UG EC S4-01
(DSC)

FOUR YEAR BACHELOR'S DEGREE COURSE

(CBCS in Accordance with NEP 2020)

SEMESTER - IV

QUANTITATIVE METHODS FOR ECONOMIC ANALYSIS - II

ECONOMICS

Surya Kumar Bhuyan School of Social Sciences

KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

Professor Madhurjya Prasad Bezbaruah, Dept. of Economics, Gauhati University

Professor Ratul Mahanta, Dept. of Economics, Gauhati University

Professor Rajashree Bedamatta, Indian Institute of Technology, Guwahati

Course Co-ordinators: Bhaskar Sarmah, Ph.D., KKHSOU

SLM Preparation Team

Units	Contributors
1 & 2	**Dr. Diwakar Singh**, Christ College, Bhopal
3	**Dr. Padmalochan Hazarika**, Gauhati University
4	**Ajanta Majumdar**, Gauhati Commerce College
5, 6	**Professor Joydeep Baruah,** Krishna Kanta Handiqui State Open University
	Dr. Tarakeswar Chaudhury, Retired Professor, Cotton College
8	**Harekrishna Deka**, KKHSOU
9, 10, 11	**Dr. Rijusmita Sarma**, HoD, Department of Statistics, L. C. B. College, Guwahati
12	**Dr. Rupam Barman**, Tezpur University
13	**Ajanta Majumdar**, Gauhati Commerce College
14	**Dr. Tarakeswar Chaudhury**, Retired Professor, Cotton College

Editorial Team

Content : **Dr. Sanjay Dutta**, PETC, Gauhati University

Structure, Format & Graphics : **Shri Harekrishna Deka**, KKHSOU

Bhaskar Sarmah, Ph.D. KKHSOU

April, 2025

This Self Learning Material (SLM) of the Krishna Kanta Handiqui State University is made available under a Creative Commons Attribution-Non Commercial-ShareAlike4.0 License (International): http.//creativecommons.org/licenses/by-nc-sa/4.0.

Printed and published by Registrar on behalf of the Krishna Kanta Handiqui State Open University.

Head Office : Patgaon, Rani Gate, Guwahati-781 017;

Sub-Head Office : Resham Nagar, Khanapara, Guwahati-22; Web: www.kkhsou.in

The University acknowledges with thanks the financial support provided by the **Distance Education Bureau, UGC**, for the preparation of this study material.

BACHELOR OF ARTS
Quantitative Methods for Economic Analysis - II

CONTENTS

UNIT 1: Introduction to Statistics **7-25**

Meaning and Importance of Statistics; Main Features of Statistics; Scope of Statistics; Application of Statistics in Business & Economics; Meaning and Types of Data; Methods of Presenting Statistical Information; Individual, Discrete and Grouped Series; Tabulation of Data: Meaning, Importance and Objectives; Parts of Table

UNIT 2: Visual Representation of Data **26-43**

Idea of Distribution and its Need; Frequency Distribution and its Classification; Method of forming Discrete and Grouped Frequency Distribution; Graphical Representation of Data; Histogram; Frequency Polygon; Ogive or Cumutative Frequency Curve; Cartogram or Mapograph; Pictorial Representation of Data

UNIT 3: Measures of Central Tendency **44-69**

Measures of Central Tendency: Definition, Characterstics of a Good Average, 'Σ' Symbol, Different Types of Measures of Central Tendency

UNIT 4: Measures of Dispersion **70-93**

Dispersion: Objectives of Studying Dispersion,Essentials of a Good Measure of Dispersion, Different Measures of Dispersion

UNIT 5: Skewness, Moments and Kurtosis **94-117**

Skewness; Measures of Skewness: Karl Pearson's Co-efficient of Skewness, Boweley's Co-efficient of Skewness; Moments: Moments About Mean, Moments About Arbitrary Point A, Relation Between Central and Raw Moments; Karl Pearson's Beta (β) and Gamma (γ) Coefficient: Coefficient of Skewness Based on Moments; Kurtosis: Measures of Kurtosis

UNIT 6: Correlation **118-131**

Correlation: Meaning and Concept; Types of Correlation; Methods of Measuring Correlation; Scatter Diagram and Correlation; Correlation Coefficient; Underlying Assumptions; Interpretation of Correlation Coefficient; Rank Correlation Coefficient

UNIT 7: Regression **132-140**

Meaning and Concept of Regression; Linear Regression; Line of Regression and Regression Equation; Coefficient of Regression; Relation between Correlation and Regression Coefficients; Correlation and Regression Analysis: A Comparison

UNIT 8: Fundamentals of Probability **141-165**

Random Experiment; Classical Definition of Probability; Set: Subset, Sample Space; Axiomatic Definition of Probability; Elementary Theorems on Probability

UNIT 9: Conditional Probability **166-186**
Conditional Probability; Multiplication Theorem on Probability; Total Probability; Baye's Theorem

UNIT 10: Random Variables and its Probability Distribution **187-199**
Random Variable: Definition or Random Variable, Discrete Random Variable, Continuous Random Variable; Probability Distribution: Probability Mass Function, Probability Density Function; Mathematical Expectation and Varience: Mathematical Expectation, Properties of Mathematical Expectation, Variance of Random Variable, Properties of Variances, Illustrated Examples

UNIT 11: Theoretical Probability Distribution (Discrete Variable) **200-225**
Moment Generating Function: Definition, Generation of Moments, Properties of Moment Generating Function, Ilustrated Examples; Binomial Distribution: Derivation of Binomial Distribution, Definition, Moments of Binomial Distribution, Moment Generating Function of Binomial Distribution, Fitting of Binomial Distribution, Properties of Binomial Distribution, Illustrated Examples; Poisson Process: Poisson Distribution as a Limiting Case of Binomial Distribution, Moments of Poisson Distribution, Moment Generating Function, Fitting of Poisson Distribution, Properties of Poisson Distribution, Application of Poisson Distribution, Illustrative Examples

UNIT 12: Theoretical Probability Distribution (Continuous Variable) **226-238**
Continuous Probability Distribution; Normal Distribution: Properties of Normal Distribution, Standard Normal Variate

UNIT 13: Index Numbers **239-264**
Uses of Index Numbers; Types of Simple Index Numbers; Properties of Relatives; Construction of Simple Index Numbers; Construction of Weighted Index Numbers; Test of Adequacy of Index Number

UNIT 14: Time Series **265-282**
Definition of Time Series; Importance of Time Series Analysis; Components of a Time Series; Methods of Measuring Secular Trend; Estimation of the Trend by the Method of Moving Average

UNIT 15: Basic Concepts in Sampling **283-304**
Definition of Population, Parameter and Sampling: Random and Non-random Sampling Methods; Sampling with and without Replacement, Large Samples; Errors in Sampling Progress

COURSE INTRODUCTION

The course *'Quantitative Methods for Economic Analysis-II'*. introduces us to the subject *Statistics*. Earlier in the previous semester, we had dealt with Mathematics, another subject which is widely used in different fields in Economics. Both the courses of Mathematics and Statistics are widely used in Economic analysis, particularly in research. And as we have already mentioned, all these three disciplines, viz., Statistics, Mathematics and Economics has resulted in the emergence of an advance branch in Economics known as Econometrics. Thus, a thorough understanding of the concepts of Mathematics and Statistics will be extremely helpful in the higher studies in Economics.

BLOCK INTRODUCTION

The Course '*Quantitative Methods for Economic Analysis – II'* has been divided into two blocks. This is the first block of the course. This block contains the first seven units of the course. The first unit deals with the discussion of a few basic concepts of the subject Statistics. The second unit will introduce you to the concept of central tendency and its different measures. After going through this unit, you will be able to define the term 'average' from the perspective of Statistics and know its different types. The next unit deliberates on the concept of dispersion. In this unit, the meaning of dispersion and its different measurements have been discussed. The fourth unit discusses the concept of dispersion. After going through this unit, you will acquire a fair knowledge about its meaning, its different types and measures. The next unit deals with the concept of skewness, moments and kurtosis. Unit sixth deliberates on the concept of correlation. After going through this unit, you will be able to define the term correlation, plot scatter diagrams and discuss its various types. The concept of correlation coefficient has also been discussed. The next unit, which is also the last unit of the block is on regression. After going through this unit, you will be able to draw the lines of regression and also differentiate between correlation and regression.

The next block deliberates on the remaining eight units of the course. The second block starts with the unit the concept of conditional probability and concludes with the discussion of a few basic concepts in sampling. Overall, the fifteen units of the course will give you a fair ground on the subject Statistics. A thorough understanding of the concepts would definitely help you in your higher studies in Economics.

While going through a unit, you will notice some along-sode boxes, which have been included to help you know some of the difficult, unseen terms. Some 'ACTIVITY' (s) has been included to help you apply your own thoughts. Again, we have included some relevant concepts in 'LET US KNOW' along with the text. At the end of sections, you will get 'CHECK YOUR PROGRESS' questions. These have been designed to self-check your progress of study. It will be better if you solve the problems put in these boxes immediately after you go through the sections of the units and then match your answers with 'ANSWERS TO CHECK YOUR PROGRESS' given at the end of each unit.

UNIT 1: INTRODUCTION TO STATISTICS

UNIT STRUCURE

1.1 Learning Objectives

1.2 Introduction

1.3 Meaning and Definition of Statistics

1.4 Main Features of Statistics

1.5 Scope of Statistics

1.6 Application of Statistics in Business & Economics

1.7 Meaning and Types of Data

1.8 Methods of Presenting Statistical Information

1.9 Individual, Discrete and Grouped Series

1.10 Tabulation of Data: Meaning, Importance and Objectives

1.11 Parts of Table

1.12 Let Us Sum Up

1.13 Further Reading

1.14 Answers to Check Your Progress

1.15 Model Questions

1.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- put forward the meaning and definition of statistics
- disucss the main features of statistics
- explain the scope of statistics
- distinguish between statistical data and statistical methods
- outline the applications of statistics in business & economics
- describe the different methods of presenting statistical information
- explain how tabulation of data is done
- discuss the different parts of a table.

1.2 INTRODUCTION

Dear learners, you must have come across various facts in your daily life involving numbers. Such as Temperature, Price of the commodity,

Quantity of items, Share market etc. These facts in statistical terms are called as data. The question arises, what to do with these facts or what type of information these facts carry? The answer to the question is – it depends on person to person that what type of information they need.

Around the world, specially the business world, most of the decisions are made with the information gathered about the companies, market place, the economic and the financial environment. Such information usually comes in the form of data or accompanied by data. Statistics provides the tool through which such data are collected, analyzed, summarized, and presented to facilitate the decision making process. This unit shall introduce to you the field of Statistics.

1.3 MEANING AND DEFINITION OF STATISTICS

The word 'Statistics' has been derived either from Latin word Status or Italian word Statistica or German word Statistick. Each carry same meaning 'An Organized Political State'. Some people consider Statistics as Data, Facts or Measurements, simply saying numerical facts while for some Statistics is analysis of figures for prediction and for some it is presentation of facts in graphs.

In this way Statistics exists in the following three forms:

a) As Numerical Facts: In this form it includes numerical facts such as National Income, Export, Import, Production, Employment Statistics etc.

b) As a subject: It exists in Singular as well as plural form.



It is used as a subject in singular form like Economics, Physics Etc. while in plural form it is the subject matter. In other words in singular sense Statistics refers to the whole body of analytical tools that are used to collect the figures, organize and interpret them and finally

to draw conclusions. In plural sense, it refers to a set of figures such as production, sale etc.

c) It may also refer to the study of, and research into the theory and principles underlying statistical methods.

According to Tate the three meanings of the word statistics are contained in the statements "You compute Statistics by Statistics from Statistics."

Definition of Statistics: According to Spiegal "Statistics is concerned with scientific method of collecting, organizing, summarizing, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis".

Secrist highlights few major characteristics of Statistics as follows:

- Statistics are aggregates of facts. This means that a single figure is not statistics.
- Statistics are affected by a number of factors. For example, sale of product depends on number of factors such as price, demand, quality etc.
- Statistics must be reasonably accurate or in other words conclusions must be based on accurate figures.
- Statistics must be collected in a systematic manner because facts collected in haphazard manner will not be reliable and will lead to misleading conclusion.
- Finally, Statistics should be placed in relation to each other. Logical collection of data helps in avoiding confusion.

A. L. Bowley defined Statistics as: "Science of Counting, Science of averages and Science of the measurement of social organism regarded as a whole".

According to P.H. Karmel "The subject Statistics is concerned with the collection, presentation, description and analysis of data which are measurable in numerical form".

1.4 MAIN FEATURES OF STATISTICS

After going through the above definitions, you must have gathered some basic idea of statistics. Now lets discuss some of its major features:

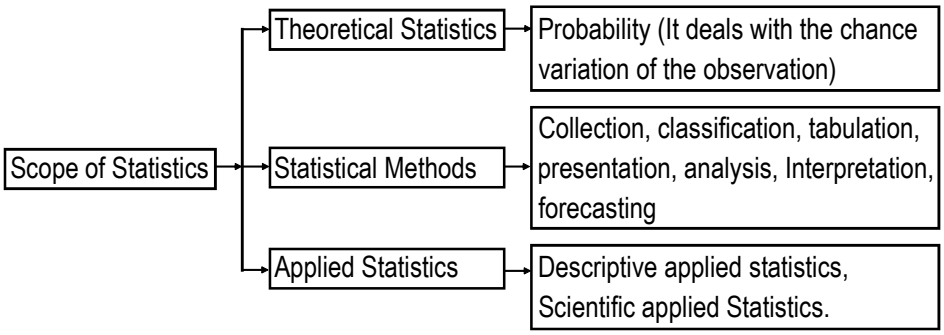
- Statistics are aggregate of facts: A single numerical record or fact is not statistics; in fact it is the aggregate of facts. For example, salary of person for a particular month, marks of a student in Economics in some exam etc.
- Statistics are expressed numerically: Statistical facts are expressed in the form of numbers. In other words it is quantitative in nature. For example, if we say that the performance of BA students in Economics this year is Average, then it is not statistics. The qualitative characteristics like honesty, beauty, intelligence etc. expressed by giving certain scores as numerical standards can be called as statistics.
- Statistics are affected to a marked extent by multiplicity of causes: Statistical data are influenced by a number of factors. One cause alone can not affect data to a marked extent. For example, facts related to agriculture production are affected by a number of factors like rain, quality of soil, quality of seed, methods etc.
- Statistics are collected by enumerations or estimation according to reasonable standards of accuracy: Statistical data are enumerated or estimated according to reasonable standard of accuracy. The reasonable standard of accuracy depends on the purpose which the statistics are to serve.
- Statistics are collected in a systematic manner: Statistics are collected in a systematic manner according to a well prepared plan. The purpose for which the data are to be collected is decided and well defined in advance.
- Statistics are being capable of being placed in relation to each other: Statistical data are collected mostly for the purpose of comparison. The data must be homogenous and uniform in nature. For example, if we want to compare standard of living of some class in India then it must be compared with the class of the same nature not with the other class or foreign country data.
- Statistics tests hypothesis and laws of other subjects: Nearly every law and theory, frequently used today, were based on deductive

method. These laws can be tested for their validity with the help of statistics.

- Statistics estimates for the present and predicts for the future: Statistical methods are not only helpful in estimating the present but also in forecasting the future. Extrapolation and interpolation are the techniques uses for the purpose. 5 year plans and other economic plans like budget etc are the examples under this heading.

1.5 SCOPE OF STATISTICS

Statistics is a tool of all science indispensable to research and intelligent judgment and has become a recognized discipline in its own right. Moreover, the scope of applied statistics is increasing day by day. The scope of statistics stretches over all those branches of human knowledge in which a group of the significance of large numbers is looked for. The scope of statistics is wide, still it can be summarized as:



Theoretical Statistics: It is the basis of the science of statistics, also called as Theory of Probability which deals with the chance variation of the observations.

Statistical Methods: In the words of Johnson and Jackson "Statistical methods are the procedures used in collection, organization, summary, analysis, interpretation and presentation of data." Simply speaking they are the devices by which complex and numerical data are analyzed.

Table 1.1: Difference between Statistical Data and Statistical Methods

S.No.	Statistical Data	Statistical Methods
1	They are available in raw form called raw data.	It helps to process the raw data.
2	They are facts expressed numerically.	It is a process.
3	They are descriptive in nature.	It is a tool of analysis.
4	They are meant for processing and processed data help in decision making.	It helps in analyzing the processed data.
5	The selection of statistical methods depends on the utility of data.	It is necessary to have a knowledge of statistical methods while collecting data.
6	They are senseless without application of statistical methods.	It remains idle for want of statistical data.

1.6 APPLICATION OF STATISTICS IN BUSINESS AND ECONOMICS

In the modern era statistics serves the base for the Mathematical approach to Economics. Statistical techniques have proved immensely useful in the solution of variety of Economics problems including trend analysis, production, consumption, GDP, income etc. cost of living index, price index, wholesale price index numbers, demand curves, forecasting techniques etc. very powerful tools in statistics used in analyzing Economics facts.

According to Marshall, “Statistics are the straw out of which, I like every other economist, have to make bricks.” According to Jevons, “I know not when we shall have a perfect system of statistics but the want of it is the only insuperable obstacle in the way of making Economics an exact Science.” According to Oskar Lange, “Economics is the science which deals with the determination by statistical methods of concrete quantitative laws occurring in economic life.” According to Tippet, “It may one day happen that Economic Department, at universities, instead of being dominated by the theorists will come under the domination of the statistical laboratory just as Physics and Chemistry Departments are dominated by the experimental laboratory.”

Today, business world is changing at rapid rate. Commerce, industry and other allied fields are also expanding. Due to industrial revolution, business activities have expanded and at the same time competition in the

market has also increased. Industries have crossed the local and national market to international market. Thus, in this situation statistics helps in handling the activities like sales, purchase, production, marketing, control, finance, promotion etc. For this purpose role of statistical data and statistical tools like probability, expectation, sampling procedures, significance tests, theory of estimation and techniques of projection and forecasting and many others are of paramount importance. Business forecasting techniques are of vital use for getting estimates which serves as an indicator to future economic events. The study of time series, index numbers, business cycles, which are supposed to be Economic Barometers and are prepared by accurate statistical data, is of great importance for businessman. Boddington rightly observed that, "The successful businessman is the one whose estimates most closely approaches accuracy."

In decision making, statistical tables and charts are frequently used by managers to present numerical facts. Businessman and industrialists rely so much on statistical methods in managerial decisions.

