = m - A$
A = assumed mean	A = assumed mean	$m = l_1 + l_2/2$
N = No. of terms.	N = sum of freq.	A = Assumed mean
		N = sum of freq
Step deviation method	**Step deviation Method**	**Step deviation Method**
	$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$	$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$
	$d' = d/i$	$d' = d/i$
	$d = X - A$	$d = X - A$
	i = size of interval.	i = size of
	N = sum of freq.	N = sum of

Coefficient of Standard Deviation :-

Individual Series	Discrete Series	Continuous series
Coff of SD = $\frac{SD}{\bar{X}} = \frac{\sigma}{\bar{X}}$	Coff of SD = $\frac{SD}{\bar{X}} = \frac{\sigma}{\bar{X}}$	Coff of SD = $\frac{SD}{\bar{X}} = \frac{\sigma}{\bar{X}}$
$\bar{X}$ = mean	$\bar{X}$ = Mean	$\bar{X}$ = Mean.
σ = standard deviation (σ).		

Variance

Variance = $\left(\text{Standard Deviation}\right)^2 = \sigma^2$

(For all individual series, discrete series & continuous series).

Coff of Variance = $\frac{SD}{\bar{X}} \times 100$

SD = σ = standard deviation

$\bar{X}$ = Mean.

Measures of Correlation :-

1). Karl Pearson's Coefficient of Correlation :

Denoted by r.

Method 1 :

Method 2 : Assumed mean method

$$r = \frac{N\Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

$dx = X - A$; $dy = y - A$; A = Assumed mean

Method 3 : Direct Method

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

N = No. of terms

Method 4 : Covariance Method

$$r = \frac{\text{Covariance } (x,y)}{\sqrt{\text{Variance } x}\sqrt{\text{Variance } y}}$$

2. Rank Coefficient Co-relation :

(Denoted by R)

- Situation 1 : Ranks are Given

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

D = diff between 2 ranks

N = No. of observation

- Situation 2 : Ranks are Not Given

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

- Situation 3 : When Equal ranks are Given

$$R = 1 - 6\left[\frac{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)}{N^3 - N}\right]$$

D = diff. between 2 ranks

N = No. of observation

m = No. of items of Equal rank

Unit 6
Time series analysis

INTRODUCTION

One of the most important tasks before economists and businessmen these days is to make estimates for the future. For example, a businessman is interested in finding out his likely sales in the year 2006 or as a long-term planning in 2010 or the year 2020 so that he could adjust his production accordingly and avoid the possibility of either unsold stocks or inadequate production to meet the demand. Similarly, an economist is interested in estimating the likely population in the coming year so that proper planning can be carried out with regard to food supply, jobs for the people, etc. However, the first step in making estimates for the future consists of gathering information from the past. In this connection one usually deals with statistical data which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as 'time series'. Thus when we observe numerical data at different points of time the set of observations is known as time series. For example, if we observe production, sales, population, imports, exports, etc., at different points of time, say, over the last 5 or 10 years, the set of observations formed shall constitute time series. Hence, in the analysis of time series, time is the most important factor because the variable is related to time which may be either year, month, week, day, hour or even minutes or seconds.

A few definitions of time series are given below :

1. "A time series is a set of statistical observations arranged in chronological order." —*Morris Hamburg*

2. "A time series consists of statistical data which are collected, recorded and observed over successive increments of time." —*Patterson*

3. "A time series may be defined as a collection of magnitudes belonging to different time periods, of some variable or composite of variables, such as production of steel, per capita income, gross national product, price of tobacco, or index of industrial production." —*Yo-Lun-Chou*

4. "When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series." —*Wessel & Wellet*

5. "A time series is a set of observations taken at specified times, usually at 'equal intervals'. Mathematically, a time series is defined by the values $Y_1, Y_2, \ldots$ of a variable Y (temperature, closing price of a share, etc.) at times $t_1, t_2, \ldots$. Thus Y is a function of t symbolised by $Y = F(t)$." —*Spiegel*

It is clear from the above definitions that time series consist of data arranged chronologically. Thus if we record the data relating to population, per capita income, prices, production, etc., for the last 5, 10, 15, 20 years or some other time period, the series so emerging would be called time series.

It should be noted that the term 'time series' is usually used with reference to economic data and the economists are largely responsible for the development of the techniques of time series analysis. However, the term 'time series' can apply to all other phenomena that are related to time such as the number of accidents occurring in a day, the variation in the temperature of a patient during a certain period, number of marriages taking place during a certain

economic and business data. But these techniques can also be used in other natural or social sciences.

The problem of time series analysis can best be appreciated with the help of the following example :

The following are the figures of sales of refrigerators of a firm in thousand units :

Year	Sales of Firm A ('000)	Year	Sales of Firm A ('000)
1996	40	2000	43
1997	42	2001	48
1998	47	2002	65
1999	41	2003	42

If we observe the above series we find that generally the sales have increased but for two years a decline is also noticed. The statistician, therefore, tries to analyse the effect of the various forces under four broad heads :

(1) Changes that have occurred as a result of general tendency of the data to increase or decrease, known as 'secular movements'.

(2) Changes that have taken place during a period of 12 months as a result of change in climate, weather conditions, festivals, etc. Such changes are called 'seasonal variations'.

(3) Changes that have taken place as a result of booms and depressions. Such changes are classified under the head 'cyclical variations'.

(4) Changes that have taken place as a result of such forces that could not be predicted like floods, earthquakes, famines, etc. Such changes are classsified under the head 'irregular or erratic variations'.

These are called components of time series and shall be discussed in detail.

UTILITY OF TIME SERIES ANALYSIS

The analysis of time series is of great significance not only to the economist and businessman but also to the scientist, astronomist, geologist, sociologist, biologist, research worker, etc., for reasons given below :

1. *It helps in understanding past behaviour.* By observing data over a period of time one can easily understand what changes have taken place in the past. Such analysis will be extremely helpful in predicting the future behaviour.

2. *It helps in planning future operations.* Plans for the future cannot be made without forecasting events and relationship they will have. Statistical techniques have been evolved which enable time series to be analysed in such a way that the influences which have determined the form of that series may be ascertained. If the regularity of occurrence of any feature over a sufficiently long period could be clearly established then, within limits, prediction of probable future variations would become possible.

3. *It helps in evaluating current accomplishments.* The actual performance can be compared with the expected performance and the cause of variation analysed. For example, if expected sale for 2003-04 was 10,000

refrigerators and the actual sale was only 9,000, one can investigate the cause for the shortfall in achievement. Time Series analysis will enable us to apply the scientific procedure of "holding other things constant" as we examine one variable at a time. For example, if we know how much is the effect of seasonality on business we may devise ways and means of ironing out the seasonal influence or decreasing it by producing commodities with complementary seasons.

4. It facilitates comparison. Different time series are often compared and important conclusions drawn therefrom.

However, one should not be led to believe that by time series analysis one can foretell with 100 per cent accuracy the course of future events. After all, statisticians are not foretellers. This could be possible only if the influence of the various forces which affect these series such as climate, customs and traditions, growth and decline factors and the complex forces which produce business cycles would have been regular in their operation. However, the facts of life reveal that this type of regularity does not exist. But this then does not mean that time series analysis is of no value. When such analysis is coupled with a careful examination of current business indicators one can undoubtedly improve substantially upon guestimates (*i.e.*, estimates based upon pure guesswork) in forecasting future business conditions.