Statistical methods also deal with researches on market and production. Market research and market surveys by sampling methods are now extremely useful for any business. Price and demand of similar products available in the present market, tastes of consumers, seasonal variations, etc, are analyzed statistically.

QC i.e. quality control is also gaining momentum in business world. Statistics has already made significant change in maintaining and improving quality through quality control techniques.

According to M. M. Blair, "If all price statistics were removed from all papers, magazines, radio and telegraphic reports for a single day, the business world would be paralysed. If all statistics now available were removed from the world for one year, utter economic chaos and ruin would result."

Thus, it can be concluded that applications of statistical methods in commerce, business and industry has become a condition of progress. Industry and business world involve risk and when it is risk, it is better to have a calculated risk.

1.7 MEANING AND TYPES OF DATA

Collection of data is the backbone of statistical investigation. Data is aggregate of facts and these facts are collected in a systematic manner for a definite and pre planned purpose.

Data can be categorized on the following bases:

1) On the basis of organization of data.
2) On the basis of sources of data.
3) On the basis of collection of data.
4) On the basis of characteristics of data.

On the basis of organization: It can be divided into two:

- **Raw data:** The data originally collected or original facts collected through questionnaire, schedules etc. are called raw data. In its original form it is also called as ungrouped data.
- **Processed data:** The raw data or ungrouped data passed through at least one statistical method, organization, classification, tabulation, computations etc. are called processed data.

On the basis of Sources: On the basis of sources it can be divided into two:

- **Internal Data:** The data collected from internal source for any specific purpose and after preparing a plan are called internal data. For example, if an economic condition of persons living in some city is to be studied, the data obtained from each person is called as internal data.
- **External data:** The data obtained from external sources are known as external data.

On the basis of Collection of data: There are two types of data based on its collection:

- **Primary data:** The data collected by the researcher itself is called as primary data. According to Horace Secrist, “By primary data are mean those data which are original, that is, those in which little or no grouping has been made, the instance being recorded or itemized as encountered. They are essentially raw materials.”

- **Secondary data:** Data which have already been collected by some other investigator/person/organization for their purpose and the present investigator uses them are called secondary data. According to M. M. Blair, "Secondary data are those already in existence and which have been collected for some other purpose than answering of the question at hand." For example, if a person uses the data regarding foreign import-export published by government for studying payment-balance, then these data are secondary data.

Note: Primary data is more reliable than secondary data because it is collected by the investigator himself/herself.

Table 1.2: Comparison of Primary and Secondary data

S.No.	Basis	Primary Data	Secondary Data
1	Originality	Primary data are original i.e. collected First time by the investigator itself.	Secondary data are not original, i.e. they are already in existence and are used by investigator.
2	Organization	Primary data are like raw material.	Secondary data are in the form of finished product. They have passed through statistical methods.
3	Purpose	Primary data are according to the object of investigation and are used without correction.	Secondary data are collected for some other purpose and are corrected before use.
4	Expenditure	The collection of primary data requires large money, energy and time.	Secondary data are easily available from secondary sources. (Published and Unpublished)
5	Precautions	Precautions are not necessary in the use of primary data.	Precautions are necessary in the use of secondary data.

On the basis of characteristics: On the basis of characteristics, data are of two types:

- **Qualitative data:** The characteristics which cannot be measured but only their presence or absence in a group of individuals can be noted are called qualitative data or attribute. For example– Poverty, Unemployment, honesty etc.

  Qualitative data are of two types– Nominal data and ranked data. Categories based on gender (male, female), residential area (rural,

urban) is examples of nominal data whereas the classification according to their height, short, average etc. is ranked data.

- **Quantitative data:** The characteristics which can be measured directly are called quantitative data. The measurable characteristics are also called as variable. For Height, weight, sales, production etc.. Quantitative data or variables are again classified into two: (i) Discrete Variable & (ii) Continuous variable

  Discrete variable: A variable which takes only finite distinct values are known as discrete variable. For example– Number of employees, number of industries etc. It is expressed in whole numbers.

  Continuous variable: A variable which can theoretically assume all values within a certain interval or intervals is called continuous variable. For example– Age, Income of employees, height of persons etc. It is expressed in decimals.

1.8 METHODS OF PRESENTING STATISTICAL INFORMATION

The data collected initially from a given source, whether they are experiments, surveys, or observation, are presented with the help of the following methods:

1) Textual method: In this, the reader acquires information by reading the data simply.
2) Tabular method: Provides a more precise, systematic and orderly presentation of data in rows and columns. (We will be discussing it in the later part of the unit).
3) Semi-Tabular Method: It uses both textual and tabular methods.
4) Graphical Method: The utilization of graphs is most effective method of visually presenting statistical results or findings. (We will be discussing it in the later part of the unit).

1.9 INDIVIDUAL, DISCRETE AND GROUPED SERIES

The process of arranging the data in a logical or systematic order is called as seriation and the data so arranged is called a statistical series. In

other words when attributes or things are counted, measured or weighed and they are placed one after another in some orderly manner, they constitute a series.

According to Horace Secrist, "A series as used statistically may be defined as things or attributes of things arranged according to some logical order."

On the basis of Shape or Construction of data, statistical series are of three types:

(a) Individual, (b) Discrete & (c) Grouped

- **Individual series:** A series of individual observations where the items are listed singly is known as an individual series. For example sales of mobile of famous brand (say A) on Monday, Tuesday and Wednesday:

Day:	Monday	Tuesday	Wednesday
No. of mobiles sold:	58	62	45

An individual series arranged in an ascending or descending order of magnitude is called as array. The process of arranging the data in ascending or descending order is known as arraying of data.

- **Discrete series:** When the distinct values of a discrete variable or a continuous variable are presented in an ascending order or descending order along with their frequencies, the series is called discrete series. For example Marks obtained by 38 students of B.A. in Economics.

Marks obtained:	20	22	23	27	30	34	38	41
No. of students:	4	3	5	8	6	5	4	3

- **Grouped Series:** When the observations or measurements are arranged in ascending or descending order in groups along with their frequencies, expressed in groups called as class intervals, is called as Grouped series. If the continuity of the group is not broken the grouped series is called as exclusive series while if the continuity is broken and the new interval starts with the next number, then it is called as inclusive series. Lets understand it with the help of examples:

Exclusive series:

Marks obtained:	0-20	20-40	40-60	60-80	80-100
No. of Students:	4	3	8	3	2

Here the continuity is maintained but when we consider a particular interval, say 20-40 then it will contain the data from 20 to less than 40 or excluding 40. Thus, it is called as exclusive series. Here the class interval is 20.

While in case of inclusive series, for example:

Marks obtained:	10-19	20-29	30-39	40-49	50-59	60-69
No. of students:	6	4	6	9	4	2

Here the both the limits of the intervals are considered, i.e. the upper as well as lower limit. Thus, it is called as **inclusive series**. While arranging them as **exclusive series** we have to subtract 0.5 from the lower limit and add 0.5 to the corresponding upper limit.

10-19 can be converted to exclusive as 9.5-19.5.

Here 0.5 is subtracted from the lower limit and 0.5 is added to upper limit to make the inclusive series an exclusive one.

Table 1.3: Comparison of Exclusive and Inclusive Series

S.No.	Exclusive Series	Inclusive Series
1	Upper limit of a class is included in the next Higher class.	Both limits are included in that class.
2	Upper limit of a class is equal to the lower limit of the next class.	Upper limit of a class is not equal to the lower limit of the next higher class.
3	This method is suitable in every situation.	This method is suitable in case of integers.
4	For further studies it is not necessary to change exclusive form to inclusive form.	Sometimes it is essential to change inclusive form to exclusive form.

1.10 TABULATION OF DATA: MEANING, IMPORTANCE AND OBJECTIVES

In broad sense, “tabulation is an orderly arrangement of data in columns and row.” In fact it is process between the collections of data on the one hand and its final analysis on the other.

According to Croxton and Cowden, "Either for one's own use or for the use of others, the data must be presented in a suitable form." This clearly shows the importance of systematic presentations of data.

According to Tuttle "A statistical table is the logical listing of related quantitative in vertical columns and horizontal rows of number with sufficient explanatory and qualifying words, phrases and statements in the form of title, headings and noted to make clear the full meaning of data and their origin." Thus, a statistical table is a presentation of numbers in a logical arrangement with some brief explanation to show what they are, i.e. in rows and columns.

In a similar way, tabulation is a scientific process involving the presentation of classified data in an orderly manner so as to bring out their essential features and chief characteristics.

Objectives: Its objectives can be understand in these words, "Tabulation is one of the most important and ingenious device of presenting the data in a condensed and readily comprehensible form and attempts to furnish the maximum information contained in the data in the minimum possible space without sacrificing the quality and usefulness of data.''

The major objectives of the tabulation are as follows:

- **Classification of the object of investigation:** Owing to presentation of data in a tabular form, the objects of investigation becomes clear and easily comprehensible. According to Bowley, "the function of tabulation in the general scheme or statistical investigation is to arrange in easily accessible form, the answer with which the investigation is concerned."
- **Clarification of Characteristics of data:** The table, after removing the vague and irrelevant information, presents clearly the major characteristics of data.
- **Presentation of facts in minimum space:** A huge amount of data can be presented in relevant way in minimum possible space, which communicates the information in far superior way than textual material.
- **Facility of Statistical Processing:** Because of categorical classification of data presented in tabular form, statistical processing-comparability, analysis etc. becomes easier.



Tabulation: Presentation of data collected in an orderly and summarized form.

Discrete series: A series of data where the data are distinct numbers with gaps between them.

Frequency distribution: The arrangement and display of data wherein the observed value is paired with its frequency.

Importance/advantages of Tabulation: Tabulation is advantageous in many ways:

- **Simplification of facts:** The data presented in a tabular form are systematic and shows a definite pattern. This helps in understanding the trends and tendencies emerging from the information obtained under investigation in similar way. In other words, by tabulation data become simpler and easily accessible.
- **Comparability:** Tabulation makes comparison easier, and figures are presented in classified way in distinct rows and columns.
- **Classification:** It becomes easier to classify the data for the presentation in tabular form. Unless the classified data are presented in different rows and columns, the distinct characteristics of the data cannot be known.
- **Economical:** As a tabulated data requires less space and time for its presentation.
- **Referencing:** Tabulated data serves as good reference.
- **Interpretation and inference:** The data, well presented with clear caption without any ambiguity, can help in interpreting the inferring in better way.
- **Computation:** Statistical measures like mean, median, mode, dispersion etc. can easily be computed by the help of a tabled data rather than textual data. It is not possible to calculate any such statistical measures and to draw graphs in the absence of data presented in tabular form.

Disadvantages/limitations of tabulation:

- A table does not possess the description of the data presented there in. Hence, the persons who are not accustomed of assimilating facts from table are not able to understand it easily.
- Specialized knowledge is required to understand a table.

Requisites of a good table: A good table must possess the following characteristics:

- **Appropriate size:** The table should be neither too big nor too small. It should be attractive and beautiful.

- **Meaningful information:** The table should not contain too much of information. It is always better to have several manageable size small tables instead of a big table. (Too much of details should be avoided).
- **Comparable and Accessible:** Since statistical table provides the basis for analysis and computation of various statistical parameters/ measures like mean, S.D., correlation etc., it should be free from all sorts of errors. Besides, wherever necessary, these statistical measures- averages, percentages etc. should be given in the table itself to facilitate comparison.
- **As per the objective:** The table should be according to the objective of statistical investigation.
- **Scientifically Prepared:** While preparing a table, all the rules of tabulation should be followed.
- **Clarity:** A table should be prepared that even layman can draw inferences from it.

According to Harry Jarome "A good statistical table is triumph ingenuity and technique, a masterpiece of economy of space combined with a maximum of clearly presented information."

1.11 PARTS OF TABLE

Raw data collected are transformed into suitable table and to have proper tabulation, it is necessary to know the main parts of it. These are as follows:

- **Table number:** Every table should be numbered so that it can be located/ identified. The number is normally indicated at the left top of the table**.**
- **Title:** Each table must bear a title indicating the type of data contained. The title should not be very lengthy so as to run in several lines. It should be clear and unambiguous.
- **Captions and Stubs:** A table consists of rows and columns. The headings or sub-headings given in columns are known as captions

while those given in rows are stubs. It is necessary that a table should have captions and stubs to indicate what columns and rows stand for. It is also desirable to provide for an extra column and row in the table for the column and row totals.

- **Main body of the table:** As this part of the table contains data, it is the most important part of it. Its size and shape should be suitable to accommodate the data. The data are entered from the top to the bottom in columns and from left to right in rows.
- **Ruling and Spacing:** A good table should invariably have proper ruling and spacing. Vertical lines are drawn to separate columns but horizontal lines are normally not drawn except in case of totals which must be separated from the main body of horizontal lines. In addition, the horizontal lines are drawn both at the top just below the title of the table and also at the bottom. Due care must be exercised in providing proper space in accordance with the dimensions or magnitude of figures.
- **Head note:** A head note is invariably given just below the title of a table indicating the unit of measurement applicable to the data displayed. This is shown generally in bracket.
- **Footnote:** Sometimes it becomes necessary to give some explanation for the data used or to explain the meaning of the abbreviation used. This should be given in the form of a footnote at the bottom of the table just below the last horizontal line.
- **Source note:** Whenever secondary data are used, it is necessary to show the source from which such data are taken. This is also in the form of footnote. Its purpose is to facilitate the reader to refer to that source if required.

Sample of the table:

Table number _____

Title _____

Stub Heading	**Captions**		**Captions**		**Total**
	Caption	Caption	Caption	Caption	
Sub Entries		Main Body			
Total					
Footnote:					
Source:					



CHECK YOUR PROGRESS

Q.1: What is meant by tabulation of data? (Answer in about 40 words)

.....

.....

.....

.....

Q.2: Mention any two distinctions between inclusive series and exclusive series. (Answer in about 60 words).

.....

.....

.....

.....

.....

.....

.....



1.12 LET US SUM UP

- The word 'Statistics' has been derived either from Latin word Status of Italian word Statistica or German word Statistick. Each carry same meaning 'An Organized Political State'.
- Statistics is a tool of all science indispensable to research and intelligent judgment and has become a recognized discipline in its own right.
- In the modern era statistics serves the base for the Mathematical approach to Economics.
- Collection of data is the backbone of statistical investigation. Data is aggregate of facts and these facts are collected in a systematic manner for a definite and pre planned purpose.
- Qualitative data are of two types– Nominal data and ranked data.
- The process of arranging the data in a logical or systematic order is called as seriation and the data so arranged is called a statistical series.
- In broad sense, "tabulation is an orderly arrangement of data in columns and row."



1.13 FURTHER READING

1) Beri, G. C. (2010). *Business Statistics* (3rd Ed.). New Delhi: Tata McGraw Hill Education Private Limited.

2) Shukla, S. M. & Sahai, S. P. (2008). *Business Statistics Theory & Practice*. Agra: Sahitya Bhawan Publications.



1.14 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: Tabulation of data means an orderly arrangement of data in columns and row. In fact, it is the process between the collections of data on the one hand and its final analysis on the other.

Ans. to Q. No. 2: Two important distinctions between inclusive series and exclusive series are: (a) in inclusive series, both limits are included in that class, while in exclusive series upper limit of a class is included in the next higher class, (b) in inclusive series, the upper limit of a class is not equal to the lower limit of the next higher class, while in exclusive series the upper limit of a class is equal to the lower limit of the next class.



1.15 MODEL QUESTIONS

A) Short questions: (Answer each question in about 150 words)

Q.1: Define the term Statistics.

Q.2: Discuss the scope of Statistics.

Q.3: Differentiate between statistical data and statistical methods.

Q.4: Discuss the different parts of a table.

B) Long questions: (Answer each question in about 300-500 words)

Q.1: Discuss the major features of Statistics.

Q.2: Discuss the process of tabulation of data.

*** ***** ***

UNIT 2: VISUAL REPRESENTATION OF DATA

UNIT STRUCTURE

2.1 Learning Objectives

2.2 Introduction

2.3 Idea of Distribution and its Need

2.4 Frequency Distribution and its Classification

2.5 Method of forming Discrete and Grouped Frequency Distribution

2.6 Graphical Representation of Data

2.7 Histogram

2.8 Frequency Polygon

2.9 Ogive or Cumutative Frequency Curve

2.10 Cartogram or Mapograph

2.11 Pictorial Representation of Data

2.12 Let Us Sum Up

2.13 Further Reading

2.14 Answer to Check Your Progress

2.15 Model Questions

2.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- discus the idea of distribution and its need
- acquire knowledge about frequency distribution and its classification
- describe the different methods of forming discrete and grouped frequency distribution
- use graphical representation of data, viz., histogram, frequency polygon, ogive and cartogram or mapograph.

2.2 INTRODUCTION

Dear learners, in this unit, we are going to discuss the distribution of data, its method and graphical representation of facts. Thus, after going through this unit, you would know the meaning of frequency distribution and

outline its need. Apart from this, the general guidelines for the use of graphs and diagrams have also been discussed in this unit. Thus, this unit will be helpful for you to be able to draw different statistical graphs like– Histogram, Frequency polygon, ogive, Cartogram and Pictograph.

2.3 IDEA OF DISTRIBUTION AND ITS NEED

A tabular presentation of the data in which the frequencies of values of a variable are given along with them is called a frequency distribution.

According to Morris Humburg, "A frequency distribution of frequency table is simply a table in which data grouped into classes and number of cases which fall in each class are recorded. The numbers in each class are referred to as frequencies."

2.4 FREQUENCY DISTRIBUTION AND ITS CLASSIFICATION

The classification according to variables can be performed in two forms:

1) Discrete Series or Discrete frequency distribution.
2) Grouped Series or grouped frequency distribution.

Frequency: The number of times each value of a variable occurs is known as its frequency.

Frequency distribution: Frequency distribution is classified (tabular form) according to the number (frequency) possessing the same values of the variable. When a frequency distribution is presented in tabular form, it is also called as frequency distribution table.

Frequency Distribution can be categorized as follows:

Univariate Frequency Distribution: A frequency distribution which shows the frequency of occurrence of different values of a single variable is called as univariate frequency distribution.

For example, marks scored by students of Class X in Economics in final exam.

Bivariate Frequency Distribution: A frequency distribution based on two variables is called as bivariate frequency distribution.

For example marks scored by students of Class X in Economic and Statistics in Final Exam.

Discrete Frequency Distribution: A frequency distribution which is formed by distinct values of a discrete variable or a continuous variable is called as discrete frequency distribution.