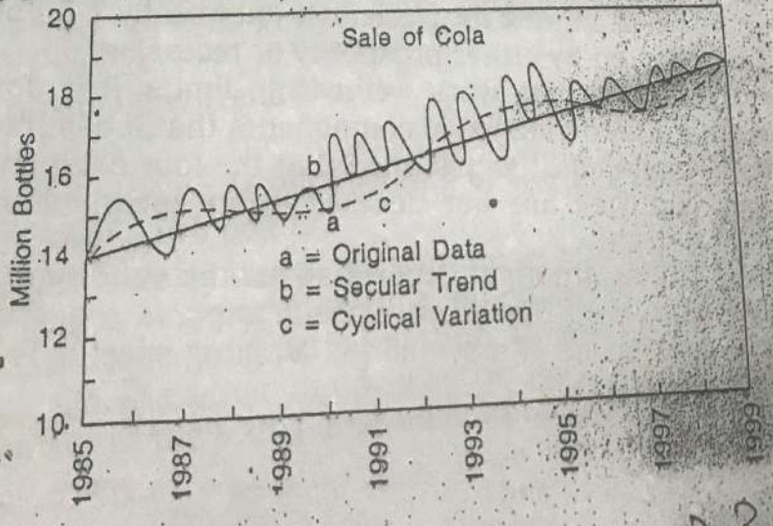
COMPONENTS OF TIME SERIES

It is customary to classify the fluctuations of a time series into four basic types of variations, which superimposed and acting all in concert account for changes in the series over a period of time. Those four types of patterns, movements, or, as they are often called, components or elements of a time series, are :

(1) Secular Trend
(2) Seasonal Variations
(3) Cyclical Variations
(4) Irregular Variations

It may be noted that any or all of these components may be present in any particular series.

The following graph gives the sale of Cola for the year 1985 to 1999.



235

The original data in this graph is represented by curve (a). The general movement over a long period of time represented by the diagonal line (b) drawn through the irregular curve is called *secular trend*.

Next, if we study the irregular curve year by year, we see that in each year the curve starts with a low figure and reaches a peak about the middle of the year and then decreases again. This type of fluctuation, which completes the whole sequence of changes within the span of a year and has about the same pattern year after year, is called *seasonal variation*.

Furthermore, looking at the broken curve superimposed on the original irregular curve, we find pronounced fluctuations moving up and down every few years throughout the length of the chart. These are known as *business cycles or cyclical fluctuations*. They are so called because they comprise a series of repeated sequence just as a wheel goes round and round.

Finally, the little saw-tooth irregularities on the original curve represent what are referred to as *irregular movements*.

In traditional or classical time series analysis, it is ordinarily assumed that there is a multiplicative relationship between these four components.. that is, it is assumed that any particular value in a series is the product of factor that can be attributed to the vaious components. Symbolically,

$$Y = T \times S \times C \times I$$

where Y denotes the result of the four elements : T = Trend; S = Seasonal Component; C = Cyclical Component; I = Irregular Component.

Another approach is to treat each observation of a time series as the sum of these four components. Symbolically,

$$Y = T + S + C + I$$

To prevent confusion between the two models it should be pointed out that in the multiplicative model S, C and I are indexes expressed as decimal per cents. In the additive model S, C and I are quantitative deviations about trend that can be expressed as seasonal, cyclical and irregular in nature.

Example. If in multiplicative model, $T = 400$, $S = 1.5$, $C = 1.2$ and $I = 0.8$ then :

$$Y = T \times S \times C \times I = 400 \times 1.5 \times 1.2 \times 0.8 = 576$$

If in the additive model, $T = 400$, $S = 120$, $C = 20$ and $I = -40$

Then $Y = 400 + 120 + 20 - 40 = 500$

The additive model assumes that all the components of the time series are independent of one another. For example, it assumes that trend has no effect on the seasonal component, no matter how high or low this value may become. Further, it assumes that the business cycle has no effect on the seasonal component. If the index for December is typically 1.50 or 150%, this per cent will not be affected by either prosperity or recession.

While the additive model may work well within limits, it is doubtful if one always can rely on the independence of components that it assumes.

In the multiplicative model, it is assumed that the four components are due to different causes but they are not necessarily independent and they can affect one another.

There is little agreement amongst experts about the validity of the different assumptions—some feel that the given classification is too crude and that here are more than four types of movements. Nothing specific is really known

There are numerous variations of these two basic models. Two such variations are :

$$Y = TCS + I \text{ and } Y = TC + SI$$

2.36

about how the components are related, how they combine to produce particular effects, or whether they are really separable. The effects of the various components might be additive, multiplicative or they might be combined in any one of indefinitely large number of other ways. Different models (assumptions or theories) will lead to different results. Although the additive assumption is undoubtedly true in some cases, the multiplicative assumption characterizes the majority of economic times series. Consequently, the multiplicative model is not only considered the standard or traditional assumption for time series analysis but it is more often employed in practice than all other possible models combined. For this reason, we shall use only the multiplicative model in our subsequent discussion.

The task of performing a time series analysis, just like the analysis of a chemist in breaking a substance into its constituent parts, is to operate on the data in such a way as to bring out separately each of the components present.

1. Secular Trend*

The term 'trend' is very commonly used in day-to-day parlance. For example, we often talk of rising trend of population, prices, etc. Trend also called secular or long-term trend, is the basic tendency of production, sales, income, employment, etc., to grow or decline over a period of time. The concept of trend does not include short-range oscillations but rather steady movements over a long period of time.

Secular trend movements are attributable to factors such as population change, technological progress and large-scale shifts in consumer tastes. The presence of more people means that more food, clothing, housing are necessary. Technological changes, discovery and exhaustion of natural resources, mass production methods, improvements in business organisation, land government intervention in the economy are other major causes for the growth or decline of many economic time series. In some cases, growth in one series involves decline in another, for example, the displacement of silk by rayon, the bullock-carts by other modes of transport like trucks, tempo, etc. Similarly, better medical facilities, improved sanitation, diet, etc., on the one hand, reduce the death rate and on the other contribute to a rise in birth rate.

There are all sorts of trends : some series increase slowly and some increase fast, others decrease at varying rates, some remain relatively constant for long periods of time, and some after a period of growth or decline reverse themselves and enter a period of decline or growth. Broadly speaking, the various types of trends are divided under two heads:

- Linear or Straight Line Trends; and
- Non-linear Trends.

For a proper understanding of the meaning of trend, the reader's attention is directed to the following two points :

(i) When we say that secular trend refers to the general tendency of the data to grow or decline over a long period of time, one may be interested in finding out as to what constitutes a long period of time. Does it mean several years? The answer is 'no'. On the other hand, whether a particular period can be regarded as long or not in the study of secular trend depends upon the nature

* The word secular is derived from the Latin

237

of the data [illegible] example, if we are studying the figures of sales of a firm for [illegible] and we find that in 1999 the sales have gone up, this increase [illegible] called as secular trend because this is too short a period of time to conclude that the sales are showing an increasing tendency. On the other hand, if we put a strong germicide into bacterial culture, and count the number of organisms still alive after each 10 seconds for 8 minutes, these 40 observations showing a general pattern would be called secular movement. It is clear from this example that in one case 2 years could not be regarded as a long period whereas in another case even 8 minutes constitute a long period. Hence, the nature of the data would dictate whether a particular period would be called as long or not.

Generally speaking, the longer the period covered, the more significant the trend. When the period is short, the secular movements cannot be expected to reveal themselves clearly and the general drift of the series may be unduly influenced by the cyclical fluctuations. This would make it difficult to separate the various series of variations in time series. *As a minimum safeguard, it may be said that to compute trend the period must cover at least two or three complete cycles.*

(ii) Another point worth mentioning is that for concluding whether the data is showing an upward tendency or downward tendency, *it is not necessary that the rise or fall must continue in the same direction throughout the period.* We have to observe the general tendency of the data. As long as we can say that the period as a whole was characterized by an upward movement or by a downward movement, we say that a secular trend was present. For example, if we observe the trend of price over a period or 20 years and find that except for a year or two the prices are continuously rising, we would call it a secular rise in prices.

2. Seasonal Variations*

Seasonal variations are those periodic movements in business activity which occur regularly every year and have their origin in the nature of the year itself. Since these variations repeat during a period of 12 months they can be predicted fairly accurately. Nearly every type of business activity is susceptible to seasonal influence to a greater or lesser degree and as such these variations are regarded as normal phenomenon recurring every year. Although the word 'seasonal' seems to imply a connection with the season of the year, the term is meant to include any kind of variation which is of periodic nature and whose repeating cycles are of relatively short duration. Seasonal variation is evident when the data are recorded at weekly or monthly or quarterly intervals. Although the amplitude of seasonal variations may vary, their period is fixed being one year. As a result, seasonal variations do not appear in series of annual figures. The factors that cause seasonal variations are :

(i) Climate and weather conditions. The most important factor causing seasonal variations is the climate. Changes in the climate and weather conditions such as rainfall, humidity, heat, etc., act on different products and industries differently. For example, during winter there is greater

* The seasonal variation in a time series is the repositive, recurrent pattern of change which occurs within a year or shorter time period. --Bllerson

demand for woollen clothes, hot drinks, etc., whereas in summer cotton clothes, cold drinks have a greater sale. Agriculture is influenced very much by the climate. The effect of the climate is that there are generally two seasons in agriculture—the growing season and harvesting season—which directly affect the income of the farmer which, in turn, affects the entire business activity.

(ii) Customs, traditions and habits. Though nature is primarily responsible for seasonal variations in time series, customs, traditions and habits also have their impact. For example, on certain occasions like Deepawali, Dussehra, Christmas, etc., there is a big demand for sweets and also there is a large demand for cash before the festivals because people want money for shopping and gifts. Similarly, on the first of every month there are heavy withdrawals and the bankers have to keep lots of cash to meet the possible demand on the basis of last month's experience. To take another example, most of the students buy books in the first few months of the opening of schools and colleges and thus the sale of books, stationery, etc., shows seasonal swings.

The study and measurement of seasonal patterns constitute a very important part of analysis of a time series. In some cases, seasonal patterns themselves are of primary concern because little, if any, intelligent planning or scheduling (of production, inventory, personnel, advertising and the like) can be done without a knowledge based on adequate statistical measures of seasonal patterns. In other cases the seasonal variation may not be of immediate concern, but it must be measured to facilitate the study of other types of variations based on adequate statistical measure of seasonal patterns. An accurate knowledge of seasonal behaviour is an aid in mitigating and ironing out seasonal movements through business policy. This may be done by introducing diversified products having different seasonal peaks, accumulating stock in slack seasons in order to manufacture at a more regular rate, cutting prices in slack seasons and advertising off-seasonal use for the products. Seasonal indices are also helpful in scheduling purchases, inventory control, personnel requirements, seasonal financing and selling and advertising programmes. For example, a housewife may buy fruits for canning or preserving at the peak of the season when the prices are low and quality high. Seasonal fluctuations may also be ironed out in order that the intra-year fluctuations may be less pronounced. Thus, attempts were made in U.S.A. to build up winter demand for ice-cream by advertising "Ice cream is one of your best foods. Eat one plate a day."

3. Cyclical Variations*

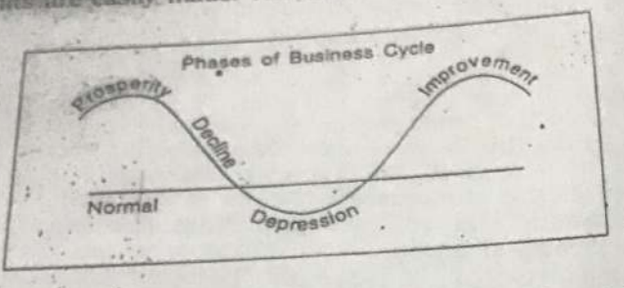
The term 'cycle' refers to the recurrent variations in time series that usually last longer than a year and are regular neither in amplitude nor in length.

Most of the time series relating to economics and business show some kind of cyclical or oscillatory variation. Cyclical fluctuations are long term movements that represent consistently recurring rises and declines [illegible] business cycle* consists of the recurrence of the [illegible]

239

business activity has some sort of statistical trend or "normal." By "normal" we mean some kind of statistical average ; we do not mean that there is anything "ideal" or desirable or special. There are four well-defined periods or phases in the business cycle, namely : (i) prosperity, (ii) decline, (iii) depression, and (iv) improvement.

Each phase changes gradually into the phase which follows it in the order given. The following diagram would illustrate a cycle. In the prosperity phase of the business cycle the public is optimistic—business is booming, prices are high and profits are easily made. There is a considerable expansion of business



activity which leads to an over-development. It is then difficult to secure deliveries and there is shortage of transportation facilities, which has a tendency to cause large inventories to be accumulated during the time of highest prices. Wages increase and labour efficiency decreases. The strong demand for money causes interest rates to rise to a high leavel while doubt enters the banker's mind as to the advisability of granting further loans. This situation causes businessmen to make price concessions in order to secure the necessary cash. Then follows the expectation of further reductions and the situation becomes worse instead of better. Buyers wait for lower prices and all this leads to a decline in business activity. Then follows period of pessimism in trade and industry ; factories close, businesses fail, there is widespread unemployment while wages and prices are low. These conditions characterize the period of depression. After a period of rigid economy, liquidation and reorganisation, money accumulates and seeks a period of improvement or recovery. The improvement period generally develops into the prosperity period and a business cycle is completed. The movements discussed above are constantly repeated in the order given as the cycle completes its swing.

The study of cyclical variations is extremely useful in framing suitable policies for stabilizing the level of business activity, *i.e.*, for avoiding periods of booms and depressions as both are bad for an economy—particularly depression which brings about a complete disaster and shatters the economy.

Business cycles are a type of fluctuations found in the aggregate economic activity of motions that organize their work mainly in business expenses: a cycle consists of expansions occurring at about the same time in many economic activities followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle: this sequence of changes is recurrent but not 'periodic: in duration business cycles vary from more than one year to ten or twelve years. they are not divisible into shorter cycles of similar charcter with amplitudes approximating their own."

--*Arthur Burns and Miller*

240

But despite the great importance of measuring cyclical variations, they are the most difficult type of economic fluctuations to measure. It is because of the following two reasons :

(*i*) Business cycles do not show regular periodicity—they differ widely in timing, amplitude and pattern which makes their study very tough and tedious.

(*ii*) Business cyclical variations are mixed with erratic, random or irregular forces which make it impracticable to isolate separately the effect of cyclical and irregular forces.

Business cycles are distinguished from seasonal variations in the following respects :

(*i*) The cyclical variations are of a longer duration than a year. A business cycle may be of any duration but normally the period of business cycle is 2–10 years. Moreover, they do not ordinarily exhibit regular periodicity as successive cycles vary widely in timing, amplitude and pattern.

For example, the 23 cycles of general business in the United States between 1854 and 1949 averaged 49 months; in duration individual cycles differed greatly—the shortest lasted only 29 months and the longest persisted for 99 months.

(*ii*) The fluctuations in a business cycle result from a different set of causes. The period of prosperity, decline, depression and improvement viewed as four phases of a business cycle are generated by factors other than weather, social customs, and those which create seasonal patterns.

4. Irregular Variations*

Irregular variations, also called 'erratic', accidental, random, refer to such variations in business activity which do not repeat in a definite pattern. In fact the category labelled irregular variation is really intended to include all types of variations other than those accounting for the trend, seasonal and cyclical movements. These latter three, if they are actually at work, act in such a way as to produce certain systematic effects. Irregular movements, on the other hand, are considered to be largely random, being the result of chance factors which, like those determining the fall of a coin, are wholly unpredictable.

Irregular variations are caused by such isolated special occurrences as floods, earthquakes, strikes and wars. Sudden changes in demand or very rapid technological progress may also be included in this category. By their very nature these movements are very irregular and unpredictable. Quantitatively it is almost impossible to separate out the irregular movements and the cyclical movements. Therefore, while analysing time series the trend and seasonal variations are measured separately and the cyclical and irregular variations are left altogether.

There are two reasons for recognizing irregular movements :

(*i*) To suggest that on occasions it may be possible to explain certain movements in the data due to specific causes and to simplify further analysis.

(*ii*) To emphasise the fact that predictions of economic conditions which are always subject to degree of error owing to the unpredictable erratic influences which may enter.

Although it is of little matter to classify the factors affecting time series into these four groups for analytical purposes, the actual application of the classification [illegible] presents serious problems. Seasonal variations are by no means always so uniform in amplitude and timing that their identification can be made with certainty. Consequently, the investigator is often hard put to distinguish seasonal influences from cyclical or random factors. Long and severe cycles may, to some observers, appear to be changes in the direction of the regular trend. During the great depression of the 1930's, for example, many leading economists interpreted the existing conditions not as a cyclical depression but as 'secular stagnation'.