Grouped Frequency Distribution: A frequency distribution which is obtained by dividing the entire range of the given observations on a discrete variable or continuous variable into groups and distributing the frequencies over these groups is called a grouped frequency distribution.

Dear students for understanding lets discuss the method of forming discrete and grouped frequency distribution in the next section.

2.5 METHOD OF FORMING DISCRETE AND GROUPED FREQUENCY DISTRIBUTION

Steps for forming discrete frequency distribution

i) Make a blank table consisting of three columns with the headings– variable, tally marks and frequency.

ii) Find the smallest observation and the largest observations; write the observations in ascending order or descending order.

iii) Read off the observations one by one from the data given and for each one, record a tally mark (I) against each observation. To facilitate the counting of tally marks, arrange them in group of five like (IIII), where every fifth tally is drawn across the preceding four. This is called 'Four and crossed method'. This process is continued till all the observations are considered.

iv) In the end, count all the tally marks in a row and write their number in the frequency column.

v) Write down the total frequency in the last row of the bottom.

Example: Prepare a discrete frequency distribution from the following data:

6	12	10	8	15	20	10	16	18	15
8	10	14	12	10	19	11	10	9	7

Discrete frequency distribution:

Variable	Tally Marks	Frequency
6	I	1
7	I	1
8	II	2
9	I	1
10	卌	5
11	I	1
12	II	2
14	I	1
15	II	2
16	I	1
18	I	1
19	I	1
20	I	1
Total	**–**	20

Steps for forming a Grouped frequency distribution: For better understanding lets understand the terminology associated with it.

Class: Various groups formed in a quantitative classification are called classes.

Class interval: The various intervals (symbols for the class) are called class-intervals (L_1–L_2).

Class frequency: The number of observations falling within a particular class or class interval is known as class frequency or frequency of the corresponding class.

Class limits: The pairs of the numbers that represent the class or class-interval are called limits. In a specified class the greatest value that a variable can take is called the upper class limit (L_2) and the least value it can take in that class is called lower class limit (L_1)

Magnitude of the class – interval or width of the class: The difference between the two limits is called as magnitude or width of the class.

Width = L_2–L_1

In case of inclusive series, the width of the class is equal to the difference between the lower class limits of the consecutive classes.

Mid-value or Mid-point of the class: The value being half way between the lower and upper class limits of a class – interval is called its mid-value or central-value or mid-point. Thus,

Mid-point = (Lower-limit + Upper-limit)/2

Now let's take one example to understand how group series is formed.

30 students appeared in an examination. The marks obtained out of 50 are as follows:

2	5	12	22	8	6	11	22	26	32	4	37	39	42	24
9	10	16	24	31	46	44	28	35	40	15	27	47	6	14

Inclusive Method			Exclusive Method		
Marks	Tally marks	No. of Students	Marks	Tally marks	No. of Students
1-10	𝍸 III	8	0-10	𝍸 II	7
11-20	𝍸	5	10-20	𝍸 I	6
21-30	𝍸 II	7	20-30	𝍸 II	7
31-40	𝍸 I	6	30-40	𝍸	5
41-50	IIII	4	40-50	𝍸	5
Total		30	Total		30



CHECK YOUR PROGRESS

Q.1: Distinguish between univariate and bivariate frequency distribution. (Answer in about 30 words)

.....

.....

.....

Q.2: What does the term class frequency mean? (Answer in about 20 words)

.....

.....

.....

2.6 GRAPHICAL REPRESENTATION OF DATA

Tabulation is one way of presenting data. Another way of presenting it is in the form of diagrams and graphs.

A. L. Boddington observes, “The wandering of a line is more powerful in its effect on the mind than a tabulated statement; it shows what is happening and what is likely to take place, just as quickly as the eye is capable of working.”

Meaning: The word ‘Graph’ could well be associated with the word ‘graphic’, which means ‘vivid’ or ‘spraining to life’. “This is exactly the kind of function which the graphic should perform for the table from which it is drawn, because most people can grasp pictures more readily than figures. Graphic methods enables statistician to present quantitative data in simple, clear and effective manner.”

Functions of Graph: The two major functions of graph are: (a) Tools of analysis & (b) Presentation of Data

As a tool of analysis it serves as a visual guide in planning of mathematical computations and general procedure of research study. It also helps the investigators to have the knowledge of development step-by-step, which becomes a good source of using his judgment and checking the result accuracy. Freehand curves may be fitted to data in more varied and flexible form than Mathematical curves.

Through graph, it is possible to present the complex data into simple and comprehensible form. Thus making comparison of the data a much easier task, as the data can be presented side by side.

Advantages of graphical presentation of data: Lowe has expressed the advantages of graphical presentation of data in his words “When properly constructed they readily show information that might otherwise be lost amid the details or chart show more than just the value of statistics, it also shows any trend that may be present and the way in which the trends change. Although this information is inherent in a table it may be quite difficult and time consuming to determine the existence and nature of trends from tabulation of data.”

The major advantages of graphical presentation of data are as under:

- **Easy to grasp and remember:** Because of their visual impact, the data presented through graphic and diagrammatic presentation are better grasped and remembered than the tabulated data.
- **Attractive and Impressive:** Graphs are always attractive and impressive than tabular form. In graph, all the features of data becomes visible at a glance and thus making it attractive and impressive.
- **Saves time and energy:** Because of its simple nature, it saves time and energy of both the statistician as well as observer.
- **Mathematical knowledge is not required:** There is no much requirement of technical knowledge or expertise in understanding the graphs, as they are very simple and easily comprehensible.
- **It makes complex thing simpler:** Even when data show highly complex relations among variables, these devices make them much clear. They thus greatly facilitate in the interpretation and analysis of data.
- **Location of positional averages:** It is possible to determine the positional averages like median, quartiles, mode etc., with the help of special kind of graphs.
- **Study of correlation:** The movements or the tendency of points (presented as graph) can easily be taken to know the correlation or regression between the variables. In simple words one can study the correlation between two sets of data, by observing the graph.
- **Forecasting and interpretation is easy:** Interpolation, extrapolation, forecasting of the data can easily be done with the help of graphs.
- **Comparisons:** Comparisons can be made between the two or more phenomena very easily with the help of graphs.

Disadvantages and limitations of graphical presentation of data:

Major disadvantages of graphical presentation are as under:

- **Accuracy is not testable:** Graphs do not reveal the accuracy of data. They only show the fluctuations of data. Hence, they cannot be used to test the accuracy of the data.

- **Incomplete information:** Graphs cannot be used to depict more than two features. It is not possible to have solutions of all the problems from the graph. Hence, they furnish incomplete and inadequate information.
- **Misleading impression:** Two different scales may show different fluctuations in data. They may often give misleading impressions.
- **Not for quotation use:** The graphs cannot be quoted in support of some statement.
- **Requires some skill:** Some special types of graphs like ratio-variation graphs etc. cannot be understood and interpreted by laymen.
- **Possibility of misuse:** A small challenge or error in measurement scale may create large difference in the shape of graph and can give misleading results.

Frequency distribution graphs in which the measurements, class-limits or mid-values are taken along X-axis and frequencies along Y-axis, are of following types:

(1) Histogram, (2) Frequency polygon, (3) Frequency curves, (4) 'Ogive' or Cummulative Frequency curves.

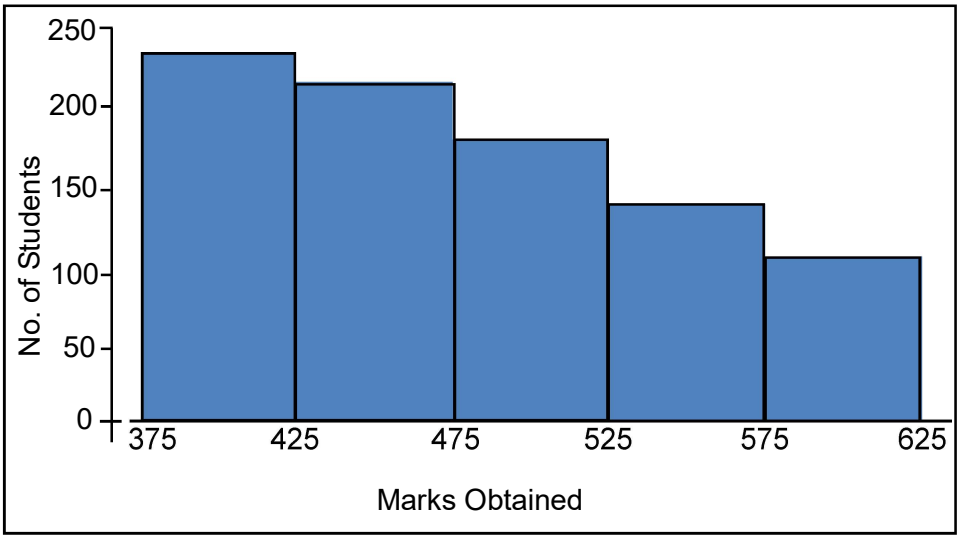
In the coming sections we will be discussing, one by one.

2.7 HISTOGRAM

Presentation of frequency distribution by a set of vertical bars-the Histograms, whose areas are proportional to the frequency represented, is the most popular and commonly used device. It consists of plotting the data as a series of rectangles. In construction of histograms, class intervals are shown on X-axis and frequencies on Y-axis. There as many rectangles as classes. The height of each rectangle represents the frequency of class. Histograms are also sometimes called as Staircase or Block Diagram.

Let's understand by taking one example:

Marks Obtained:	375-425	425-475	475-525	525-575	575-625
No. of Students:	225	180	175	125	100



Here the marks obtained shown by Class intervals are on X-axis while the number of students falls in the corresponding groups are on Y-axis.



LET US KNOW

- The distinction between a bar chart and a histogram can be graphically represented as follows:



Bar Chart Histogram

- A histogram is a two dimensional diagram, where both the width (base) and the length (height) of the rectangle are important. On the contrary, a bar diagram is one dimensional diagram, in which only length (height) of the bar occupies significant place as the width of the bar is arbitrary.

- The bar diagrams are prepared at proper spacing between two bars, while in histograms, the bars are adjacent to each other.
- In the histogram, class frequencies are shown by the area of rectangles, whereas in bar diagrams, they are shown by the height of the bars.

2.8 FREQUENCY POLYGON

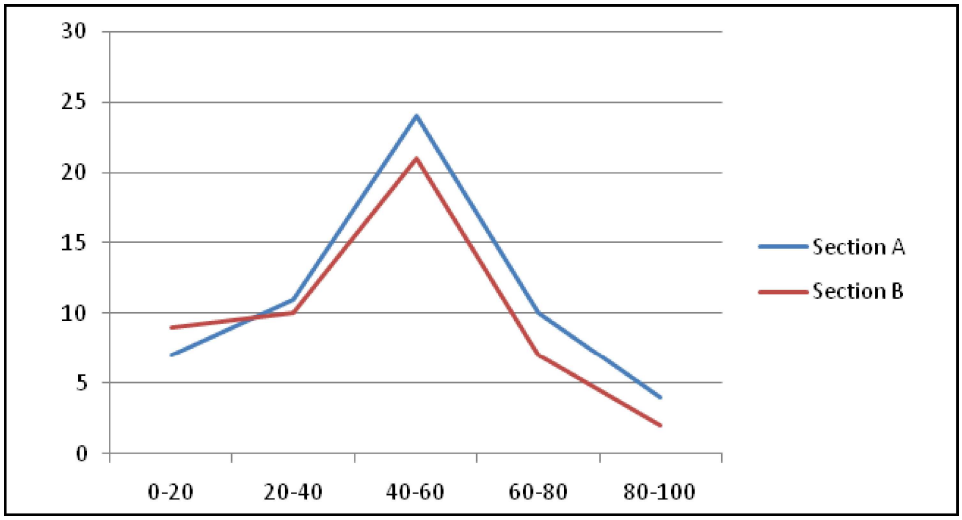
A frequency polygon is another way of graphic presentation of a frequency distribution, whether it is discrete or grouped or continuous. It can be effectively used to compare two or more frequency distribution.

A frequency polygon like any polygon consists of many angles. A histogram can be easily transformed into a frequency polygon by joining the mid-points of the rectangles by straight lines.

Even if required, one can draw straightway a frequency polygon by taking the mid-point of each class-interval and by joining the mid-points by the straight lines. Another point to be noted is that this can be done only when we have continuous series. In case of discrete distribution, this is not possible.

Let's clarify the concepts by taking one example:

Marks Obtained:	0-20	20-40	40-60	60-80	80-100
No. of Students (Section A):	7	11	24	10	4
No. of Students (Section B):	9	10	21	7	2



In the above frequency polygon the comparison can be done very easily as the two lines belonging to the two sections– Section A and Section B, are of different colours and can be easily understood.

Comparison of Frequency Polygon and Histogram:

- Frequency Polygon is a line graph where as Histogram is a collection of adjacent rectangles.
- Frequency Polygon is multi-dimensional, whereas Histogram is two-dimensional figure.
- Several frequency distributions can be plotted on the same axis as Frequency Polygon. This makes comparison- easier and meaningful. This is not possible in case of Histogram. However for the study of the relationship of the individual class frequencies to the total frequency, histogram is preferred to the frequency polygon, as the former gives a better picture.
- Frequency Polygon is a continuous curve and has all the distinct advantage of graphic representation, viz, it may be easily used to determine the scope, rate of change, estimates (interpolation and extrapolation), etc., wherever admissible. This is not possible in the case of Histogram.

2.9 OGIVE OR CUMULATIVE FREQUENCY CURVE

Till now, we have learnt about graphical representation of frequency distributions by using Histogram and Frequency Polygon. In this section, we will learn about the construction of cumulative frequency polygon or cumulative frequency curve or Ogive.

The technique of drawing the cumulative frequency polygons and cumulative frequency curves or Ogives is more or less the same. The only difference is that in case of simple frequency curves and polygons the frequencies are plotted against class marks of the class intervals where as in case of a cumulative frequency curve, the cumulative frequencies are plotted against the lower or upper limits of the class intervals, depending upon the manner in which the series has been cumulated. There are two methods of constructing an Ogive: Less than method and More than method.

Steps of constructing less than method ogive:

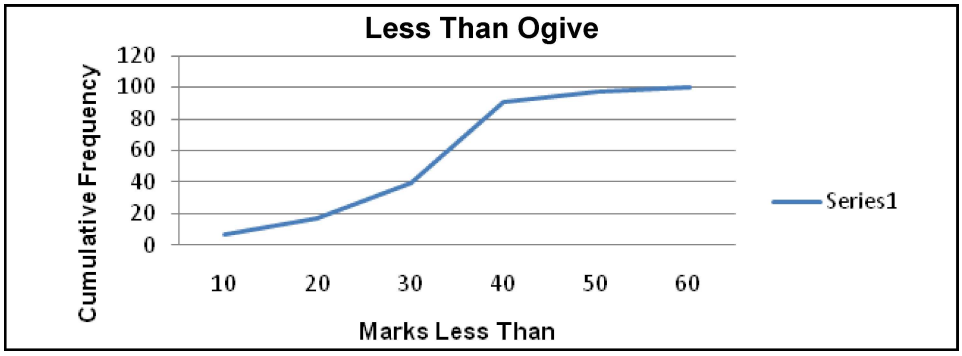
- Start with the upper limits of class intervals and add class frequencies to obtain the cumulative frequency distribution.
- Mark Upper class limits along X-axis on a suitable scale.
- Mark cumulative frequencies along Y-axis on a suitable scale.
- Plot the points (X_1, f_1), where X_1 is the upper limit of a class and f_1 is the corresponding cumulative frequency.
- Join the points obtained in Step 4 by a free hand curve to get the ogive.

Example: Draw an ogive for the following frequency distribution by less than method.

Marks:	0-10	10-20	20-30	30-40	40-50	50-60
No. of Students:	7	10	23	51	6	3

Ogive table:

Marks	No. of Students	Marks less than	Cumulative Frequency
0-10	7	10	7
10-20	10	20	17
20-30	23	30	40
30-40	51	40	91
40-50	6	50	97
50-60	3	60	100
Total	100		



Steps of constructing more than method ogive:

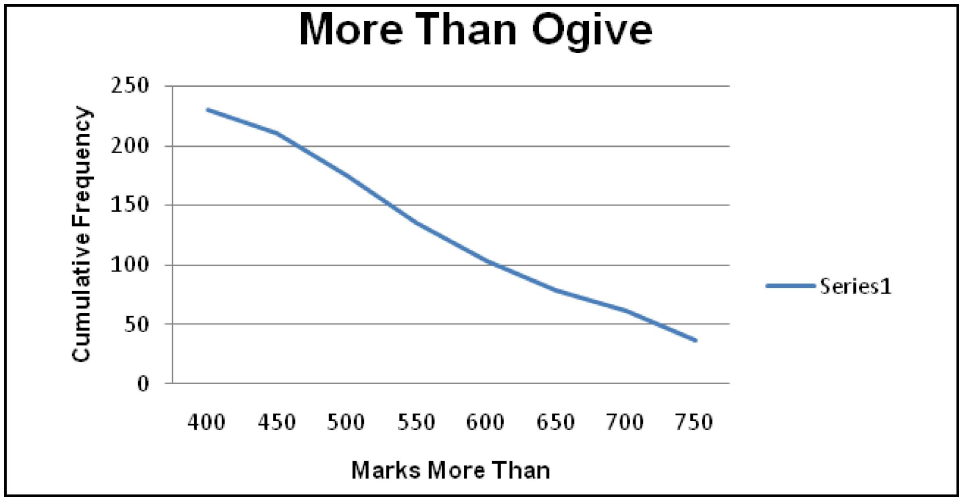
- Start with the lower limits of the class intervals and from the total frequency subtract the frequency of each class to obtain the cumulative frequency distribution.

- Mark the lower class limits along X-axis on a suitable scale.
- Mark the cumulative frequencies along Y-axis on a suitable scale.
- Plot the points (X_1, f_1), where X_1 is the lower limit of a class and f_1 is the corresponding cumulative frequency.
- Join the points obtained in Step 4 by a free hand curve to get the ogive.