Another difficulty arises because the four components of time series data are not mutually independent of one another. An exceedingly severe seasonal influence may aggravate or even precipitate a change in the cyclical movement. Conversely, cyclical influence may seriously affect the seasonal. A very rapidly rising trend virtually eliminates seasonal and cyclical variations.

Finally, the fourfold breakdown of time series data when applied to general economic conditions has frequently been challenged on analytical grounds. Bratt* sees not one trend, but two : a primary trend representing the long-term growth of productive capacity and the drift away from it, which he calls secondary trend. Schumpeter developed an even more detailed breakdown by identifying three cyclical components, the 3-year Kitchin cycle, the 10-year Juglar cycle and the 50-year Kondratieff cycle.† The divergence of opinion among eminent scholars indicates clearly that the fourfold breakdown is mere approximation, convenient to employ but frequently subject to modification.

PRELIMINARY ADJUSTMENTS BEFORE ANALYSING TIME SERIES

Before beginning the actual work of analysing a time series it is necessary to make certain adjustments in the raw data. The adjustments may be needed for :

- Calendar Variations.
- Population Changes.
- Price Changes.
- Comparability.

(i) Calendar Variations. A vast proportion of the important time series is available in a monthly form and it is necessary to recognise that the month is a variable time unit. The actual length of the shortest month is about 10 per cent less than that of the longest, and if we take into account holidays and weekends, the variation may be even greater. Thus, the production or sales for the month of February may be less not because of any real drop in activity but because of the fact that February has fewer days. Thus the purpose of adjusting for calendar variation is to eliminate certain spurious differences which are caused by peculiarities of our calendar. The adjustment for calendar variations is made by dividing each monthly total by the number of days in the month (sometimes by the number of working days in the month) thus arriving at daily average

---

* Emler C. Bratt : *Business Cycles and Forecasting.*
† J.A. Schumpeter : *Business Cycles.*

Merits and Limitations

Merits. This method is associated with the following advantages:

- This method is simple to understand as compared to the moving average method and the method of least squares.
- This is an objective method of measuring trend as everyone who applies the method is bound to get the same result (of course, leaving aside the arithmetical mistakes).

Limitations. Though a simple and objective, this method has some limitations too. These are :

- This method assumes straight line relationship between the plotted points regardless of the fact whether that relationship exists or not.
- The limitations of arithmetic average shall automatically apply. If there are extremes in either half or both halves of the series, then the trend line will not give a true picture of the growth factor. This danger is greatest when the time period represented by the average is small. Consequently, trend values obtained are not precise enough for the purpose either of forecasting the future trend or of eliminating trend from original data.

For the above reasons if the arithmetic averages of the data are to be used in estimating the secular movement, it is sometimes better to use moving averages than semi-averages.

Method of Moving Averages*

When a trend is to be determined by the method of moving averages, the average value for a number of years (or month or weeks) is secured, and this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the averages. The effect of averaging is to give a smoother curve, lessening the influence of the fluctuations that pull the annual figures away from the general trend.

While applying this method, it is necessary to select a period for moving average such as 3-yearly moving average, 5-yearly moving average, 8-yearly moving average, etc. The period of moving average is to be decided in the light of the length of the cycle. Since the moving average method is most commonly applied to data which are characterised by cyclical movements, it is necessary to select a period for moving average which concides with the length of the cycle, otherwise the cycle will not be entirely removed. The danger is more severe, the shorter the time period represented by the average. When the period of moving average and the period of the cycle do not coincide, the moving average will display a cycle which has the same period as the cycle in the data, but which has less amplitude than the cycle in the data. Often we find that the cycles in the data are not of uniform length. In such a case we should take a moving average period equal to or somewhat greater than the average period of the cycle in the data. Ordinarily the necessary period will

... and ten years for general business series but even longer ... certain types of data.
... average shall be computed as follows :

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3}, \ldots$$

and for 5-yearly moving average :

$$\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5}, \ldots\ldots$$

Illustration 5. (a) Calculate the 3-yearly moving averages of the production figures given below and draw the trend :

Year	Production (in m. tonnes)	Year	Production (in m. tonnes)
1989	15	1997	63
1990	21	1998	70
1991	30	1999	74
1992	36	2000	82
1993	42	2001	90
1994	46	2002	95
1995	50	2003	102
1996	56		

Solution.

CALCULATION OF 3-YEARLY MOVING AVERAGES

Year	Production (in m. tonnes)	3-yearly total (in m. tonnes)	3-yearly moving average
1989	15	—	—
1990	21	66	22.00
1991	30	87	29.00
1992	36	108	36.00
1993	42	124	41.33
1994	46	138	46.00
1995	50	152	50.67
1996	56	169	56.33
1997	63	189	63.00
1998	70	207	69.00
1999	74	226	75.33
2000	82	246	82.00
2001	90	267	89.00
2002	95	287	95.67
2003	102	—	—

(b) Construct 5-yearly moving averages of the number of students studying in a college shown below :

Year	No. of students	Year	No. of students
1994	332	1999	405
1995	317	2000	410
1996	357	2001	427
1997	392	2002	405
1998	402	2003	438

(B.Com., Madras Univ.; B. Com., Calcutta Univ.)

olution.

CALCULATION OF 5-YEARLY MOVING AVERAGES

Year	No. of students	5-yearly total	5-yearly moving average
1994	332	—	—
1995	317	—	—
1996	357	1800	360.0
1997	392	1873	374.6
1998	402	1966	393.2
1999	405	2036	407.2
2000	416	2049	409.8
2001	427	2085	417.0
2002	405	—	—
2003	438	—	—

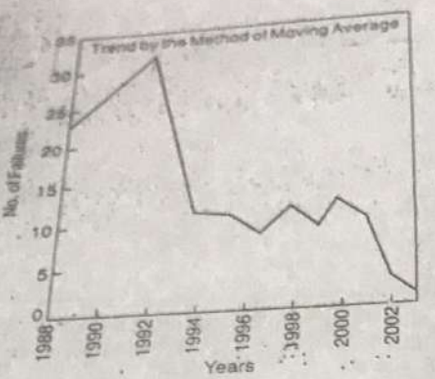
lustration 6. Calculate 5-yearly and 7-yearly moving averages for the following data of the mbers of commercial and industrial failures in a country during 1988 to 2003:

Year	No. of failures	Year	No. of failures
1988	23	1996	9
1989	26	1997	13
1990	28	1998	11
1991	32	1999	14
1992	20	2000	12
1993	12	2001	9
1994	12	2002	3
1995	10	2003	1

Also plot the actual and trend values on a graph.

lution. CALCULATION OF 5-YEARLY AND 7-YEARLY MOVING AVERAGES

Year	No. of failures	5-yearly moving total	5-yearly moving average	7-yearly moving total	7-yearly moving average
1988	23	—	—	—	—
1989	26	—	—	—	—
1990	28	129	25.8	—	—
1991	32	118	23.6	153	21.9
1992	20	104	20.8	140	20.0
1993	12	86	17.2	123	17.6
1994	12	63	12.6	108	15.4
1995	10	56	11.2	87	12.4
1996	9	55	11.0	81	11.6
1997	13	57	11.4	81	11.6
1998	11	59	11.8	78	11.1
1999	14	59	11.8	71	10.1
2000	12	49	9.8	63	9.0
2001	9	39	7.8	—	—
2002	3	—	—	—	—



Even Period of Moving Average If the moving average is an even period moving average, say, four-yearly or six-yearly, the moving total and moving average which are placed at the centre of the time span from which they are computed fall between two time periods. This placement is inconvenient since the moving average so placed would not coincide with the original time period. We, therefore, synchronise moving averages and original data. This process is called *centering* and always consists of taking a two-period moving average of the moving averages.

Illustration 7. Estimate the trend values using the data given by taking a four-yearly moving average :

Year	Value	Year	Value
1990	12	1997	100
1991	25	1998	82
1992	39	1999	65
1993	54	2000	49
1994	70	2001	34
1995	87	2002	20
1996	105	2003	7

(M.Com., Madras Univ.)