For example: The frequency distribution of scores obtained by 230 students of a college in an Attitude test is as follows:

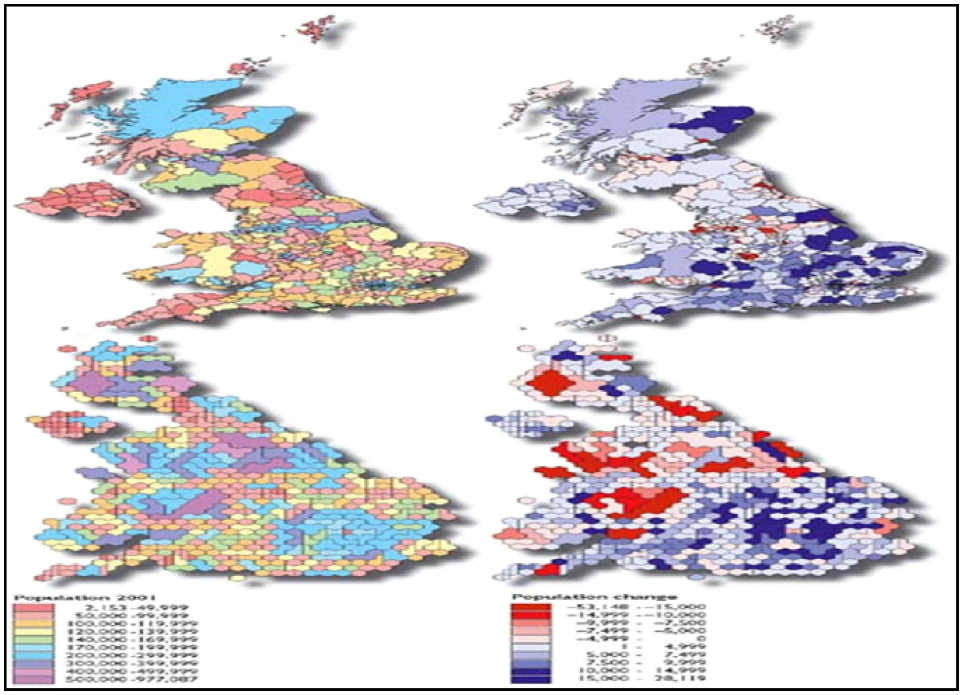
Scores:	400-450	450-500	500-550	550-600	600-650	650-700	700-750	750-800
No. of Students:	20	35	40	32	24	27	18	24

Scores	Number of Candidates	Scores More than	Cumulative Frequency
400-450	20	400	230
450-500	35	450	210
500-550	40	500	175
550-600	32	550	135
600-650	24	600	103
650-700	27	650	79
700-750	18	700	61
750-800	24	750	37
Total	230		



2.10 CARTOGRAM OR MONOGRAPH

In this technique, statistical facts are presented through maps accompanied by various types of diagrammatic presentation. They are generally used to present the facts according to geographical regions. Population and its other constituents like births, deaths, growth, density etc., production, imports, exports and several other facts are required to be compared on the regional as well as country basis. These facts can well be presented on the map with certain colours, dots, cross, points etc. Cartograms are simple and elementary form of visual presentations and easily understandable.



Description of Cartogram:

Population: The population of the UK in 2001, as counted in the 2001 Census, was the infamous One Number: 58,789,194, infamous because it was released as being the most accurate count ever made. It certainly was the most accurate, but it will still be several hundred thousand people adrift of reality in one direction or the other. The comparable figure, following counting and correcting in 1991, as 57,770,226. Thus in 10 years the population is thought to have grown by roughly one million people.The first map and cartogram pair on this page show the population of each local authority in the UK in 2001. At most almost a million people are living in

Birmingham, fewer than in 1991 as most major cities have experienced declining populations other than London . At the least, just over 2,000 people are now recorded as being resident in the Isles of Scilly.Local authorities tend to be more populous in the North as large metropolitan boroughs were created there in the 1970s. London was split in the 1960s into slightly smaller boroughs. Local authorities have their smallest populations in rural England and Northern Ireland. Outside of London it is these that are growing most in population. Authorities tend to be a little larger in population in Wales and Scotland as these were reorganised in the 1990s to be so. Within both of these countries the population is falling in some parts and rising in others. In general, however, the population of the UK is slowly moving towards the South and particular parts of the South that are now growing very quickly.

2.11 PICTORIAL PRESENTATION OF DATA

Also called as Pictogram. The technique of presenting the data through pictures was developed by Dr. Otto Neurath of Vienna. This is the reason that this method is sometimes called as Vienna Method. This is also known as Iso-type method.

According to New World Encyclopedia “A pictogram or pictograph is a symbol representing a concept, object, activity, place or event by illustration. Pictography is form of writing in which ideas are transmitted through drawing. Such pictograms are characterized by their simplified style, which omits all details that are unnecessary to the desired communication. It is the basis for some of the earliest forms of structured written languages, such as Cuneiform and, to some extent, Hieroglyphs.”

In this method the magnitude of the particular phenomenon, being studied, is drawn. The size of pictures is kept proportional to the values of the different magnitudes to be presented. Pictures are more attractive and of lasting impression on mind. Accordingly, the use of pictogram techniques is being frequently used in every field of knowledge, specially in advertising world.

Example: Information about 300 children of a school who come to school by different modes of transportation.

→1 face represents 10 children

See the picture graph or pictograph to understand the status:

Modes	School-going children
Auto-rickshaw	
Car	
Bicycle	
Bus	
On foot	

Source: http://www.math-only-math.com/examples-of-pictographs.html

Information gathered from the above table:

- Number of students going to school by different modes of transportation:

  Auto-rickshaw = 6 × 10 = 60, Car = 4 × 10 = 40,

  Bicycle = 7 × 10 = 70, Bus = 10 × 10 = 100,

  On foot = 3 × 10 = 30

- Total number = 60 + 40 + 70 + 100 + 30 = 300

The above explanation on the examples of pictographs or pictorial representation will help us to solve different types of questions on pictograph.



CHECK YOUR PROGRESS

Q.3: Why is pictorial presentation of data also called as Vienna method? (Answer in about 20 words)

.....

.....

.....

Q.4: Mention the steps involve in the construction of a less than ogive curve. (Answer in about 40 words)

.....

.....

.....

.....



2.12 LET US SUM UP

- A tabular presentation of the data in which the frequencies of values of a variable are given along with them is called a frequency distribution.
- Tabulation is one way of presenting data. Another way of presenting it is in the form of diagrams and graphs.
- Presentation of frequency distribution by a set of vertical bars-the Histograms, whose areas are proportional to the frequency represented, is the most popular and commonly used device.
- A frequency polygon is another way of graphic presentation of a frequency distribution, whether it is discrete or grouped or continuous.
- There are two methods of constructing an Ogive: Less than method and More than method.
- The technique of presenting the data through pictures was developed by Dr. Otto Neurath of Vienna.



2.13 FURTHER READING

1) Black, Ken (2007). *Business Statistics for Contemporary Decision Making*. New Delhi: Wiley India Pvt. Ltd.
2) Garrett, H. E. (2005). *Statistics in Psychology & Education*. New Delhi: Paragon International Publishers.
3) Sharma, R. D. (2009). *Mathematics for Class X*. New Delhi: Dhanpat Rai Publications.
4) Srivastava, S. C. & Srivastava, S. (2009). *Fundamentals of Statistics*. New Delhi: Anmol Publications Pvt. Ltd.



2.14 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A frequency distribution which shows the frequency of occurrence of different values of a single variable is called as

univariate frequency distribution. On the other hand, a frequency distribution based on two variables is called as bivariate frequency distribution.

Ans. to Q. No. 2: The number of observations falling within a particular class or class interval is known as class frequency or frequency of the corresponding class.

Ans. to Q. No. 3: The technique of presenting the data through pictures was developed by Dr. Otto Neurath of Vienna. This is the reason that this method is sometimes called as Vienna Method.

Ans. to Q. No. 4: The steps involved in the construction of a less than method ogive are:

1) Start with the upper limits of class intervals and add class frequencies to obtain the cumulative frequency distribution.
2) Mark Upper class limits along X-axis on a suitable scale.
3) Mark cumulative frequencies along Y-axis on a suitable scale.
4) Plot the points (X1, f1), where X1 is the upper limit of a class and f1 is the corresponding cumulative frequency.
5) Join the points obtained in Step 4 by a free hand curve to get the ogive.



2.15 MODEL QUESTIONS

Q.1: What are the advantages of Histogram?

Q.2: Draw Histogram and Frequency Polygon in the same graph, of the given data:

C.I.:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F:	8	5	16	14	22	14	12	9

Q.3: Write the main features of Ogive.

Q.4: What are the advantages of Cartogram?

*** ***** ***

UNIT 3: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE

3.1 Learning Objectives

3.2 Introduction

3.3 Measures of Central Tendency

- 3.3.1 Definition
- 3.3.2 Characterstics of a Good Average
- 3.3.3 'Σ' Symbol
- 3.3.4 Different Types of Measures of Central Tendency

3.4 Let Us Sum Up

3.5 Further Reading

3.6 Answers to Check Your Progress

3.7 Model Questions

3.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the meaning of measures of central tendency or averages
- describe the various types of averages along with their advantages, limitations and uses
- explain the meaning of weighted mean.

3.2 INTRODUCTION

In the earlier two units, we have discussed the classification, tabulation, diagrammatic and graphic representations of data that has considerably advanced our knowledge of the real significance of statistical data. We need a numerical expression which summaries the characteristics of the whole set of data under study. 'Measures of Central Location' or 'Measures of Central Tendency' popularly known as 'Averages' serve this purpose. A figure which represents a series of values should obviously be greater than the lowest value and less than the highest value. It should be a value somewhere between these two limits, possibly at the centre where

most of the values of the series center. Such a figure is called a Measure of Central Tendency.

Measures of central tendency, very often, are not fully representative of a given set of data. This happens when the extent of variation of individual values in relation to the average, or in relation to the other values is large. As an illustration, let us observe the following three series:

Series A	40	40	40	40	40
Series B	35	39	41	42	43
Series C	10	18	35	57	80

In the first series the arithmetic mean or simply mean (sum of the values divided by the number of values) is 40 and the values of all the items are identical each being equal to 40. The mean fully represents the series in general and the individual items in particular. The data (items or observations) are not at all scattered. In the second series, although the mean is 40, all the observations are not very much scattered as the minimum value of the series is 35 and the maximum value is 43. Hence in case of the second series also the mean is a good representation of the series. Although none of the observations of the series is equal to the mean of the series yet the discrepancy between the mean and any other observation is not so significant. In case of the third series, we observe that all the items or observations of the series are different. This series also has the same mean 40. In case of this series the observations are widely scattered. Clearly, in case of this series the mean neither satisfactorily represents the entire series in general nor the individual items of the series in particular. Thus we have observed that although all the above three series have the same average (arithmetic mean is a method of measuring average) yet they widely differ from one another in terms of their formation. When the extent of variation (deviation or scatteredness) of the individual values (items or observations) of a distribution or series in relation to their average or in relation to the other values is large then measures of central tendency or averages cannot be representative of the distribution. Hence it is important for any investigation not only to know the average of any type (mean, median or mode) but also the scatteredness of the various observations of a distribution.

3.3 MEASURES OF CENTRAL TENDENCY OR AVERAGES

3.3.1 Definition

An average of a distribution (i.e. a distribution of the values of a variable like height, weight, income etc.) is a representative value of that distribution. This representative value usually lies at the central part of distribution. ‘‘A measure of central tendency or an average of a certain distribution is nothing but a representative value of that distribution which enables us to comprehend in a single effort the significance of the whole.’’

Unit of average: The unit of average of a distribution is the unit of that distribution.

3.3.2 Characteristics of a Good Average

An average will be termed as a good if it possesses the following characteristics:

i) It should be easy to understand and calculate.

ii) It should be rigidly defined which means that the definition should be so clear that its interpretation does not differ from person to person.

iii) The average should be based on all the values of the variable.

iv) The average should have sampling stability. This means that the value of an average calculated from various independent random samples of the same size from a given population should not vary much from another.

v) It should be capable of further algebric treatment.

vi) The average should not be unduly affected by extreme values.

3.3.3 ‘Σ’ Symbol

In order to denote sum (i.e., a total of certain quantities) the Greek letter ‘Σ’ (capital sigma) is used. For example, if variable x

takes the values $x_1, x_2, x_3, \ldots, \ldots, ..., x_n$ then the sum of these values of the variable i.e. $(x_1, x_2, x_3, \ldots, \ldots, ..., x_n)$ is denoted by $\sum_{i=1}^{n} x_i$ or $\sum x$. The symbol $\sum_{i=1}^{n} x_i$ means that the lower limit of i is 1 and the symbol Σ means that all the values of x_i for i = 1, 2, 3, …, …, ..., n are to added. Thus $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots \ldots ... + x_n$. Again, the symbol $\sum x$ implies 'sum of the values of x'.

3.3.4 Different Types of Measures of Central Tendency

The following three types of measures of central tendency or averages are in use.

(a) Mean, (b) Median, and (c) Mode.

a) MEAN: There are three types of mean, namely,

(i) Arithmetic Mean (A.M.), (ii) Geometric Mean (G.M.) and

(iii) Harmonic Mean (H.M.)

So far as mean is concerned, we shall discuss arithmetic mean only since this is the most popular technique among the different types of mean.

In unit two, we have already discussed about arithmetic mean and geometric mean. Here we will discuss some special case related to Arithmetic mean.

The A.M. of a variable x is generally denoted by $\bar{x}$ and is defined as the sum of the values of divided by the total number of values of x.

Thus, if $x_1, x_2, x_3, ..., x_n$ be n values of x then A.M ($\bar{x}$) will be given by,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

In case of ungrouped frequency distribution table:

$x : x_1 \quad x_2 \quad x_3 \quad ... \quad x_n$

$f : f_1 \quad f_2 \quad f_3 \quad ... \quad f_n$

the A.M. ($\bar{x}$) will be given by,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + ... + f_nx_n}{f_1 + f_2 + f_3 + ... + f_n}$$

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

$$= \frac{\sum fx}{\sum f} = \frac{\sum fx}{N}$$

where, $N = f_1 + f_2 + f_3 + ... + f_n$

- **Arithmetic Mean (A.M.) of grouped frequency distributions:**

In order to determine the A.M. of a grouped frequency distribution an assumption is made that the observations included in a class represented by a class interval are concentrated around the centre of that class interval. For example, if in a distribution of marks of some students the frequency of the class interval 50-60 is 8 (say) then we assume that each of 8 students is getting marks around 55. Consequently we may approximately take that each of the 8 students is getting 55 marks which is the mid value of the class interval 50-60 and we say '8 is the frequency of 55'. To obtain the A.M. of a grouped frequency distribution, we take the frequencies of the different class intervals to be the frequencies of the mid-values of the corresponding classes. This converts a grouped frequency distribution to a discrete (or ungrouped) frequency distribution. Here by applying the arithmetic mean formula for a discrete or ungrouped frequency distribution we can find the arithmetic mean of a grouped frequency distribution.

Thus, in case of grouped frequency distribution A.M. ($\bar{x}$) is defined as $\bar{x} = \frac{\sum fx}{N}$

where, $N = f_1 + f_2 + f_3 + ... + f_n = \sum f$ and x is the mid value of a class.

Example 1: Find A.M. of the following frequency distribution:

x	1	2	3	4	5	6	7	8	9
y	7	11	16	17	26	31	11	1	1

Solution: First of all we shall prepare the following frequency table:

x	f	fx
1	7	7
2	11	22
3	16	48
4	17	68
5	26	130
6	31	186
7	11	77
8	1	8
9	1	9

N = 121 $\Sigma fx = 555$

A.M. $(\bar{x}) = \frac{\sum fx}{N} = \frac{555}{121} = 4.59$ (Approx.)

Example 2: Determine mean of the following distribution:

Daily wages (in Rs.):	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of workers:	6	5	8	15	7	6	3

Solution: Since there are three methods of obtaining mean, namely arithmetic mean (A.M.) method, geometric mean (G.M.) method and harmonic mean (H.M.) method, we shall apply A.M. method to find mean of the given distribution. (We shall see subsequently that A.M. is the most popularly used measure of central tendency.)

To find mean by applying arithmetic mean technique we form the following table:

Wages	Mid value ($\bar{x}$) = $\frac{l_1 + l_2}{2}$	No. of workers (f)	fx
0-10	5	6	30
10-20	15	5	75
20-30	25	8	200
30-40	35	15	525
40-50	45	7	315
50-60	55	6	330
60-70	65	3	195
		N = 50	Σfx = 1670

Mean ($\bar{x}$) = $\frac{\sum fx}{N} = \frac{1670}{N} = 33.4$

i.e.; The required mean (i.e. average) wage = **Rs. 33.40**

Note: l_1 and l_2 imply respectively the lower limit and the upper limit of a class interval.

ALTERNATIVE METHODS OF FINDING A.M.:

a) Assumed Mean Method or Short-Cut Method: In this method A.M. $\bar{x}$ is calculated by using the following formula:

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

where A = Assumed mean of x, d = x – A, N = total frequency. This formula is used when the values of 'd' do not have a common factor.

b) Step-deviation Method: If the values of (d = x – A) in the above formula have one or more common factors then this method is applied. The formula is:

$$\bar{x} = A + \frac{\sum fd'}{N} \times h$$

where $d' = \frac{d}{h}$, h = Highest common factor of 'd'.

In case of grouped frequency distributions with equal class intervals, h is equal to the length of the class intervals.

A note on assumed mean: Any value of a variable x can be considered to be the assumed mean of x. However, the assumed mean should be taken from the central part of the values

Example 3: Determine arithmetic mean of the following distribution:

Height (in cm)	130-134	135-139	140-144	145-149
Frequency	5	15	28	24
Height (in cm)	150-154	15-159	160-164	
Frequency	17	10	1	

Solution: We denote height by the variable x and we take the assumed mean A of x to be 147.

Class Interval	Mid value (x)	Frequency (f)	$d = x - A$ A = 147	$d' = \frac{d}{h}$ h = 5	fd'
130-134	132	5	–15	–3	–15
135-139	137	15	–10	–2	–30
140-144	142	28	–5	–1	–28
145-149	147	24	0	0	0
150-154	152	17	5	1	17
155-159	157	10	10	2	20
160-164	162	1	15	3	3
		N = 100			$\Sigma fd' = -33$

Now, A.M. $(\bar{x}) = A + \frac{\sum fd'}{N} \times h$

$= 147 + \frac{-33}{100} \times 5$

$= 147 + \frac{-33}{20}$

$= 147 - 1.65$

$= 145.35$

i.e. the required A.M. = **145.35 cm**

PROPERTIES OF A.M.:

Property 1: If each value of a variable is increased (decreased) by a constant c, then the A.M. of the new values is increased (decreased) by c. if each value of a variable is multiplied by a constant c then the A.M. of the new values is c times the A.M. of the original values. Again, if each value of a variable is divided by a constant c ($c \neq 0$) then the A.M. of the new values is equal to the A.M. of the original values divided by c.

Note that if $u = \frac{x \pm a}{b}$, then $\bar{u} = \frac{\bar{x} \pm a}{b}$ where a and b are constants.

Property 2: The sum of the deviations measured from the mean is zero.

In case of individual series $\Sigma(x - \bar{x}) = 0$, where $\bar{x} = \Sigma x/n$ and in case of frequency distributions $\Sigma f(x - \bar{x}) = 0$,

where $\bar{x} = \Sigma fx/N$, $N = \Sigma f$.

Property 3: If the A.M. of a distribution having n_1 value is $\bar{x}_1$ and the A.M. of another distribution having n_2 values is $\bar{x}_2$ then the A.M. $\bar{x}$ of the combined distribution is:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

This property can be extended for any number of distributions.

Example 4 (a): If $\bar{u} = \frac{x - 55}{10}$ and $\bar{x} = 59$, what is the value of $\bar{u}$.

b) The A.M. of the values of a variable x is 25.

i) If each value is increased by 5, what will be the new A.M.?

ii) If each value is decreased by 7, what will be the new A.M.?

iii) If each value is multiplied by 2.5 what will be the new A.M.?

iv) If each value is divided by 4, what will be the new A.M.?

c) Two series with 28 and 36 observations have means 2.9 and 5.6 respectively. Find the mean of the combined series.

Solution (a) : $u = \frac{x - 55}{10}$

$$\Rightarrow \bar{u} = \frac{\bar{x} - 55}{10}$$

When $\bar{x} = 59$, $\bar{u} = \frac{59 - 55}{10} = \frac{4}{10} = 0.4$

Solution (b) :

i) The A.M. of x i.e. $\bar{x} = 25$, Let u = x + 5.

(u is the variable each of whose values is more than the corresponding value of x by 5.)

$\therefore \bar{u} = \bar{x} + 5 = 25 + 5 = 30$

The new A.M. = 30

ii) Let u be the variable each of whose values is less than the corresponding value of x by 7.

$\therefore u = x - 7$

$\Rightarrow \bar{u} = \bar{x} - 7 = 25 - 7 = 18$

i.e. the new A.M. = 18

iii) If u is the variable each of whose values is 2.5 times the corresponding value x then u = 2.5x

$\Rightarrow \bar{u} = 2.5\ \bar{x} = 2.5 \times 2.5 = 6.25$

Thus the new A.M. = 6.25

iv) Let $u = \frac{x}{4}$

$$\Rightarrow \bar{u} = \frac{\bar{x}}{4} = \frac{25}{4} = 6.25$$

Thus the new A.M. = 6.25

Solution (c): Let $\bar{x}$ be the mean of the combined series.

Then $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

where $n_1 = 28$, $n_2 = 36$, $\bar{x}_1 = 2.9$, $\bar{x}_2 = 5.6$.

Now, $\bar{x} = \frac{28 \times 2.9 + 36 \times 5.6}{28 + 36}$

$= \frac{81.2 + 201.6}{64}$

$= \frac{282.8}{64}$

$\therefore$ A.M = 4.42

Determination of Arithmetic Mean in case of Cumulative Frequency Distributions: We shall illustrate with the following example how arithmetic mean of cumulative frequency distributions can be determined.

Example 5: The following are the marks obtained by the students of class XII of a certain Higher Secondary School. Find the average marks using arithmetic mean technique.

Marks	No. of students
Less than 10	5
Less than 20	17
Less than 30	31
Less than 40	41
Less than 50	49

Solution: The above cumulative frequency distribution should first be converted into a simple frequency distribution as under:

We convert the given cumulative frequency distribution to the simple frequency distribution as follows:

Marks	No. of students
0 - 10	5
10 - 20	17 - 5 = 12
20 - 30	31 - 17 = 14
30 - 40	41 - 31 = 10
40 - 50	59 - 41 = 8

Now arithmetic mean of the data can be obtained by using the Direct Method as under:

Calculation of A.M.:

Marks	Mid value (x) = $\frac{l_1 + l_2}{2}$	No. of workers (f)	fx
0 - 10	5	5	25
10 - 20	15	12	180
20 - 30	25	14	350
30 - 40	35	10	350
40 - 50	45	8	360
		N = 49	fx = 1265

$$\text{A.M.} = \left(\bar{x}\right) = \frac{\sum fx}{N} = \frac{1265}{49} = 25.81$$

Note: Student will solve the problem by using step-deviation method.

Example 6: Calculate Arithmatic Mean from the following data using short cut method.

Marks (out of 50):	1–10	11–20	21–30	31–40	41–50
No. of students:	5	7	10	6	2

Solution: Calculation of Arithmatic Mean

Here A = 25.5 (Assumed mean)

Marks	f	Mid-value x	fx	d = x – A	fd
1-10	5	5.5	27.5	–20	–100
11-20	7	15.5	108.5	–10	–70
21-30	10	25.5	255.0	0	0
31-40	6	35.5	213.0	+10	60
41-50	2	45.5	91.0	+ 20	40
Total	N = 30		Σfx = 695.0		Σfd = –70

By short cut method

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

$$= 25.5 + \frac{-70}{30}$$

$$= 25.5 - 2.33$$

$$= 23.17 \text{ marks.}$$

Example 7: Calculate mean from the following data.

Value	frequency
Less than 10	4
Less than 20	10
Less than 30	15
Less than 40	25
Less than 50	30
Less than 60	35
Less than 70	45
Less than 80	65

Solution: In the problem cumulative frequencies and classes are given. We will first convert the data in simple series from the given cumulative frequencies. After this, we will calculate the mean.

Here Assumed mean, A = 35.

Calculation of Mean

Value	Individual frequency	Mid-value (xi)	$d' = \frac{x_i - 35}{10}$	fd'
0–10	4	5	–3	–12
10–20	10 – 4 = 6	15	–2	–12
20–30	15 – 10 = 5	25	–1	–5
30–40	25 – 15 = 10	35	0	0
40–50	30 – 25 = 5	45	1	5
50–60	35 – 30 = 5	55	2	10
60–70	45 – 35 = 10	65	3	30
70–80	65 – 45 = 20	75	4	80
	N = 65			$\Sigma fd' = 96$

$$\therefore \bar{x} = a + \frac{\sum fd'}{N} \times H$$

$$\therefore A = 35, \Sigma fd' = 96, N = 65, h = 10$$

$$\therefore \bar{x} = 35 + \frac{96}{65} \times 10 = 35 + 14.77 = 49.77$$



LET US KNOW

Arithmetic mean is the most popularly used measure of averages. Most people are familiar with this technique. The basic advantages of this measure over other measures of central tendency are: (i) It is comparatively easier to understand and calculate; (ii) It provides a good basis for comparison; (iii) It can be determined for all values namely positive, negative or zero'; (iv) It is amenable for further mathematical treatment. Mainly because of its easiness in understanding and calculating, it has become most popular and hence it is called an ideal average. Usually mean or average means arithmetic mean. It is widely used in social, economic and business problems.

Weighted Arithmetic Mean : When we determine the arithmetic mean of series by assigning weights to the different quantities of the series depending upon their relative importance then this arithmetic mean will be called the weighted arithmetic mean. For example, if we want to know the change in the cost of living of a particular community over a period of time then we must assign appropriate weights to the quantities of the different items of consumption of that community while constructing the index. Since all the commodities consumed by them are not of equal importance hence simple mean of the prices of the commodities consumed by them will not reflect the true cost.

Let the weights attached to the quantities $x_1, x_2, ..., x_n$ be $w_1, w_2, ..., w_n$ respectively. The weighted mean of these quantities is denoted by $\bar{x}_w$ and is given by:

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots w_n} = \frac{\sum wx}{\sum w}$$

Note: (i) This formula is similar to the formula for obtaining A.M. of a frequency distribution. Here instead of the frequencies f_1, f_2, ..., ..., ..., f_n of x_1, x_2, ..., ..., ..., x_n respectively we take their respective weights w_1, w_2, ..., ..., ..., w_n.

USES OF WEIGHTED MEAN:

i) In order to determine the mean of the quantities whose weights are not equal the formula for weighted mean is applied.

ii) In order to determine the mean of the sub-series of a series, the formula for weighted mean is used.

Example 6: The marks obtained by three students A, B and C in physics, Chemistry and Mathematics out of 100 in each subject in a certain entrance test are:

	Physics	Chemistry	Mathematics
A	60	65	70
B	75	60	50
C	55	60	65

i) Rank the three students on the basis of their performance if equal weights are given to the subjects.

ii) Rank the students if weights are given as below:
Physics: 30%, Chemistry: 20%, Mathematics: 50%

Solution: Let x, y and z denote the marks obtained by A, B, and C respectively is Physics, Chemistry and Mathematics.

Let $\bar{x}(\bar{x}_w)$, $\bar{y}(\bar{y}_w)$ and $\bar{z}(\bar{z}_w)$ denote the simple (weighted) average of marks obtained by A, B and C respectively.

[Recall that when equal weights are given to all observations then it is the case of simple average or average and when different weights are assigned, then it is the case of weighted average. Again, average is usually obtained by arithmetic mean technique.]

i) $\bar{x} = \frac{60+65+70}{3} = \frac{195}{3} = 65$

ii) $\bar{y} = \frac{75+60+50}{3} = \frac{185}{3} = 61.67$

iii) $\bar{z} = \frac{55+60+65}{3} = \frac{180}{3} = 60$

From the average marks obtained by A, B and C we find that the ranking positions of A, B and C are 1st, 2nd and 3rd respectively. Again,

$$\bar{x}_w = \frac{60\times 30\% + 65\times 20\% + 70\times 50\%}{30\% + 20\% + 50\%}$$

$$= \frac{60\times 0.3 + 65\times 0.2 + 70\times 0.5}{0.3+0.2+0.5} \left[\because 30\% = \frac{30}{100} = 0.3\text{etc.}\right]$$

$$= \frac{18+13+35}{1} = 66$$

$$\bar{y}_w = \frac{75\times 0.3 + 60\times 0.2 + 50\times 0.5}{0.3+0.2+0.5}$$

$$= \frac{22.5+12+25}{1} = 59.5$$

$$\bar{z}_w = \frac{55\times 0.3 + 60\times 0.2 + 65\times 0.5}{0.3+0.2+0.5}$$

$$= \frac{16.5+12+32.5}{1} = 61$$

We find from above that when the weights are assigned as given to the marks obtained in Physics, Chemistry and Mathematics, the ranking positions of A, B and C becomes 1st, 2nd and 3rd respectively.

b) MEDIAN: The median of a series or distribution in ascending or descending order is that observation of the distribution which divides the distribution into two equal parts. Thus there are equal number of observations on the right and on the left of the median value.

In order to determine the median of an individual series, first of all we have to observe whether the values (observations) are

in a define order or not i.e. whether the values are in ascending or in descending order or not. If the values are not in a define order, then these values are to be arranged either in ascending or in descending order. If there are odd number of values in the series, then the $\left(\frac{n+1}{2}\right)$th value from the beginning (and also from the end) will be the median. If the number of values is even then the arithmetic mean of the $\frac{n}{2}$th value and the $\left(\frac{n}{2}+1\right)$th value will be the median.

Example 7: Determine median for the following series:

i) 77, 73, 72, 70, 75, 79, 78

ii) 94, 33, 86, 68, 32, 80, 48, 70

Solution:

i) Arranging the values of the series in ascending order, we get: 70, 72, 73, 75, 78, 79

No. of terms in the series = 7 = An odd number

The required median = $\frac{7+1}{2}$th term i.e; 4th term = 75

ii) Arranging the data (values or observations) in ascending order, we get: 32, 33, 48, 68, 70, 80, 86, 94

No. of term in the series = 8 = An even number

Now, $\frac{n}{2}$th term = $\frac{8}{2}$th term = 4th term and $\left(\frac{n}{2}+1\right)$th term = 5th term

$\therefore$ The required median = $\frac{68+70}{2}$ = 69.

MEDIAN OF AN UNGROUPED FREQUENCY DISTRIBUTION:

In order to determine the median of an ungrouped frequency distribution which is in ascending order we have to, first of all, form a cumulative frequency table. If the number of observations is odd then the $\frac{N=1}{2}$th (N being the total frequency) term

(observation) will be the median. If we find that $\frac{N=1}{2}$ is greater then x but less than or equal to y where x and y are two consecutive values in the cumulative frequency column, then the observation whose cumulative frequency is y will be the median. Again, if the number of observations is even (i.e., if the total frequency is even) then as in case of individual series we are to determine the $\frac{n}{2}$th and the $\left(\frac{n}{2}+1\right)$th terms (observations). The A.M. of these two terms will be the median. Very often these two terms are the same and consequently when these two terms are equal, the $\frac{n}{2}$th term can be taken as the median.

Example 8: Determine median for the following distribution:

Wages (Rs.):	20	21	22	23	24	25	26	27	28
No. of workers:	8	10	11	16	20	25	19	9	6

Solution:

Wages (Rs.)	No. of workers (f)	Cumulative frequency (f)
20	8	8
21	10	18
22	11	29
23	16	45
24	20	65
25	25	90
26	19	109
27	9	118
28	6	124
	N = 124	

Here total frequency (i.e. total no. of observations) = 124 which is even. Hence A.M. of the $\frac{n}{2}$th and the $\left(\frac{n}{2}+1\right)$th terms will be the median.

Now $\frac{n}{2} = \frac{124}{2} = 62$ and $\frac{N=1}{2} = 63$. We find from the cumulative frequency column that 62 and 63 lie between 45 and 65. Since 65 is the cumulative frequency of 24 hence each of the 62th and the 63th terms will be 24.

Hence the required median = Rs. 24.

Note: The 62th term = Rs. 24, and the 63th term = Rs. 24 and their A.M.= $\frac{Rs.24 + Rs.24}{2} = \frac{Rs.48}{2}$ = Rs. 24.

MEDIAN OF A GROUPED FREQUENCY DISTRIBUTION: In case of grouped frequency distributions, one may consider the $\frac{n}{2}$th observation as the median if N is even. When N is odd, $\left(\frac{n}{2}+1\right)$th observation will be the median.

Median (Me) = $L + \frac{\frac{N}{2} - f_c}{f} \times 1$

Where L = Lower class limit (lower class boundary) in case of exclusive (inclusive) classification.

f = Frequency i.e. simple frequency of the median class

f_c = Cumulative frequency of the class preceeding the median class.

N = Total frequency

I = length of the median class

Note: This formula for obtaining median of a grouped frequency distribution holds only when the distribution is in ascending order. If the distribution is in descending order then it is to be first of all arranged in ascending order in order to apply the above formula.

Example 9: Determine median for the following distribution:

Weekly wages (Rs.)	50-55	55-60	60-65	65-70	70-75
No. of workers	6	10	22	30	16
Weekly wages (Rs.)	75-80	80-85			
No. of workers	12	15			

Solution:

Weekly wages (Rs.)	No. of workers (f)	Cumulative frequency (fc)
50-55	6	6
55-60	10	16
60-65	22	38
65-70	30	68
70-75	16	84
75-80	12	96
80-85	15	111
	N = 111	

Here N = 111, which is odd.

Now, $\frac{N=1}{2}$ th term = $\frac{111+1}{2}$ th term = 56th term. From the cumulative frequency table we find that the 56th term lies in the class 65 - 70. 65 - 70 is the median class.

Now, median = $L+\frac{\frac{N}{2}-fc}{f}\times 1$

Here L = 65, f = 30, fc = 38, N = 111, l = 5

Median = $65+\frac{\frac{111}{2}-38}{30}\times 5$

$= 65+\frac{55.5-38}{30}\times 5$

$= 65+\frac{17.5}{6}$

= 65 + 2.92

= 67.92

i.e. the required median = Rs. 67.92

Advantages and Limitations of Median:

- **Advantages:**
  - i) Extreme values do not affect median.
  - ii) Median is easy to understand; it is easy also to determine.
  - iii) Median can also be determined graphically

iv) Median can be determined for distributions having open end class intervals.

➢ **Limitations:**

i) In order to determine the median of a distribution, the distribution must be arranged in a define order if it is not in an order. This is not needed in other measures of central tendency.

ii) Median of a distribution is not based on all the observations of the distribution.

Uses of median: In order to determine the average of distribution having open-end class intervals median is the best measure. In case of income distribution the use of median gives better results.

c) MODE: The mode of a distribution is that observation of the distribution whose frequency is the maximum. It is to be noted that mode is not unique which means that a distribution may have more than one modes.

Clearly an individual or single series (distribution) does not have a mode. In many cases of ungrouped frequency distributions, mode can be detected simply by observation. Mode of a grouped frequency distribution is obtained by using the following formula.

$$\text{Mode } (M_o) = L + \frac{f_1 = f_0}{2f_1 - f_0 - f_2} \times I$$

Where L = Lower limit/lower boundary of the model class

f_o = Frequency of the class preceeding the model class

f_1 = Frequency of the model class

f_2 = Frequency of the class succeeding the model class

And I = Length of the model class

➢ **Note:** i) The class (specified by a class interval) whose frequency is the maximum is called the model class.

ii) The above formula is used when all the classes are of equal length.

Example 10: Determine mode for the following distribution:

Marks	1-5	6-10	11-15	16-20	21-25	26-30
No. of students	7	10	16	32	24	18
Marks	31-35	36-40	41-45			
No. of students	10	5	1			

Solution: Since the frequency of the class 16-20 is the maximum, hence this class is the model class. The class intervals of the given distribution are as per the inclusive method of classification and hence in determining mode we must take the lower boundary of the model class.

Now, Mode $= L + \frac{f_1 = f_0}{2f_1 - f_0 - f_2} \times I$

$= 15.5 + \frac{32 - 16}{2 \times 32 - 16 - 24} \times 5$

$= 15.5 + \frac{16}{64 - 40} \times 5$

$= 15.5 + \frac{16}{24} \times 5$

= 15.5 + 3.33 = **18.33 marks**

Example 11: Determine mode/modes, if any, of the following series:

i) 3, 4, 5, 2, 3, 4, 1, 6, 4;

ii) 7, 9, 11, 7, 6,5, 9, 13;

iii) 3, 5, 6, 7, 9, 12, 3, 6, 5, 9, 12, 7

Solution:

i) The number 4 is repeated the maximum of number times (thrice). Hence 4 is the mode of the given series.

ii) We find that the frequency of 7 and 9 is the maximum each being equal to 2. Hence the two modes of the series are 7 and 9.

iii) Since the frequency of each observation is the same (being 2 in each case), hence the given series has no mode.