Solution. ESTIMATING THE TREND VALUES

Year	Value	4-yearly moving total	4-yearly moving average	4-yearly moving average centered
1990	12	—	—	—
1991	25	—	—	
		← 130	32.5	
1992	39			← 39.75
		← 188	47.0	
1993	54			← 54.75

There is another method of centering the moving averages. If we are calculating 4-yearly moving average, we will then take four-yearly totals and of these totals we will again take 2-yearly totals and divide these totals by 8.

Year	Value	4-yearly moving total	4-yearly moving average	4-yearly moving average centered
1994	70	← 250	62.5	← 70.75
1995	87	← 316	79.0	← 84.75
1996	105	← 362	90.5	← 92.00
1997	100	← 374	93.5	← 90.75
1998	82	← 352	88.0	← 81.00
1999	65	← 296	74.0	← 65.75
2000	49	← 230	57.5	← 49.75
2001	34	← 168	42.0	← 34.75
2002	20	← 110	27.5	
2003	7	—	—	

Illustration 8. Assume a four-yearly cycle and calculate the trend by the method of moving averages from the following data relating to the production of tea in India:

Year	Production (m. lbs.)	Year	Production (m. lbs.)
1994	464	1999	540
1995	515	2000	557
1996	518	2001	571
1997	467	2002	586
1998	502	2003	612

(M.Com., Madras Univ.; B.Com., MD Univ.)

Solution. CALCULATION OF TREND BY THE MOVING AVERAGE METHOD

Year	Production (m. lbs.)	4-yearly moving totals	4-yearly moving average	4-yearly moving average centered
1994	464	—	—	—
1995	515	← 1964	491.00	← 495.75
1996	518	← 2002	500.50	← 508.62
1997	467	← 2027	506.75	← 511.62
1998	502	← 2066	516.50	← 529.50
1999	540	← 2170	542.50	← 556.00
2000	557	← 2254	563.50	← 572.50
2001	571	← 2326	581.50	

Merits and Limitations

Merits. The method of moving averages has the following advantages :

- This method is simple as compared to the method of least squares.
- It is a flexible method of measuring trend for the reason that if a few more figures are added to the data, the entire calculations are not changed—we only get some more trend values.
- If the period of moving average happens to coincide with the period of cyclical fluctuations in the data, such fluctuations are automatically eliminated.
- The moving average has the advantage that it follows the general movements of the data and that its shape is determined by the data rather than the statistician's choice of a mathematical function.
- It is particularly effective if the trend of a series is very irregular.

Limitations. There are, however, some limitations of this method too. These are :

- Trend values cannot be computed for all the years. The longer the period of moving average, the greater the number of years for which trend values cannot be obtained. For example, in a three-yearly moving average, trend values cannot be obtained for the first year and last year, in a five-yearly moving average for the first two years and the last two years, and so on.
- Great care has to be exercised in selecting the period of moving average. No hard and fast rules are available for the choice of the period and one has to use his own judgment.
- Since the moving average is not represented by a mathematical function this method cannot be used in forecasting which is one of the main objectives of trend analysis.
- Although theoretically we say that if the period of moving average happens to coincide with the period of cycle, the cyclical fluctuations are completely eliminated, but in practice since the cycles are by no means perfectly periodic, the lengths of the various cycles in any given series will usually vary considerably and, therefore, no moving average can completely remove the cycle. The best result would be obtained by a moving average whose period was equal to the average length of all the cycles in the given series. However, it is difficult to determine the average length of the cycle until the cycles are isolated from the series.
- Finally, when the trend situation is not linear (a straight line) the moving average lies either above or below the true sweep of the data. Consequently, the moving average is appropriate for trend computations only when :
  (a) the purpose of investigation does not call for current analysis or forecasting,

The Method of least squares

This method is most widely used in practice. It is a mathematical method and with its help a trend line is fitted to the data in such a manner that the following two conditions are satisfied:

(1) $\Sigma (Y - Y_c) = 0$.

i.e., the sum of deviations of the actual values of Y and the computed values of Y is zero.

(2) $\Sigma (Y - Y_c)^2$ is least.

i.e., the sum of the squares of the deviations of the actual and computed values is least from this line and hence the name method of least squares. The line obtained by this method is known as the line of best fit.

The method of least squares may be used either to fit a straight line trend or a parabolic trend.

The straight line trend is represented by the equation

$$Y_c = a + bX$$

where Y_c is used to designate the trend values to distinguish them from the actual Y values, a is the Y intercept or the computed trend figure of the Y variable when $X = 0$. b represents the slope of the trend line or amount of change in Y variable that is associated with a change of one unit in X variable. The X variable in time series analysis represents time. Whenever we fit any straight line trend by the least squares method, three things should be specified :

(1) Which year was selected as the origin ?

(2) What is the unit of time represented by X ? Is it half year, one year or five years ?

(3) In what kind of units is Y being measured? Is it production in tonnes, sales in rupees, price in rupees, employment in thousands of workers ?

In order to determine the values of the constants a and b the following two normal equations are to be solved :

$$\Sigma Y = Na + b\Sigma X \quad \ldots(i)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \ldots(ii)$$

where N represents number of years (months or any other period) for which data are given.

It should be noted that the first equation is merely the summation of the given function, the second is the summation of X multiplied by the given function.

We can measure the variable X from any point of time in origin such as the first year. But the calculations are very much simplified when the mid-point in time is taken as the origin because in that case the negative values in the first half of the series balance out the positive values in the second half so that $\Sigma X = 0$. In other words, the time variable is measured as a deviation from its mean. Since $\Sigma X = 0$ the above two normal equations would take the form.

$$\Sigma Y = Na \quad \ldots(i)$$

$$\Sigma XY = b\Sigma X^2 \quad \ldots(ii)$$

The values of a and b can now be determined easily.

Since

$$\Sigma Y = Na$$

Since $a = \frac{\Sigma Y}{N}$ or $\bar{Y}$

$$\Sigma XY = b\Sigma X^2 \quad \therefore \quad b = \frac{\Sigma XY}{\Sigma X^2}$$

The constant 'a' is simply equal to the mean of Y values and the constant 'b' gives the rate of change.

It should be noted that in case of odd number of years, when the deviations are taken from the middle year ΣX would always be zero provided there is no gap in the data given. However, in case of even years also ΣX will be zero if the X origin is placed midway between the two middle years. For example, if the years are 1998, 1999, 2000, 2001, 2002 and 2003, we can take deviations from the middle year 2000.5. Thus the deviations would be – 2.5, – 1.5, – 0.5, + 0.5, + 1.5, + 2.5 for the various years and the total ΣX would be zero. Hence both in odd as well as in even number of years we can use the simple procedure of determining the values of the constants a and b.

The arithmetic straight line trend fitted by the method of least squares is by far the most widely applied trend curve. This particular trend curve is applicable for those series in which period-to-period changes are constant in absolute amount.

A quick method for assessing the appropriateness of the straight line model is the *method of first differences*. If the differences between successive observations of a series are constant (or nearly so), the arithmetic straight line should be taken to be an appropriate representation of the trend component. The method of first differences is explained below :

METHOD OF FIRST DIFFERENCES

Year	Sales ('000 units)	First differences
1996	120	
		20
1997	140	
		22
1998	162	
		18
1999	180	
		23
2000	203	
		22
2001	225	
		20
2002	245	
		23
2003	268	

It is clear from the above series that the differences in successive observations are nearly constant and hence the arithmetic straight line is an appropriate model for assessing the trend component of the series.

It may be pointed out that very few time series exhibit this type of constant change over a period of time, say, over a period of several business cycles. It is often necessary to fit other types of lines or curves.

Illustration 10. Below are given the figures of production (in thousand quintals) of a sugar factory :

Year	1997	1998	1999	2000	2001	2002	2003
Production (in '000 qtls.)	80	90	92	83	94	99	92

(i) Fit a straight line trend to these figures.