Advantages and Limitations of Mode:

- **Advantages:**
  - i) In most cases the mode/modes of an ungrouped frequency distribution can be determined simply by observation.
  - ii) Mode is not affected by extreme values
  - iii) Mode is easy to understand
  - iv) Mode can be determined graphically.
- **Limitations:**
  - i) Mode is not based on all the observations
  - ii) It is not suitable for further mathematical treatment.
  - iii) Like arithmetic mean we cannot know the sum of the observations of a distribution if we know the mode and the number of observations of the distribution.

Uses of Mode: The concept of mode is used by manufacturers, businessmen, agriculturists, etc. For example, a manufacture of shoes is interested in the model size of shoes and manufactures them in large quantities. It is useful in industry and business. Weather forecasts are based on mode. The concept of mode is used in socio-economic surveys besides being used in business and commerce.

RELATIONSHIP AMONG MEAN, MEDIAN AND MODE: Prof. Kari Pearson has established an empirical (experimental) relationship among mean, median and mode. For any distribution the following relationship known as empirical relationship approximately holds:

Mean – Mode = 3 (Mean – Median)



CHECK YOUR PROGRESS

Q.1: What do you mean by measures of central tendency or averages?

Q.2: Define weighted arithmetic mean.

Q.3: Find mean and median from the following distribution:

Age (year):	20	19	18	17	16	15	14	13	12	11
No. of students:	1	2	4	8	11	10	7	4	2	1

Q.4: Determine mean, median and mode from the following data:

Weekly wages (Rs.):	15	16	17	18	19	20
No. of workers:	6	12	23	30	9	1

Q.5: Determine mean, median and mode from the following statistical distribution.

Class interval:	6.5-7.5	7.5-8.5	8.5-9.5	9.5-10.5
Frequency:	5	12	25	48
Class interval:	10.5-11.5	11.5-12.5	12.5-13.5	
Frequency:	32	6	1	

Q.6: The weekly expenditures of a few families are given below. Calculate arithmetic mean and median.

Weekly expenditures (in Rs.):	110-120	120-130	130-140	140-150	150-160
No. of families:	6	15	38	62	106
Weekly expenditures (in Rs.):	160-170	170-180	180-190	190-200	
No. of families:	50	18	12	3	

Q.7: Calculate arithmetic mean and median of the distribution given below. Calculate mode using the empirical relation among the three.

Class limit:	130-134	135-139	140-144	145-149
Frequency:	7	12	15	6



3.4 LET US SUM UP

In this unit we have learnt the meaning of measure of central tendency (mean, media and mode). Then we discussed some properties of arithmetic mean. Here we also discussed the relationship between mean, medium and mode.



3.5 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta, S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan, R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill.
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



3.6 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A measure of central tendency or an average of a certain distribution is nothing but a representative value of that distribution which enables us to comprehend in a single effort the significance of the whole.

Ans. to Q. No. 2: When we determine the arithmetic mean of series by assigning weights to the different quantities of the series depending upon their relative importance then this arithmetic mean will be called the weighted arithmetic mean.

Ans. to Q. No. 3: Mean (A.M.) = 15.54 years; Median = 16 years

Ans. to Q. No. 4: Men (A.M.) = Rs. 17.33; Median = Rs. 17; Mode = Rs. 18

Ans. to Q. No. 5: Mean (A.M.) = 9.98; Median = 9.97; Mode = 10.17

Ans. to Q. No. 6: Mean (A.M.) = Rs. 152.64; Median = Rs. 153.67

Ans. to Q. No. 7: 139.5, 139.63, 139.89



3.7 MODEL QUESTIONS

Q.1: Define median of a distribution. Discuss its advantages and limitations.

Q.2: Define mode. What are its advantages and limitations?

Q.3: Write down the empirical relationship among mean, median and mode.

Q.4: What are the properties of A.M?

*** ***** ***

UNIT 4: MEASURES OF DISPERSION

UNIT STRUCTURE

4.1 Learning Objectives

4.2 Introduction

4.3 Dispersion

- 4.3.1 Objectives of Studying Dispersion
- 4.3.2 Essentials of a Good Measure of Dispersion
- 4.3.3 Different Measures of Dispersion

4.4 Let Us Sum Up

4.5 Further Reading

4.6 Answer To Check Your Progress

4.7 Model Questions

4.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the meaning of dispersion or variation
- describe the various methods of obtaining dispersion for various types of distribution
- describe the concept of skewness and types of skewness
- describe the methods of measuring skewness.

4.2 INTRODUCTION

In this unit, we are going to discuss about dispersion or variation. In the previous unit we discussed about one single figure that represents the entire data. Here we shall discuss about the degree to which numerical data tend to spread about an average value. Dispersion measures the extent to which the items vary from central value. Thus through this unit you will also able to explain the concept of skewness and types of skewness.

4.3 DISPERSION

Scatteredness of data around an average is termed as dispersion or variation. To quite Spiegel, "The degree to which numerical data tend to spread about an average value is called variation or dispersion".

4.3.1 Objectives of Studying Dispersion

The objectives of studying Dispersion are as follows:

i) To study the reliability of average.
ii) To control the variation.
iii) To make the comparison among series of observations with regards to variations.
iv) To make further statistical analysis.

4.3.2 Essentials of a Good Measure of Dispersion

i) A measure of dispersion should be based on all the observations of a series.
ii) It should be rigidly defined.
iii) It should be easy to understand and calculate.
iv) It should not extremely affected by extreme values.
v) It should not be affected much by fluctuations of sampling.
vi) It should be usable for further statistical analysis.

4.3.3 Different Measures of Dispersion

The following are the various measures of dispersion: (a) Range (b) Interquartile Range and Quartile Deviation (c) Mean Deviation (d) Standard Deviation and Variance (e) Coefficient of variation. While measures (a), (b), (c) and (d) are called absoluate measures, measure (e) is called a relative measure of dispersion. An absolute measure is that which possesses the same unit of the distribution for which the measure is obtained. On the other hand a relative measure is a ratio or percentage and as such it is a pure number. It is to be mentioned that all measures of central tendency are absolute measures.

a) Range: The range of a distribution is the difference between the largest and the smallest observations of the distribution. Thus if L denotes the largest observation and S denotes the smallest observation of a distribution then the range R of the distribution will be: R = L – S

Example 1: If the marks obtained by six students are 24, 12, 16, 11, 40 and 42, find the range of these marks.

Solution: Here L = 42, S = 11

Range R = L – S = 42 – 11 = 31 (marks)

Example 2: Determine range of the following distribution:

Weights (in kg):	40	47	56	62	70
No. of students:	4	7	11	3	1

Solution: Here L = 70kg, S = 40kg

Range R = L – S = (70 – 40)kg = 30kg

Example 3. The following distribution is a distribution of height. Determine the range of the distribution:

Height (in cm.):	120-129	130-139	140-149	150-159
No. of students:	10	17	23	8

Solution: Range R = (Upper limit of 150-159) – (Lower limit of 120-129) = (159-120) cm = 39 cm.

Note: In case of a grouped frequency distribution with inclusive classification of data like the above distribution, we may also obtain range as below:

R = (Upper class boundary of 150-159) – (Lower class boundary of 120-129) = (159.5 - 119.5) cm = 40cm.

b) Interquartile Range and Quartile Deviation (Q.D.): The interquartile range of a distribution is the difference between the third quartile Q_3 and the first quartile Q_1 of the distribution. Again, half the interquartile range of a distribution is called the quartile deviation (Q.D.) of the distribution. Thus,

Interquartile range = $Q_3 - Q_1$, and Q.D. = $\frac{Q_3 - Q_1}{2}$

Formulas to Calculate Interquartile Range and Quartile Deviation:

For Individual Servies:

Inrerquartile Range = $Q_3 - Q_1$

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

Where Q_1 = size of $\frac{n+1}{4}^{th}$ item

n = number of observations

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

n = number of observations

For Discrete Series:

Interquartile Range = $Q_3 - Q_1$

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

Where Q_1 = Locate the size of $\frac{N+1}{4}^{th}$ item in the cumulative frequency column and corresponding x value is Q_1.

N = Sum of frequencies.

Q_3 = Locate the size of $\left(\frac{3(N+1)}{4}\right)^{th}$ item in the cumulative frequency column and corresponding x value is Q_3.

N = Sum of frequencies.

For Continuous Series:

Interquartile Rang = $Q_3 - Q_1$

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

Here we first locate the size of $\frac{N}{4}^{th}$ item in c.f column and the value of Q_1 will lie in the corresponding class interval. (i.e. Q_1 class)

Where $Q_1 = l_1 + \frac{\frac{N}{4} - c.f.}{f} \times i$

l_1 = lower limit of class interval.

N = Sum of frequencies.

c.f. = Cumulative frequencies of the class preceding the Q_1 class.

i = Class interval.

First, we locate the size of $\frac{3N^{th}}{4}$ item in c.f column and the value of Q_3 will lie in the corresponding class interval.

$$Q_3 = l_1 + \frac{3\left(\frac{N}{4}\right) - c.f.}{f} \times i$$

Where l_1 = lower limit of class interval.

N = Sum of frequencies.

c.f. = Cumulative frequencies of the class preceding the Q_3 class.

i = Class interval.

Example 1: Calculate Range and Q.D. of the following observation:

20 25 29 30 35 39 41 48 51 60 and 70

Solution: Range is 70 – 20 = 50

For Q. D., we need to calculate values of Q_3 and Q_1.

Q_1 is the size of $\frac{N+1^{th}}{4}$ value

Here N being 11. Q_1 is the size of 3rd value.

As the values are already arranged, in ascending order, it can be seen that Q_1, the 3rd value is 29.

Similarly. Q_3 is size of $\frac{3(N+1)^{th}}{4}$ value : i.e. 9th value which is 51

Here Q_3= 51

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{51 - 29}{2} = 11$$

Example 2: For the following distribution of marks scored by a class of 40 students. Calculate the range and Q.D.

Class interval	No of students (f)
0–10	5
10–20	8
20–40	16
40–60	7
60–90	4
	40

Solution: Range is just the difference between the upper limit of the highest class and the lower limit of the lowest class. So, Range is 90 – 0 = 90

For Q.D., first calculate cumulative frequencies as follows:

(Class Interval) (c.i.)	(Frequency) (f.)	(Cumulative Frequency) (c.f.)
0–10	5	05
10–20	8	13
20–40	16	29
40–60	7	36
60–90	4	40
	N = 40	

Q_1 is the size of $\frac{N}{4}^{th}$ value in a continuous series. Thus, it is the size of the 10^{th} value. The class containing the 10^{th} value is 10-20. Here Q_1 lies in class 10-20. Now, to calculate the exact value of Q_1, the following formula is used:

$$Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} \times i$$

Where L = 10 (Lower limit of Quartile class)

c.f. = 5 (value of c.f. for the class preceding the Quartile class)

i = 10 (interval of the Quartile class) and

f = 8 (frequency of the Quartile class)

Thus, $Q_1 = 10 + \frac{10-5}{8} \times 10 = 16.25$

Similarly, Q_3 is the size of $\frac{3N^{th}}{4}$ value i.e., 30^{th} value which lies in class 40 – 60. Now using the formula for Q_3 its value can be calculated as follows:

$$Q_1 = L + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$$= 40 + \frac{30-29}{7} \times 20$$

$$= 42.87$$

$$\therefore Q.D. = \frac{42.87 - 16.25}{2}$$

$$= 13.31$$

Example 3: Calculate Quartile deviation from the following.

Age in years	No. of members
20	3
30	61
40	132
50	153
60	140
70	51
80	3

Solution:

Age in years	No. of members	c.f.
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543

$$Q_1 = \text{Value of } \left(\frac{N+1}{4}\right)^{th} \text{ item}$$

$$= \text{Value of } \left(\frac{543+1}{4}\right)^{th} \text{ item}$$

$$= \text{Value of } \frac{544^{th}}{4} \text{ item}$$

= Value of 136^{th} item

= 40 years

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

$$= \text{Value of } 3\left(\frac{543+1}{4}\right)^{th} \text{ item}$$

= Value of 3×136^{th} item

= 408^{th} item which is 60 years

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

$$= \frac{60-20}{2}$$

$$= \frac{20}{2}$$

= 10 years

Example 4: Calculate the range and Quartile deviation

Wages (Rs.)	Labouress
30–32	12
32–34	18
34–36	16
36–38	14
38–40	12
40–42	8
42–44	6

Solution: Range = L – S

= 44 – 30 = 14 years

x	f	c.f.
30-32	12	12
32-34	18	30
34–36	16	46
36–38	14	60
38–40	12	72
40–42	8	80
42–44	6	86

Q_1 = Size of $\frac{N}{4}^{th}$ item

$= \frac{86}{4} = 21.5$

$\therefore Q_1$, lies in the group 32 – 34

$$Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$= 32 + \frac{21.5 - 12}{18} \times 2$

$= 32 + \frac{19}{18}$

$= 32 + 1.06 = 33.06$

Q_3 = Size of $\frac{3N}{4}^{th}$ item

$= \frac{3 \times 86}{4}$

$= 64.5^{th}$ item

Q_3, lies in the group 38 – 40

$$Q_3 = L + \frac{\frac{3N}{4} - c.f.}{f} \times i$$

$= 38 + \frac{64.5 - 60}{12} \times 2$

$= 38 + 0.75$

$= 38.75$

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

$$= \frac{38.75 - 33.06}{2}$$

$$= \frac{5.69}{2} = 2.85$$

Note: When a distribution is in increasing order then the three quantities which divide the distribution into four equal parts are called the quartiles of the distribution. In increasing order these quartiles are denoted by Q_1, Q_2 and Q_3. Q_1 is called the first or lower quartile Q_2 is called the second or middle quartile and Q_3 is called the third or upper quartile. The second quartile Q_2 and the median (Me) of a distribution are the same.

c) Mean Deviation (M.D.): The arithmetic mean of the absolute deviations of the observations of a distribution from its mean, median or mode is known as mean deviation.

If a variable x takes n values then

$$\text{M.D.} = \frac{|x_1 - A| + |x_2 - A| + ... + |x_n - A|}{n}$$

$$\text{or} \quad \text{M.D.} = \frac{\sum |x - A|}{n} = \frac{\sum |d|}{n}$$

Here A = mean, median or mode of x and d = x – A

Again, if the frequencies of $x_1, x_2, ..., ..., ..., x_n$ are $f_1, f_2, ..., ..., ..., f_n$ respectively then

$$\text{M.D.} = \frac{f_1|x_1 - A| + f_2|x_2 - A| + ... + f_n|x_n - A|}{N}, \quad N = \Sigma f$$

$$\text{or} \quad \text{M.D.} = \frac{\sum f|x - A|}{N} = \frac{\sum f|d|}{N}$$

Here A = mean, median or mode of the distribution, d = x – A

If A = mean, then we get mean deviation from mean. Likewise, we get mean deviation from median and mean deviation from mode. Mean deviation from mean ($\bar{x}$) can be denoted by the symbol M.D. ($\bar{x}$). Likewise, mean deviations from median and mode can be denoted by M.D. (Me) and M.D. (Mo) respectively.

Note:

i) In case of mean deviation from mean, 'A' may be A.M., G.M. or H.M. But usually A is taken as A.M.

ii) $|x - A|$ is called the absolute value of the deviation. By absolute value we mean the magnitude of the value without considering the sign. Thus $x - A$ may be positive or negative but is $|x - A|$ always positive. Thus $|5 - 2| = |3| = 3$, $|7 - 9| = |-2| = 2$ etc.

iii) Since the sum of the deviations measured from the mean is zero, hence in case of mean deviation we always take absolute deviations.

Advantages and Limitations of Mean Deviations:

Advantages:

i) It is based on all the observations.

ii) It is less affected by extreme values in comparison to standard deviation.

iii) Since deviations are taken from average (mean, median or mode), therefore mean deviation is considered to be a good measure for comparing the variability among two or more distributions.

Limitations:

i) In mean deviation actual signs of deviations are discarded by taking absolute values of the deviations.

ii) Mean deviation from mode is not considered to be a good measure of dispersion.

iii) One cannot determine mean deviation for a grouped frequency distribution containing open-end class interval.

Example 5: For the following distribution determine mean deviation (M.D.) from mean:

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

Solution: First of all, we form the following table.

x	f	fx	$\|d\| = \|x - \bar{x}\|$	f|d|
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
	N = 48	$\Sigma fx = 576$		$\Sigma f\|d\| = 36$

A.M. $(\bar{x}) = \frac{\sum fx}{N} = \frac{576}{48} = 12$

(Usually mean implies arithmetic mean.)

Now, M.D. from mean = $\frac{\sum f|d|}{N} = \frac{36}{48} = 0.75$

d) Standard Deviation (S.D.): The positive square root of the arithmetic mean of the squares of the deviations of the values of a variable from its arithmetic mean is called the standard deviation of that variable. The symbol σ (a Greek letter pronounced as 'sigma') is used to denote standard deviation.

If a variable x takes n values $x_1, x_2, ..., ..., ..., x_n$ and if $\bar{x}$ be the arithmetic mean of these values, then

$$\sigma = \sqrt{\frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + ... + f_n(x_n - \bar{x})^2}{n}}$$

i.e. $\sigma = \sqrt{\frac{\sum(x_1 - \bar{x})^2}{n}}$, $\bar{x} = \frac{\sum x}{n}$

Again, in case of frequency distribution:

x	x_1	x_2	.....	x_n
f	f_1	f_2	.....	f_n

Standard deviation will be:

$$\sigma = \sqrt{\frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + ... + f_n(x_n - \bar{x})^2}{N}}, \quad N = \Sigma f$$

i.e. $\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N}}$, $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$

Instead of the symbol σ we may use the symbol σ_x to clearly signify the standard deviation of the variable x.

Alternative methods of obtaining standard deviation:

a) Assumed Mean Method: In case of individual series we have the following formula:

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}, \quad d = x - A$$

And A = Assumed mean of x.

In case of frequency distribution,

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

b) Step deviation method: If the deviations from assumed mean A denoted by d have some common factors then these deviations are divided by the highest common factor h (say) and the step deviations denoted by d' are obtained. Thus $d' = d/h$, where h is the H.C.F. of the deviations d. In this method, we have the following formula:

For individual series, $\sigma = \sqrt{\frac{\sum f'^2}{n} - \left(\frac{\sum f'}{n}\right)^2} \times h$

For frequency distribution $\sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times h$

$d' = \frac{d}{h} = \frac{x - A}{h}$, A = Assumed mean of the variable x whose values are the various values of the distribution.

Empirical relationship among Q.D., M.D. and S.D.: It can be verified empirically that for a distribution the following relationship among Q.D., M.D. and S.D. approximately holds. However, in case of normal distribution, this relationship is exact.

$$\text{Q.D.} = \frac{5}{6}, \text{ M.D.} = \frac{2}{3}\text{S.D.}$$

or, 6Q.D. = 5 M.D. = 4 S.D.

Advantages and Limitations of S.D.: Among all the measures of dispersion standard deviation is considered to be the best measure. Standard deviation possesses almost all the qualities that a good measure of dispersion should have. It is rigidly defined. It is based on all the observations. The sampling stability of this measure is the maximum in comparison to all other measures of dispersion. While mean deviation ignores the algebraic signs of deviations, standard deviation is free from this demerit as in standard deviations squares of the deviations are taken. The formula for standard deviation is amenable for further mathematical treatment. Normal curve can be analysed with the help of standard deviation. Standard deviation plays a significant role in sampling theory and correlation analysis.

The main limitation of standard deviation is that it is difficult to calculate in comparison to other measures of dispersion. Moreover, in comparison to mean deviation it is more affected by extreme values. If we make a comparative study of the advantages and limitations of various measures of dispersion, we find that standard deviation is the best of all the measures of dispersion. Hence it is widely used as a measure of dispersion. It is called an ideal measure of dispersion.

An Important Property of Standard Deviation (S.D.): The S.D. of a series remains unaltered if a constant is added to or subtracted from each value of the series. For example, if the standard deviation of the series $x_1, x_2, ..., ..., ..., x_n$ is σ then the S.D. of the series $x_1 \pm c, x_2 \pm c, ..., ..., ..., x_n \pm c$, will also be where c ($c \neq 0$) is a constant. On the other hand, if each value of a series is multiplied by a constant c then the S.D. of the new series will be equal to c times the S.D. of the original series. Again, if each value of a series is divided by a constant c then the S.D. of the new series will be equal to the S.D. of the original series divided by c.

e) Coefficient of Variation (C.V.): It is a relative measure of dispersion. It is expressed as a percentage and is useful in comparing the variability of two or more sets of data especially when they are expressed in different units of measurement. According to Karl Pearson, ''Coefficient of variation is the percentage variation in mean, standard deviation being considered as the total variation in the mean.'' Thus coefficient of variation abbreviated as C.V. is defined as below:

$$C.V. = \frac{\sigma}{\bar{x}} \times 100\%$$

Although we have several relative measures of dispersion yet the coefficient of variation is the most commonly used relative measure of dispersion. Other relative measures are rarely used in practice. While comparing the variability between distributions, the distribution having minimum C.V. is said to be less variable, more stable, more uniform, more consistent or more homogeneous. On the other hand, the other distribution is said to be more variable, less stable, less uniform, less consistent or less homogeneous.

Note: Suppose the scores of a cricket player A in three matches are 0, 10 and 80. And the scores of another cricket player B in these matches be 28, 30 and 32. Although the mean scores of both the players are same both being 30, we say that B is more consistent then A. Even people without having the knowledge of variability or dispersion say like this. If we calculate the C.V. for both the series of scores we shall find that the C.V. of scores of B is less than the C.V. of scores of A. In fact, the magnitude of any measure of dispersion will be less in case of scores attained by B than the scores attained by A.

Example 6: Find standard deviation of the following observations: 8,10, 12, 14, 16, 18, 20, 22, 24, 26

Solution: We shall solve this problem by Direct Method as well as by Step-Deviation Method.

Direct Method:

Value x	Deviation from the mean $x = \bar{x}$, ($\bar{x} = 17$)	$(x-\bar{x})^2$
8	–9	81
10	–7	49
12	–5	25
14	–3	9
16	–1	1
18	1	1
20	3	9
22	5	25
24	7	49
26	9	81
$\Sigma x = 170$		$\Sigma(x-\bar{x})^2 = 330$

$$\bar{x} = \frac{\sum x}{n} = \frac{170}{10} = 17$$

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

$$= \sqrt{\frac{330}{10}}$$

$$= \sqrt{33}$$

$$= 5.74$$

Note: When the mean $\bar{x}$ is an integer the formula $\sigma = \sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n}}$ should be used to find S.D. When $\bar{x}$ is not an integer then the $\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$ or $\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$ formula should be used. Because when $\bar{x}$ is not an integer determination of the values of $(x - \bar{x})^2$ becomes vary time consuming.

Step-Deviation Method: Let the given values be the values of the variable x and let the assumed mean of x be 18.

Calculation of S.D.:

Value	d = x – 18	$d' = \frac{d}{2}$	d'^2
8	–10	–5	25
10	–8	–4	16
12	–6	–3	9
14	–4	–2	4
16	–2	–1	1
18	0	0	0
20	2	1	1
22	4	2	4
24	6	3	9
26	8	4	16
Total		$\Sigma d' = -5$	$\Sigma d'^2 = 85$

$$\because \sigma = \sqrt{\frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2} \times h$$

$$= \sqrt{\frac{85}{10} - \left(\frac{-5}{10}\right)^2} \times 2$$

$$= \sqrt{8.50 - 0.25} \times 2$$

$$= \sqrt{8.25} \times 2$$

$$= 2.87 \times 2$$

$$= 5.74$$

Note: If we use assumed mean method then we will have to apply the following formula:

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

But if the values of d have a common factor then we shall have to divide the values of d by that common factor and the values obtained thereby are to be assumed to be the values of the variable d'. Then we get the formula for step-deviation method. If the values of d have two or more common factors

then we are to divide the values of d by its H.C.F. and the values obtained thereby will be considered to be the values of d. If the values of d have only one common factor then that will be the highest common factor.

Example 7: Find S.D. from the following record of number of car accidents in a street.

No. of accidents (x):	1	2	4	5	6
No. of days:	2	3	3	1	1

Solution: We shall offer the solution by the Assumed Mean Method.

x	f	d = x – 4	fd	fd^2
1	2	–3	–6	18
2	3	–2	–6	–12
4	3	0	0	0
5	1	1	1	1
6	1	2	2	4
	N = 10		$\sqrt{\sum fd'}$ = –9	$\sqrt{\sum fd'^2}$ = 35

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

$$= \sqrt{\frac{35}{10} - \left(\frac{-9}{10}\right)^2}$$

$$= \sqrt{3.5 - 0.81}$$

$$= \sqrt{2.69}$$

$$= 1.64$$

Example 8: Calculate mean and standard deviation from the following distribution:

Age (year):	20-25	25-30	30-35	35-40	40-45	45-50
No. of Persons:	170	110	80	45	40	35

Solution:

Age (year)	Mid value (x)	No. of persons (f)	d = x – 32.5	$d' = \frac{d}{2}$	fd'	fd'^2
20-25	22.5	170	–10	–2	–340	680
25-30	22.7	110	–5	–1	–110	110
30-35	32.5	80	0	0	0	0
35-40	37.5	45	5	1	45	45
40-45	42.5	49	10	2	80	160
45-50	47.5	35	15	3	105	315
		N = 480		$\Sigma fd'$ = –220	$\Sigma fd'^2$ =1310	

Here, assumed mean A = 32.5

Now, A.M.$(\bar{x})$ = $A + \frac{\Sigma fd'}{N} \times h$

$= 32.5 + \frac{-220}{480} \times 5$

$= 32.5 - 2.99$

$= 29.51$

The required A.M. = 29.51 years.

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

$= \sqrt{\frac{1310}{480} - \left(\frac{220}{480}\right)^2} \times 5$

$= \sqrt{2.7292 - 0.2101} \times 5$

$= \sqrt{2.5191} \times 5$

$= 1.5871 \times 5$

$= 7.936$

∴ The required S.D. = 7.936 years.

Example 9: Calculate mean and standard deviation from the following data:

Value (Rs.):	90-99	80-89	70-79	60-69	50-59	40-49	30-39
No. of Persons:	2	12	22	20	14	4	1

Solution:

Determine of A.M. and S.D.:

Class interval	Mid value (x)	Frequency (f)	d = x – 64.5	$d' = \frac{d}{10}$	fd'	fd'^2
90-99	94.5	2	30	3	6	18
80-89	84.5	12	20	2	24	48
70-79	74.5	22	10	1	22	22
60-69	64.5	20	0	0	0	0
50-59	54.5	14	–10	–1	–14	14
40-49	44.5	4	–20	–2	–8	16
30-39	34.5	1	–30	–3	–3	9
	N = 75				$\Sigma fd'$ = 27	$\Sigma fd'^2$ = 127

Now, A.M. $(\bar{x}) = A + \frac{\Sigma fd'}{N} \times h$

$= 64.7 + \frac{27}{75} \times 10$

= 64.5 + 3.6

= 68.10

∴ The required A.M. = Rs. 68.10

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

$= \sqrt{\frac{127}{75} - \left(\frac{27}{75}\right)^2} \times 10$

$= \sqrt{1.6933 - 0.1296} \times 10$

$= \sqrt{1.5637} \times 10$

= 1.2505 x 10

= 12.505

= 12.51

∴ The required S.D. = Rs. 12.51

Example 10: Calculate standard deviation and coefficient of variation from the following data:

Age (year):	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Persons:	3	61	132	153	140	51	2

Solution:

Calculation of S.D. and C.V.:

Age	Mid value (x)	Frequency (f)	d = x–55	$d' = \frac{d}{10}$	fd'	fd'^2
20-30	25	3	–30	–3	–9	27
30-40	35	61	–20	–2	–122	244
40-50	45	132	–10	–1	–132	132
50-60	55	153	0	0	0	0
60-70	65	140	10	1	140	140
70-80	75	51	20	2	102	204
80-90	85	2	30	3	6	18
N = 542					$\Sigma fd' = -15$	$\Sigma fd'^2 = 765$

$$\text{S.D. } (\sigma) = \sqrt{\frac{765}{542} - \left(\frac{-15}{542}\right)^2} \times 10$$

$$= \sqrt{1.4114 - 0.0001} \times 10$$

$$= \sqrt{1.4105} \times 10$$

$$= 11.876$$

i.e. S.D. = **11.876 years.**

$$\text{Again, C.V.} = \frac{\sigma}{\bar{x}} \times 100\%$$

$$\text{Now, } \bar{x} = A + \frac{\Sigma fd'}{N} \times h$$

$$= 55 + \frac{-15}{542} \times 10$$

$$= 55 - 0.277 = 54.723$$

i.e. A.M. = $\bar{x}$ = 54.723 years

$$\therefore \text{ C.V.} = \frac{11.876}{54.723} \times 100\% = \textbf{21.76\%}$$



CHECK YOUR PROGRESS

Q.1: Determine standard deviation and coefficient of variation for the following distribution:

Age (year):	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Persons:	6	42	86	122	78	30	5

Q.2: Below are given yearly profits of a few small companies. Determine arithmetic mean and standard deviation:

Profit (in thousand Rs.):	20-30	30-40	40-50	50-60	60-70
No. of Companies:	30	58	62	85	112
Profit (in thousand Rs.):	70-80	80-90	90-100		
No. of Companies:	112	57	26		

Q.3: The following distribution is the distribution of ages of 100 students. Find the standard deviation of the distribution.

Age (in years):	16-17	17-18	18-19	19-20	20-21
No. of Students:	4	14	18	28	20
Age (in years):	21-22	22-23			
No. of Students:	12	4			

Q.4: i) What is the S.D. of the following series:

5, 5, 5, 5, 5, 5, 5

ii) Suppose the S.D. of the series a_1, a_2, ..., ..., ..., a_n is σ. What is the S.D. of the series $a_1 + 5$, $a_2 + 5$, ..., ..., ..., $a_n + 5$?

iii) What is the relationship between the S.D.s of the following two series?

10, 20, 30, 70, 100, 110

And 1, 2, 3, 7, 10, 11

Q.5: Chose the correct answer:

i) Which of the following is a unitless measure of dispersion?

a) Standard derivation b) Mean deviation

c) Coefficient of variation d) Range

ii) The relationship between mean deviation and standard deviation is:

a) 3 M.D. = 2 S.D. b) 6 M.D. = 5 S.D.

c) 5 M.D. = 4 S.D. d) M.D. = S.D.

iii) If the minimum value in a series is 20 and its range is 47, the maximum value of the series is:

a) 67 b) 57

c) 48 d) None of the above.



4.4 LET US SUM UP

- In this unit we have learnt the meaning of measure of dispersion.
- Then we have discussed different measures of range, interquartile range and quartile deviation, mean deviation, standard deviation.
- The unit also encompass the concept of variance and coefficient of variation.



4.5 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.

2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.

3) Rajagopalan, S. P. & Sattanthan R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill

4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.

5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



4.6 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: S.D. = 119, C.V. = 21.74;

Ans. to Q. No. 2: A.M. = Rs. 63.32, S.D. = Rs. 18.69;

Ans. to Q. No. 3: 1.47 years

Ans. to Q. No. 4: (i) 0; (ii) σ; (iii) S.D. of the 2nd series = $\frac{1}{10}$ x S.D. of the first series

Ans. to Q. No. 5: (i) (c), (ii) (c), (iii) (a).



4.7 MODEL QUESTIONS

Q.1: What do you mean by dispersion? What are the various measures of dispersion?

Q.2: Distinguish between absolute and relative measures of dispersion.

Q.3: Define standard deviation. Why is standard deviation most widely used as a measure of dispersion?

Q.4: Write a note on coefficient of variation.

Class:	0-10	10-20	20-30	30-40	40-50
Frequency:	6	12	22	48	56
Class:	50-60	60-70	70-80		
Frequency:	32	18	6		

*** ***** ***

UNIT 5 : SKEWNESS, MOMENTS AND KURTOSIS

UNIT STRUCTURE

5.1 Learning Objectives

5.2 Introduction

5.3 Skewness

5.4 Measures of Skewness

- 5.4.1 Karl Pearson's Co-efficient of Skewness
- 5.4.2 Boweley's Co-efficient of Skewness

5.5 Moments

- 5.5.1 Moments About Mean
- 5.5.2 Moments About Arbitrary Point A
- 5.5.3 Relation Between Central and Raw Moments

5.6 Karl Pearson's Beta (β) and Gamma (γ) Coefficient

- 5.6.1 Coefficient of Skewness Based on Moments

5.7 Kurtosis

- 5.7.1 Measures of Kurtosis

5.8 Let Us Sum Up

5.9 Further Reading

5.10 Answers To Check Your Progress

5.11 Model Questions

5.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- Know the complementary relationship of skewness with measures of central tendency and dispersion in describing a set of data.
- understand 'moments' as a convenient and unifying method for summerizing several descriptive statistical measures.

5.2 INTRODUCTION

In order to make proper comparison between two or more distributions and to reveal clearly the salient features of a frequency distribution, we need

to study certain statistical measures.The measures of central tendency tells us about the concentration of the observations about the middle of the distribution and the measure of dispersion give us an idea about the speead and scatter of the observation about some measure of central tendency. We may come across frequency distributions which differ very widely in their nature and composition and yet may have the same central tendency and dispersion.

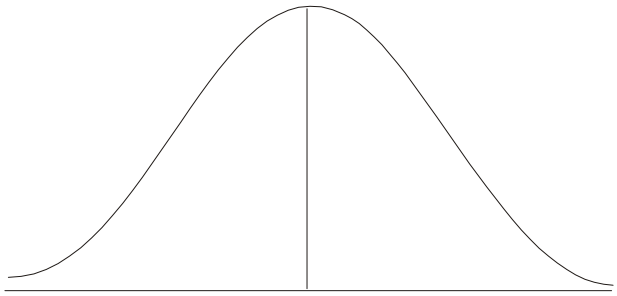
Thus these two measures viz. central tendency and dispersion are inadequate to characterise a distribution completely and these must be supported and supplemented by two more measures viz skewness and Kurtosis. **Skewness** helps us to study the shape i.e. symmetry or asymmetry of the distribution. While **Kurtosis** refers to the flatness or peakedness of the curve which may be drawn with the help of the given data. These four measures viz central tendency, dispersion, skewness and kurtosis are sufficient to describe a frequency distribution completely.

5.3 SKEWNESS

Literal meaning of skewness is lack of symmetry. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given frequency distribution. It helps us to determine the nature and extent of the concentration of the observations towards the higher or lower values of the variable. In a symmetrical frequency distribution, which is unimodal, if the frequency curve is folded about the ordinate at the mean, the two halves so obtained will coincide with each other.

A frequency distribution, which is not symmetrical is called **asymmetrical** or **skewed**. In a skewed distribution, extreme values in a data set move towards one side or tail of the distribution, thereby lergthening that tail. A frequency distribution. For which the curve has a longer tail towards the right is said to be **positively skewed** (Figure 5.2) and if the longer tail lies towards the left, it is said to be negatively skewed (Figure 5.3).

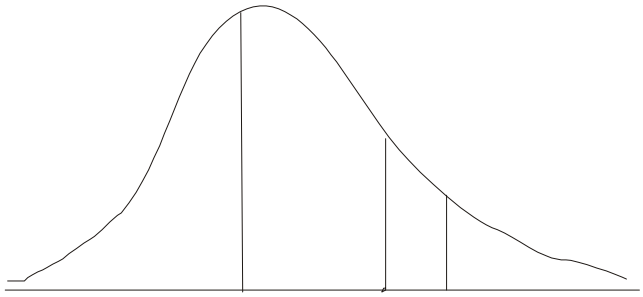
Note : 1) For a symmetrical distribution, a Bell-shaped (normal curve) is obtained for which Mean = Median = Mode i.e. these three values fall at the same point. (Figure 5.1)



Mean = Mode = Median

(Fig. 5.1)

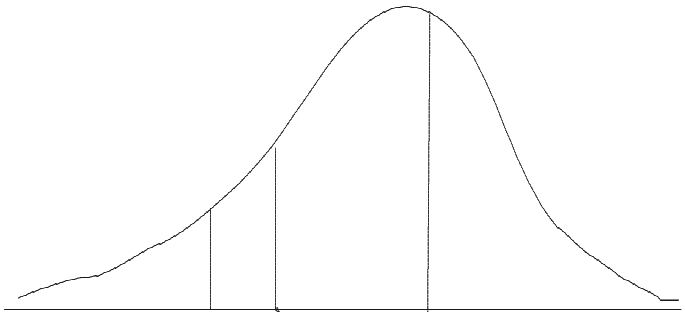
(Symmetrical Distribution)



Mode = Median = Mean

(Fig. 5.2)

(Positively Skewed Distribution)



$\bar{x}$ Median Mode

(Fig. 5.3)

(Negatively Skewed Distribution)

2) For a positively skewed distribution A.M > Median > Mode, and for a negatively skewed distribution AM < Median < Mode.