251

Solution. FITTING THE STRAIGHT LINE TREND

Year	Y	X	XY	X^2	Trend values Y_c
1997	90	−3	−240	9	84
1998	96	−2	−180	4	86
1999	92	−1	−92	1	88
2000	88	0	0	0	90
2001	94	+1	+94	1	92
2002	99	+2	+198	4	94
2003	92	+3	+276	9	96
N = 7	Σ Y = 630	Σ X = 0	Σ XY = 56	Σ X^2 = 28	Σ Y_c = 630

The equation of the straight line is $Y_c = a + bX$

Since $\Sigma X = 0$; $a = \frac{\Sigma Y}{N}$, $b = \frac{\Sigma XY}{\Sigma X^2}$

$\Sigma Y = 630$, $N = 7$, $\Sigma XY = 56$, $\Sigma X^2 = 28$,

$a = \frac{630}{7} = 90$; and $b = \frac{56}{28} = 2$

Hence the equation of the straight line trend is $Y_c = 90 + 2X$.

Origin, 2000 ; X units, one year; Y units, production in thousand quintals.

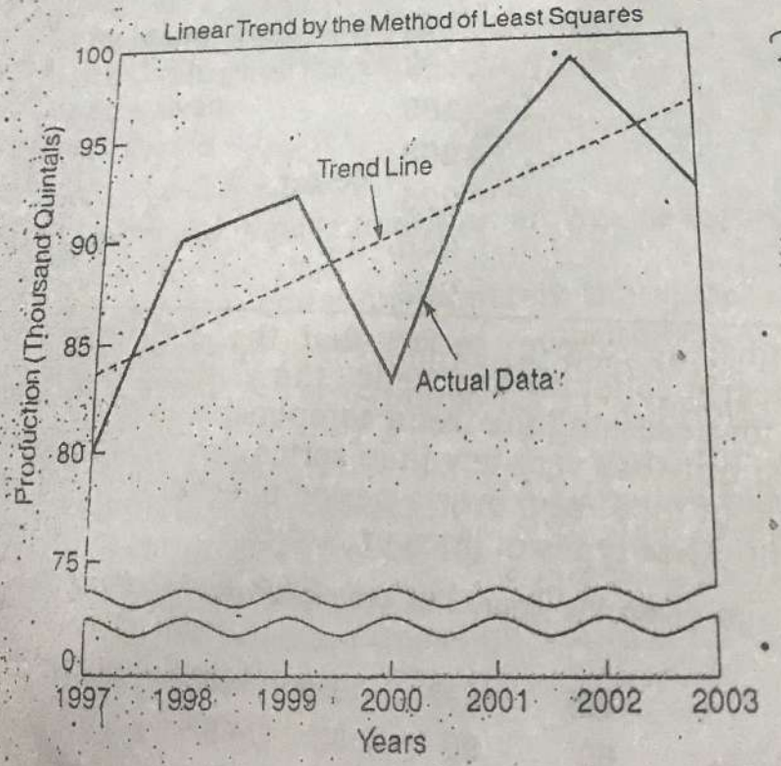
For $X = -3$, $Y_c = 90 + 2(-3) = 84$

For $X = -2$, $Y_c = 90 + 2(-2) = 86$

For $X = -1$, $Y_c = 90 + 2(-1) = 88$.

Similarly, by putting $X = 0$, 1, 2, 3, we can obtain other trend values. However, since the value of b is constant, first trend value need be obtained and then if the value b is positive we may continue adding the value of b to every preceding value. For 1998 it will be 84 + 2 = 86, for 1999 it will be 86 + 2 = 88, and so on. If b is negative then instead of adding we will deduct.

(ii) The graph of the above data is given below :



If instead of middle year as origin, we take first year as origin the solution would be as follows :

Year	Production (7000 qtls) Y	X	XY	X^2	Y_c
1997	80	0	0	0	84
1998	90	1	90	1	86
1999	92	2	184	4	88
2000	83	3	249	9	90
2001	94	4	376	16	92
2002	99	5	495	25	94
2003	92	6	552	36	96
$N = 7$	$\Sigma Y = 630$	$\Sigma X = 21$	$\Sigma XY = 1,946$	$\Sigma X^2 = 91$	$\Sigma Y_c = 630$

$Y_c = a + bX$; $\Sigma Y = Na + b\Sigma X$; $\Sigma XY = a\Sigma X + b\Sigma X^2$

Substituting the value $630 = 7a + 21b$ (i)

$1,946 = 21a + 91b$ (ii)

Multiplying Eqn. (i) by 3 : $1,890 = 21a + 63b$

$1,946 = 21a + 91b$

$-28b = -56$ or $b = 2$

Substituting the value of b in Eqn. (i) :

$630 = 7a + 21(2)$

$7a = 630 - 42 = 588$ or $a = 84$

Thus the equation is $Y_c = 84 + 2X$.

Origin 1997; X units, one year ; Y units, production in thousand quintals.

For $X = 0$, $Y = 84$

Note : The difference in the two equations is because of the difference in origin. In the first case 2000 was taken as origin whereas in the second case 1997 was taken as origin. However, end values are the same.

Illustration 11. Fit a straight line trend for the following series. Estimate the value for 2004.

Year	1997	1998	1999	2000	2001	2002	2003
Production of Steel (m. tonnes)	60	72	75	65	80	85	95

[B.A. (H) Econ., DU; M.Com., Madurai Univ.]

Solution :

		X		
1997	60	−3	−180	9
1998	72	−2	−144	4
1999	75	−1	−75	1
2000	65	0	0	0
2001	80	+1	+80	1
2002	85	+2	+170	4
	95	+3	+285	9
	$\Sigma Y = 532$	$\Sigma X = 0$	$\Sigma XY = 136$	$\Sigma X^2 = 28$

Since $\Sigma X = 0$; $a = \frac{\Sigma Y}{N} = \frac{532}{7} = 76$

Last Year Question Papers

265

Bharati Vidyapeeth (Deemed to be University)
Institute of Management and Research (BVIMR), New Delhi
Internal Backlog

Course: BCA
Subject: Statistics
Max. Marks: 40

Semester: IV
Course Code: 404
Max. Time: 2 Hours

Instructions :-
1) Question No. 1 is compulsory. Attempt any two questions from Q2 to Q5.
2) Attempt any two question from section 2.
3) Bring your own simple calculator and stationary during the exam.

Section 1

Answer in 400 words. Each question carry 06 marks.

Q1. The following data shows the marks obtained by 100 students in an examination:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	10	9	25	30	10	16

a) Construct a less than cumulative frequency distribution for the following data.
b) Construct a "less than" ogive of the cumulative frequency distribution of the above data.

Q2. a) Discuss the concept of median
b) Calculate median for following data:

Marks	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of Students	7	15	18	25	30	20	16	7	2

Q3. a) The mean weight of 100 students (boys and girls) in a class is 50 kg. The mean weight of boy students is 52 kg and that of girl students is 42 kg. Find the number of boys and girls in that class.
b) Calculate Karl Pearson coefficient of correlation between the marks in statistics and mathematics.

Marks in Statistics	66	68	69	72	65	59	62	67	61	71
Marks in Mathematics	65	64	67	69	64	60	59	68	60	64

Q4. Find the mean deviation and its coefficient about the median for the following data:

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of students	2	6	12	18	25	20	10	7

Q5. **Write Short Note on any two. Answer in 300 words. Each carry 03 marks.**

a) What is time series? Explain the components of time series.
b) What is the difference between correlation and regression analysis?
c) Discuss the concept of measure of central tendency and also discuss its different types.

Section 2

Answer in 800 words. Attempt any 2 questions. Each question carry 11 marks

Q6. An incomplete distribution is given below:

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
Frequency	12	30	x	65	y	25	18	229

a) Find out missing frequencies if median value is 46. : (8 Marks)
b) Calculate the arithmetic mean of the completed table. : (3 Marks)

Q7. a) Calculate mode from the following data (grouping method): (8 Marks)

Height in Inches	56	58	59	60	61	62	63	64	66	68
No of Persons	3	7	6	9	20	22	24	5	3	1

b) In a moderately skewed distribution, the value of mode is 120 and that of median is 140. Find the value of arithmetic mean. (3 Marks)

Q8. A purchasing agent obtained samples of lamps from two suppliers. He had the samples tested in his own laboratory for the lengths of life with the following results:

Length of life (in hours)	Company A	Company B
700-900	10	3
900-1100	16	42
1100-1300	26	12
1300-1500	8	3

a) Which company's lamps have greater average life?
b) Which company's lamps are more uniform?

Bharati Vidyapeeth (Deemed to be University)
Institute of Management and Research (BVIMR), New Delhi
2nd Internal Examination (2019)

BVIMR

Course: BCA
Subject: Statistics
Max. Marks: 40

Semester: IV
Course Code: 404
Max. Time: 2 Hours

Instructions:
1) Question No. 1 is compulsory. Attempt any two questions from Q2 to Q5.
2) Attempt any two question from section 2.
3) Bring your own simple calculator and stationary during the exam.

Section 1

Answer in 400 words. Each question carry 06 marks.

Q1. Calculate standard of deviation and coefficient of variation from the following data:

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	10	15	25	25	10	10	5

Q2. a) Find the regression coefficient B_{xy} and B_{yx} of X on Y and Y on X respectively. If S.D. of X and Y are 4 and 3 respectively and coefficient of correlation between X and Y is 0.8.
b) Find the mean deviation and its coefficient about the median for the following data:

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

Q3. The sales of a commodity (in '000 of Rs.) are given below:

Year	1999	2000	2001	2002	2003	2004	2005
Sales	82	86	81	86	92	90	99

a) Using the method of least square, fit a straight line trend equation to the data.
b) What are the expected sales for the year 2010?

Q4. An experiment conducted on 9 different cigarette smoking subjects resulted in the following data.

Subject Number	Cigarettes smoked per week	Number of years lived
1	25	63
2	35	68
3	10	72
4	40	62
5	85	65
6	75	46
7	60	51
8	45	60
9	50	55

Calculate the correlation of coefficient between the number of cigarettes smoked and the longevity of a test subject.

Q5. **Write Short Note on any two. Answer in 300 words. Each carry 03 marks**

a) Components of Time Series.
b) Types of Correlation.
c) Correlation v/s Regression Analysis.

Answer in 800 words. Attempt any 2 questions. Each question carry 11 marks

Section 2

Q6. The following is the record number of bricks laid each day for 10 days by two brick layers A and B. calculate the coefficient of variation in each case and discuss the relative consistency of the two brick layers.

A	700	675	725	625	650	700	650	700	600	650
B	550	600	575	550	650	600	550	525	625	600

If each of the values in respect of worker A is decreased by 10 and each of the values for worker B is increased by 50, how will it affect the results obtained earlier?

Q7. b) Estimate a) the sale for advertising of Rs. 100 lakhs
The advertisement expenses for sales of Rs. 47 crores from the data given below:

Sales (Rs. Crores)	14	16	18	20	24	30	32
Advertisement expenses (Rs. Lakhs)	52	62	65	70	76	80	78

Q8. Find out the three and five year moving averages. Also plot the original time series along with the 3 yearly and 5 yearly moving average.

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Sales(Rs. in million)	10	15	20	25	15	12	15	24	15	21	15	24

Bharati Vidyapeeth (Deemed to be University)
Institute of Management and Research (BVIMR), New Delhi
1st Internal Examination (2019)

Course: BCA
Subject: Statistics
Max. Marks: 40

Semester: IV
Course Code: 404
Max. Time: 2 Hours

Instructions:

1) Question No. 1 is compulsory. Attempt any two questions from Q2 to Q5.
2) Attempt any two question from section 2.
3) Bring your own simple calculator and stationary during the exam.

Section 1

Answer in 400 words. Each question carry 06 marks.

Q1. The following data shows the marks obtained by 100 students in an examination:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	10	9	25	30	10	16

a) Construct a less than cumulative frequency distribution for the following data.

b) Construct a "less than" ogive of the cumulative frequency distribution of the above data.

Q2. a) Discuss the concept of median

b) Calculate median for following data:

Marks	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of Students	7	15	18	25	30	20	16	7	2

Q3. a) Discuss three limitations of mean.

b) Calculate mean for the following data by short cut method.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students	10	9	25	30	16	10

Q4. The mean weight of 100 students (boys and girls) in a class is 50 kg. The mean weight of boy students is 52 kg and that of girl students is 42 kg. Find the number of boys and girls in that class.

Q5. **Write Short Note on any two. Answer in 300 words. Each carry 03 marks**

a) Define statistics and its importance in detail.
b) Explain discrete and continuous variables in detail with the help of suitable example.
c) Discuss the concept of measure of central tendency and also discuss its different types.

Section 2

Answer in 800 words. Attempt any 2 questions. Each question carry 11 marks

Q6. An incomplete distribution is given below:

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80	Total
Frequency	12	30	x	65	y	25	18	229

a) Find out missing frequencies if median value is 46. : (8 Marks)
b) Calculate the arithmetic mean of the completed table. : (3 Marks)

Q7. a) Calculate mode from the following data (grouping method): (8 Marks)

Height in Inches	56	58	59	60	61	62	63	64	66	68
No of Persons	3	7	6	9	20	22	24	5	3	1

b) In a moderately skewed distribution, the value of mode is 120 and that of median is 140. Find the value of arithmetic mean. (3 Marks)

Q8. a) Weight in Kgs. of 50 students in a class are given below: (6 Marks)

45	76	70	75	47	90	70	40	77	95
90	97	79	75	55	95	60	55	99	60
67	55	80	40	85	64	65	78	55	85
65	47	98	95	60	69	57	80	95	99
60	59	76	90	70	84	60	56	98	70

Prepare a frequency distribution table using class intervals as 40-45, 45-50 ... etc. and also find the relative frequency.

b) Discuss the different methods of Diagrammatic representation of frequency distribution. (5 Marks)

Bharati Vidyapeeth Deemed University,
Institute of Management and Research (BVIMR), New Delhi
1stInternal Examination (January, 2017)

Course BCA
Subject STATISTICS
Semester IV
Max. Marks: 40
Course Code 404
Max. Time: 2 Hours

Instructions (if any):-

Q. 1 Write short notes on any five questions. Answer in 50 words (Recall) [5 x 2]

a) Advantages of Statistics

b) Limitations of Statistics

c) Utility of Mode

d) Discrete and Continuous Variables

e) Distinguish between ordinal and cardinal scale

f) Arithmetic Mean

g) Bar Diagram

h) Frequency Polygon

Q. 2 attempt any two question. Answer in 200 words (Theoretical Concept) [2 x 5]

a) What are the various methods of graphical presentation of data?

b) Explain any two methods used for collecting primary data.

c) What are the chief characteristics of an ideal measure of central tendency?

Q.3 Attempt any two questions. Answer in 200 words (Practical/Application oriented) [2 x 5]

a) The areas of the various continents of the world (in millions of square miles) are as follows:

11.7 for Africa;
10.4 for Asia;
1.9 for Europe;
9.4 for North America;
3.3 Australia;
6.9 South America;
7.9 Antartica.

Draw a bar chart representing the above data.

Bharati Vidyapeeth Deemed University,
Institute of Management and Research (BVIMR), New Delhi
2nd Internal Examination (March, 2016)

Course : BCA
Subject : STATISTICS

Semester : IV
Course Code: 404

Max. Marks: 40

Max. Time: 2 Hours

Instructions: 1. Non programmable calculators are allowed
2. Draw diagrams wherever required.
3. All questions are compulsory, however internal choices have been provided.

Q. 1 Attempt any five questions. Answer in 50 words. [5 x 2]

a) Define Coefficient of variation.

b) Explain Scatter diagram.

c) Write the properties of regression.

d) What do you mean by trend?

e) Find the standard deviation for data: 14, 22, 9, 15, 20, 17, 12, 11

f) Draw trend line by method of Semi Averages

Year	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Sale ('000)	210	200	215	205	220	235	210	235	225	245

g) From given two regression equations, find mean values of X and Y

$12X - 5Y + 99 = 0$; $60X - 27Y = 321$

h) What do you mean by Regression?

i) Explain the term 'Moving Average'.

j) What is 'Spurious' Correlation. Explain with example.

Q. 2 Attempt any two questions. Answer in 200 words. [2 x 5]

a) What is a "Time Series". Explain the components of time series.

b) Distinguish between:

(i) Positive and negative correlation

(ii) Linear and Non Linear correlation

(iii) Simple, Partial and Multiple correlations

c) Write the difference between Correlation and Regression Analysis.

and 12 and 12.7 respectively. What is the mode of the distribution?

(ii). Obtain the value of median from the following data:

391, 384, 591, 407, 672, 522, 777, 253, 2,488 1,490

c) Calculate the most suitable average for the following data:

Size of the item	Below 50	50-100	100-150	150-200	200 and above
Frequency	15	20	36	40	10

Q.4 Attempt any one. Answer in 600 words (Analytical Question / Case Study / Essay Type Question to test analytical and Comprehensive Skills) [10 x 1]

a) Calculate the Mean, Median and Mode for the data given below :

Daily earnings (Rs.)	No. of persons
50-53	3
53-56	8
56-59	14
59-62	30
62-65	36
65-68	28
68-71	16
71-74	10
74-77	5

b) (i) Display the following data in a pie chart . The total world wool production was distributed over various countries as follows in 2014: Country % of wool world production produced Australia 30% USSR 30% New Zealand 20% Argentina 10% SA 10% Others 10%

(ii)Represent the following data in a frequency polygon. The data gives the mass of apples from one tree.

Mass of Apple	Frequency
20-30	6
30-40	18
40-50	34
0-60	30
1-70	12

ANOOR- IV (2014 Course) [BE] - [S] MAY'17 - 2017

Subject : Statistics

Day : Wednesday
Date : 19/04/2017

35638

Time : 10.00 AM TO 01.00 PM
Max Marks : 100 Total Pages : 2

N.B.:
1) Attempt any **FOUR** questions from Section –I and any **TWO** questions from Section -II
2) Figures to the right indicate **FULL** marks.
3) Answers to both the sections should be written in **SAME** answer book.
4) Use of non programmable **CALCULATOR** is allowed.
5) Graphs should be drawn on **GRAPH PAPERS** only.

SECTION-I

Q.1 a) In a survey of 20 families, the following data regarding the number of children in a family was collected: (07)
1, 0, 1, 3, 4, 2, 2, 3, 4, 1, 0, 4, 2, 3, 3, 5, 3, 1, 3, 1
Represent the data as a discrete frequency distribution.

b) What do you mean by frequency distribution? Explain its types. (08)

Q.2 For the following data: (15)

X	35	25	29	31	27	24	33	36
Y	23	27	26	21	24	20	29	30

i) Obtain the two regression lines.
ii) Estimate the most likely value of X when Y = 20 and most likely value Y when X = 22.

Q.3 The median and the mode for the following data are 33.5 and 34 respectively. (15)
Find the missing frequencies.

Value	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
Frequency(f)	4	16	?	?	?	6	4	230

Q.4 Draw histogram, frequency polygon, ogive curves for the following distribution: (15)

Marks (less than)	10	20	30	40	50	60	70	80	90
No of students	4	6	24	46	67	86	96	99	100

Q.5 Find out the mean, median and mode from the following series: (15)

Size(above)	0	5	10	15	20	25	30
Frequency	38	37	35	32	21	11	02

P. T. O.

Q.6 a) The following table gives the age distribution of a group of 50 individuals: (07)

Age (in years)	16-20	21-25	26-30	31-35	36-40
No. of persons	10	15	17	8	5

Hence calculate range and the coefficient of range.

b) Find out mean deviation (about median) and its coefficient from the following series: (08)

Weight of package (kg)	4	6	8	10	12	14	16
Frequency	2	4	5	6	4	3	1

Q.7 Write short notes on the following: (15)

a) Limitations of Statistics.
b) Types of correlation.
c) Difference between absolute and relative measures of dispersion.

SECTION-II

Q.8 a) What is time series? Explain different components of time series in brief. (10)

b) Two ladies were asked to rank seven different types of lipsticks. The ranks given by them are: (10)

lipsticks	A	B	C	D	E	F	G
Priya	2	4	1	3	5	7	6
Nisha	1	3	2	5	4	7	6

Hence calculate Spearman's rank correlation coefficient.

Q.9 Compute Karl Pearson's correlation coefficient using the following data: (20)

Supply	152	158	169	182	160	166	182
Price	198	178	167	152	180	170	162

Hence interpret.

Q.10 The following data give the sales (in '000 Rs) of a company for the years 1985-1994. (20)

Years(t)	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
Sales(y)	50	82	65	86	70	52	90	65	87	43

Hence calculate:

i) Three yearly moving averages.
ii) Five yearly moving averages.
iii) Plot the original time series along with the 3 yearly and 5 yearly moving averages.

WINTER 2016

Subject : Statistics

32283

Time : 10.00 AM TO 01.00 PM
Max Marks : 100 Total Pages : 2

any **FOUR** questions from Section –I and any **TWO** questions from
to the right indicate **FULL** marks
to both the sections should be written in **SAME** answer book.
non programmable **CALCULATOR** is allowed.
should be drawn on **GRAPH PAPERS** only.

SECTION-I

istics and discuss its importance and limitations. (15)

ious methods that are used in the collection of primary data. (07)

frequency distribution table taking class intervals as: 20-24, 25-29, (08)
so on from the following data

20	55	39	48	46	36	54	42	30
42	32	40	34	31	35	37	52	44
45	37	33	51	55	52	46	43	47
26	52	48	25	34	37	33	36	27
36	41	33	23	39	28	44	45	38

istogram and frequency polygon for the following data: (07)

Obtained	0-10	10-20	20-30	30-40	40-50	50-60
Students	20	50	80	100	70	10

idend given by a software company from 2005 to 2011 is given below: (08)

Year	2005	2006	2007	2008	2009	2010	2011
idend (%)	20	30	32	50	60	50	40

ent the data using bar diagram.

late coefficient of variation from the following frequency distribution: (15)

rks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Students	05	15	30	50	35	15	10

late Spearman's coefficient of correlation between marks assigned to ten (07)
nts by judge X and judge Y in a certain competitive test as shown below:

No.	1	2	3	4	5	6	7	8	9	10
by Judge X	52	53	42	60	45	41	37	38	25	27
by Judge Y	65	68	43	38	77	48	35	30	25	50

P. T. O.

(08)

b) You are given the following data:

	X	Y
Mean	36	85
Standard Deviation	11	8

Correlation coefficient between X and Y is 0.66. Find:
i) Two regression equations
ii) Estimate value of X when Y = 75.

(15)

Q.6 Calculate the mean, median and mode for following data:

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	3	4	6	8	10	9	5

(15)

Q.7 Write short notes on the following:

a) Merits and demerits of median
b) Graphs and diagrams
c) Uses of Time Series

SECTION-II

Q.8 a) What do you understand by central tendency? Explain the characteristics of good measures of central tendency. (10)

b) The following table shows the distribution of 100 families according to their expenditure per week. The mode of the distribution is Rs.23. Calculate the missing frequencies and the arithmetic mean. (10)

Expenditure per week (in rupees)	0-10	10-20	20-30	30-40	40-50
No. of families	14	x	27	y	15

Q.9 a) Define Time Series Analysis and explain the components of Time Series. (10)

b) Calculate 3- years moving averages of the production figures given below and draw the trend: (10)

Year	Production (in tonnes)	Year	Production (in tonnes)
1995	15	2003	63
1996	21	2004	70
1997	30	2005	74
1998	36	2006	82
1999	42	2007	90
2000	46	2008	95
2001	50	2009	100
2002	56	2010	102

Q.10 Compute Karl Pearson's correlation coefficient between height (X) and weight (Y) from the following data. Also obtain the two regression lines: (20)

Height (X) :	61	65	68	62	60	63	64	69	70	66
Weight (Y) :	62	67	70	64	63	68	70	76	75	70