3) For a skewed distribution, Quartiles Q_1 and Q_3 are not equidistant from the Median i.e.

Q_3 – Median # Median – Q_1

7.4 MEASURES OF SKEWNESS

For a skewed distribution, the difference between AM and Mode i.e (AM – Mode) can also be taken as a measure of skewness as for a positively skewed distribution AM > Mode and for a negatively skewed distribution AM < Mode. But the statistician observed that this measure may not be desirable as the diffence between mean and mode is expressed in the same unit as the distribution and therefore can not be used for comparing skewness of two or more distributions having different units of measurement.

In order to overcome this shortcoming, relative measure of skewness (free from units of measurement), also commonly known as Co-efficient of skewness which is pure number suggested by Karl pearson is preferred for measuring skewness.

5.4.1 Karl Pearson's Co-efficient of Skewness

Karl Pearson's Co-efficient of Skewness is given by the Formula:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{\bar{x} - M_o}{\sigma}$$

But quite often, Mode is ill-defined. In such a situation, we use the empieical relation between Mean, Mode and Median for moderately skewed distribution: Mean – Mode = 3 (Mean – Median)

$$\text{In that case, } Sk_p = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

Remarks:

1) Theoretically, Karl Pearson's Co-efficient of Skewness lies between ± 3. But in practice these limits are rasely attained. For a moderately skewed distribution, SK_p lies between ± 1.
2) For a symmetrical distribution Mean = Mode = Median. Therefore, $SK_p = 0$. In other words, for a symmetrical distribution $SK_p = 0$.
3) SK > 0 if $\bar{x}$ > Med > Mo

   $\therefore$ For a positively skewed distribution, the value of Mean is the greatest of the three measures and value of Mode is the leasl. On the other hand, SK < 0 if Mo > Med > $\bar{x}$

   $\therefore$ For a negatively skewed distribution the value of Mode is the greatest of the three measures and value of Mean is the leasl.

5.4.2 Boweley's Co-efficient of Skewness

There is another method suggested by Prof. Bowely known as **Bowleys'** Co-efficient of Skewness. It is given by

$$Sk_B = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

Remarks:

1) The values of Sk_3 lies between ± 1.
2) This method of measuring skewness is quite useful when the distribution has open end classes.

Note: It is important to note that–

1) These two measures of Skewness namely Karl Pearson's and Bowley's co-efficient of Skewness can not be compared. On certain occasions it is possible that one of them gives a possible that one of them gives a positive value while the other gives a negative value.
2) Out of the two measures, SK_p is more reliable due to the fact that sometime it may happen that $SK_B = 0$ but the distribution is not perfectly symmetrical.

Example 1: Given for a distribution, pearson's measure of skewness is 0.4, AM = 30 and standard deviation = 8. Find Median and Mode.

Solution: Given, $\bar{x} = 30$, $\sigma = 8$ and $SK_p = 0.4$

We know that, $SK_p = \frac{\bar{x} - \text{Mode}}{\sigma}$

$$\Rightarrow 0.4 = \frac{30 - \text{Mode}}{8}$$

$$\Rightarrow 3.2 = 30 - \text{Mode}$$

$$\Rightarrow \text{Mode} = 30 - 3.2 = 26.8$$

Also, $SK_p = \frac{3(\bar{x} - \text{Median})}{\sigma}$

$$\Rightarrow 0.4 = \frac{3(30 - \text{Median})}{8}$$

$$\Rightarrow 3.2 = 90 - 3 \text{ Median}$$

$$\Rightarrow 3 \text{ Median} = 86.8$$

$$\Rightarrow \text{Median} = 28.93$$

Example 2: From the data given below, calculate the co-efficient of variation–

AM = 86, Median = 80, Pearsonian measure of skewness = 0.42.

Solution: Given $\bar{x}$ = 86, Median = 80, SK_p = 0.42.

We know that, $SK_p = \frac{3(\bar{x} - \text{Median})}{\sigma}$

$$\Rightarrow 0.42 = \frac{3(86-80)}{\sigma}$$

$$\Rightarrow \sigma = \frac{18}{0.42} = 42.86$$

$\therefore$ Co-efficient of variation = $\frac{\sigma}{\bar{x}} \times 100$

$$= \frac{42.86}{86} \times 100$$

$$= 49.84\%$$

Example 3: Given below are the Arithmetic mean, the Median and Standard Deviation of two distributions, Determine which distribution is more skewed.

i) AM = 22, Median = 24, SD = 10

ii) AM = 22, Median = 25, SD = 12

Solution: Here Pearson's Co-efficient of Skewness (SK_p) is applicable.

$$SK_p = \frac{3(\text{AM} - \text{Median})}{\text{SD.}}$$

For distribution (i), $SK_p = \frac{3(22-24)}{10} = -0.6$

For distribution (ii), $SK_p = \frac{3(22-25)}{12} = -0.75$

$\therefore$ For (i) $|SK_p| = 0.6$

For (ii) $|SK_p| = 0.75$

Since, $|SK_p|$ for (ii) > $|SK_p|$ for (i), hence distribution (ii) is more skewed.

Example 4: For a group 10 observations, sum of the observations is 452 and sum of squares of the observations is 24,270 and Mode is 43.7. Find the Pearson's co-efficient of Skewness.

Solution: Given n = 10, $\Sigma x = 452$, $\Sigma x^2 = 452$, $\Sigma x^2 = 24{,}270$ and Mode = 43.7

$\therefore$ Mean $(\bar{x}) = \frac{\Sigma x}{n} = \frac{452}{10} = 45.2$

$$S.D.(\sigma) = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{24270}{10} - (45.2)^2} = 19.59$$

$$\therefore\ SK_p = \frac{\bar{x} - Mode}{\sigma} = \frac{45.2 - 43.7}{19.59} = -0.08$$

Example 5: In a frequency distribution, the co-efficient of Skewness lased on quartiles is 0.6. If the sum of upper and lower quaitiles is 100 and Median is 38, find Q_3.

Solution : Given $SK_B = 0.6$, $Q_3 + Q_1 = 100$, Median = 38

We know that, $SK_B = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$

$$\Rightarrow 0.6 = \frac{100 - 76}{Q_3 - Q_1}$$

$$\Rightarrow Q_3 - Q_1 = \frac{24}{0.6} = 40$$

Given $Q_3 + Q_1 = 100$

$\therefore\ (Q_3 - Q_1) + (Q_3 + Q_1) = 40 + 100$

$\Rightarrow\ 2Q_3 = 140$

$\Rightarrow\ Q_3 = 70$

Example 6: Calculate Karl Pearson's Co-efficient of Skewness from the data given below.

Wt (Kg):	10	20	30	40	50	60	70
Frequency:	1	5	12	22	17	9	4

Solution: Table for calculation of co-efficient of Skewness

Wt (kg) (x)	Frequency (f)	d = x – A	$d' = d/10$	fd'	fd'^2
10	1	–30	–3	–3	9
20	5	–20	–2	–10	20
30	12	–10	–1	–12	12
40	22	0	0	0	0
50	17	10	1	17	17
60	9	20	2	18	36
70	4	30	3	12	36
	70 = N			$\Sigma fd' = 22$	$\Sigma fd'^2 = 130$

Let assumed mean = A = 40

Here i = 10

$$\therefore \bar{x} = A + \frac{\Sigma fd'}{N} \times i$$

$$= 40 + \frac{22}{70} \times 10$$

= 43.14 kg.

Mode = 40 (By inspection)

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

$$= \sqrt{\frac{130}{70} - \left(\frac{22}{70}\right)^2} \times 10 = 13.27$$

$$\therefore \ SK_P = \frac{\bar{x} - Mode}{\sigma} = \frac{43.1 - 40}{13.26} = 0.23$$

Example 7: From the data given below, calculate Karl Pearson's co-efficient of Skewness.

Ht (cm):	130-134	135-139	140-144	145-149	150-154	155-159	160-164
f:	3	12	21	28	19	12	5

Solution: Table for calculating Karl Pearson's co-efficient of Skewness.

Ht (cm) value(x)	(f)	Mid	d = x–A	$d'=d/5$	fd^1 Boundarie	fd^{12}	Class
130-134	3	132	–15	–3	–9	27	129.5–134.5
135–139	12	137	–10	–2	–24	48	134.5–139.5
140–144	21	142	–5	–1	–21	21	139.5–144.5
145–149	28	14.7	0	0	0	0	144.5–154.5
150–154	19	152	5	1	19	19	149.5–154.5
155–159	12	157	10	2	24	48	154.5–159.5
160–164	5	162	15	3	15	45	159.5–164.5
	100 = N				4 = $\Sigma fd'$	208 = Σfd^{12}	

Let A = 147 Here i = 5

$$\therefore \bar{x} = A + \frac{\Sigma fd'}{N} \times i = 147 + \frac{4 \times 5}{100} = 147.2 \text{ cm}$$

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

$$= \sqrt{\frac{208}{100} - \left(\frac{4}{100}\right)^2} \times 5$$

$$= \sqrt{2.08 - .0016} \times 5 = 7.21 \text{ cm}$$

Here Modal Class is 144.5 – 149.5

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

$$= 144.5 + \frac{28 - 21}{56 - 21 - 19} \times 5$$

$$= 144.5 + \frac{35}{16} = 146.69 \text{ cm}$$

$$\therefore SK_P = \frac{\bar{x} - \text{Mode}}{\sigma} = \frac{147.2 - 146.69}{7.21} = 0.07$$

Example 8: Fine the most suitable measure of Skewness from the data given below.

Age (yrs):	Below 20	20–25	25–30	30–35	35–40	40–45	45–50
No. of employee 50 and above 21	13	29	46	60	112	94	45

Solution: This is an opended series. So. the most suitable measure of Skewness is Bowley's Co-efficient of Skewness.

Age (yrs)	No. employee (f)	c.f.
Below 20	13	13
20 – 25	29	42
25 – 30	46	88
30 – 35	60	148
35 – 40	112	260
40 – 45	94	354
45 – 50	45	399
50 and above	21	420
	420 = N	

We know that, Bowley's coefficiant of Skewness

$$= SK_B = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Here N = 420

$\therefore$ Q_1 = Value of $\left(\frac{N}{4}\right)^{th} = \frac{420}{4} = 105^{th}$ item

Median = Value of $\left(\frac{N}{2}\right)^{th} = \frac{420}{2} = 210^{th}$ item

Q_3 = Value of $\left(\frac{3N}{4}\right)^{th} = \frac{3 \times 420}{4} = 315^{th}$ item

$\therefore$ Q_1 class is 30 – 35

Median class is 35 – 40

Q_3 class is 40 – 45

$$\therefore Q_1 = l + \frac{N/4 - cf}{f} \times i$$

$$= 30 + \frac{105 - 88}{60} \times 5$$

= 30 + 1.42

= 31.42 yrs

$$\text{Median} = l + \frac{N/2 - cf}{f} \times i$$

$$= 35 + \frac{210 - 148}{112} \times 5$$

= 34 + 2.77

= 37.77 yrs

$$\text{and } Q_3 = l + \frac{3N/4 - cf}{f} \times i$$

$$= 40 + \frac{315 - 260}{94} \times 5$$

= 42 + 2.93 = 42.93 yrs

$$\therefore SK_B = \frac{42.93 + 31.42 - 2 \times 37.77}{42.93 - 31.42}$$

$$= \frac{-1.19}{11.51} = -0.103$$

5.5 MOMENTS

In statistics, 'Moments' are used to describe the various characteristics of a frequency distribution viz central tendency, dispersion, Skwness and Kurtosis.

5.5.1 Moments About Mean

Let a variable x takes the values $x_1, x_2, ..., ..., x_n$. Then $\bar{x} = \frac{\Sigma x}{n}$ is its arithmetic mean.

The r^{th} moment of x about the mean $\bar{x}$, is usually denoted by μr [where μ is the letter (mu) of the Greek alphabet] is defined as

$$\mu_r = \frac{1}{n}\Sigma(x - \bar{x})^r \quad r = 1, 2, 3, 4 \quad 5.5.1 \text{ (a)}$$

Similarly, for a frequency distribution of a variable x

x	x_1 x_2..... x_n	
y	f_1 f_2.....f_n	$\Sigma f = N$

Then, $\bar{x} = \frac{\Sigma fx}{N}$ is the arithmetic mean

Then r^{th} moment about $\bar{x}$ is defined as

$$\mu_r = \frac{1}{N}\Sigma f(x - \bar{x})^r, \quad r = 1, 2, 3, 4 \quad 5.5.1 \text{ (b)}$$

Putting r = 1 in 5.5.1 (a) and 5.5.1 (b) we get

$$\mu_1 = \frac{1}{n}\Sigma(x - \bar{x})^r = 0 \quad \rightarrow \text{For individual series}$$

$$\mu_1 = \frac{1}{N}\Sigma f(x - \bar{x})^r = 0 \quad \rightarrow \text{For frequency distribution}$$

[Because, the algebric sum of deviations of a given set of observations from their AM is zero]

Thus, the first moment about mean is always zero.

Putting r = 2.

$$\mu_2 = \frac{1}{n}\Sigma(x - \bar{x})^2 = \sigma x^2 \quad \rightarrow \text{For individual series.}$$

$$\mu_2 = \frac{1}{N}\Sigma f(x - \bar{x})^2 = \sigma x^2 \quad \rightarrow \text{For frequency distribution.}$$

Thus, the second moment about mean is the variance of the distribution.

Also, $\left.\begin{aligned}\mu_3 &= \frac{1}{n}\Sigma(x-\bar{x})^3 \\ \mu_4 &= \frac{1}{n}\Sigma(x-\bar{x})^4\end{aligned}\right\}$ For individual series

and $\left.\begin{aligned}\mu_3 &= \frac{1}{N}\Sigma f(x-\bar{x})^3 \\ \mu_4 &= \frac{1}{N}\Sigma f(x-\bar{x})^4\end{aligned}\right\}$ For frequency distribution

5.5.2 Moments About Arbitrary Point A

When actual mean is in fractions, moments are first calculated about an assumed mean A (arbitrary point) and then are converted about the actual mean.

The r^{th} moment of a variable x about an arbitrary point A is given by

$$\mu'_r = \frac{\Sigma(x-A)^r}{n} \rightarrow \text{For individual series}$$

and $\mu'_r = \frac{\Sigma f(x-A)^r}{N} \rightarrow$ For frequency distribution.

Now, $\mu'_1 = \frac{1}{n}\Sigma(x-A)^2 = \frac{1}{n}\Sigma x - \frac{1}{n}\Sigma A$

$$= \bar{x} - \frac{1}{n}\times nA = \bar{x} - A$$

Also, $\mu'_1 = \frac{1}{N}\Sigma f(x-A)$

$$= \frac{1}{N}\Sigma fz - \frac{1}{N}\Sigma fA$$

$$= \bar{x} - \frac{A}{N}\Sigma f = \bar{x} - \frac{A}{N}\times N = \bar{x} - A$$

$\therefore$ For both individual series and frequency distribution.

$$\boxed{\mu'_1 = \bar{x} - A}$$

Note : 1) μ_r , the r^{th} moment about the mean; r = 1, 2, 3, 4, ..., ... are also called **Central** moments and μ'_r , the r^{th} moment about any arbitrary point A are called **raw moments** or **non-central moments.**

2) The moments (raw or central) of higher order other than fourth order are seldom used.

5.5.3 Relation Between Central and Raw Moments

$$\text{Central Moment } = \mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$$

Raw Moments about an arbitrary point A

$$\mu'_r = \frac{1}{n}\Sigma(x-A)^r$$

Now, $\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$

$$= \frac{1}{n}\Sigma\{x-A-(\bar{x}-A)\}^r$$

$$= \frac{1}{n}\Sigma\left(x-A-\mu'_1\right)^r$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\,\mu'_1\Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\,\Sigma(x-A)^{r-2} + \ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\,\mu'_1\frac{1}{n}\Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\,\Sigma(x-A)^{r-2} + \ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \mu'_r - {}^rC_1\,\mu'_1\mu'_{r-1} + {}^rC_2\,(\mu'_1)^2\mu'_{r-2} + \ldots\ldots + (-1)^r(\mu'_1)^r$$

From the above relation, for various values of r we have

For r = 1, $\mu_1 = \mu'_1$

For r = 2, $\mu_2 = \mu_2' - {}^2C_1\mu'^2{}_1 + {}^2C_2\,\mu_1'^2\mu_0'$

$$\Rightarrow \boxed{\mu_2 = \mu_2' - (\mu_1')^2}$$

$$\mu_0 = \frac{1}{n}\Sigma(x-x)^0$$

$$= \frac{1}{n}\Sigma 1$$

$$= \frac{n}{n} = 1$$

$$\mu_0' = \frac{1}{n}\Sigma(x - A)^0$$

$$= \frac{1}{n}\Sigma 1$$

$$= \frac{n}{n} = 1$$

For r = 3, $\mu_3 = \mu_3' - {}^3C_1\mu_1'\mu_2' + {}^3C_2(\mu_1')^2\mu_1' - {}^3C_3(\mu_1')^3\mu_0'$

$$\Rightarrow \boxed{\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3}$$

Similarly,

For r = 4, $\boxed{\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - (\mu_1')^4}$

Example 9: The first two moments of a distribution about the value 5 of the variable are 2 and 20 respectively. Find the mean and variance of the distribution.

Solution: given A = 5, $\mu_1' = 2, \mu_2' = 20$

We know that mean $(\bar{x}) = A + \mu_1' = 5 + 2 - 7$

Variance $(\sigma^2) = \mu_2 = \mu_2' - (\mu_1')^2$ = 20 – (2)² = 16

Example 10: Find the first four moment about the mean for the set of numbers 2, 4, 6 and 8.

Solution: $\mu_r = r^{th}$ Central moment $= \frac{\Sigma(x - \bar{x})^r}{n}$, r = 1, 2, 3, 4

Table for Calculation of First Four Moments

x	$x - \bar{x} = x - 5$	$(x - \bar{x})^2$	$(x - \bar{x})^3$	$(x - \bar{x})^4$
2	–3	9	–27	81
4	–1	1	–1	1
6	1	1	1	1
8	3	9	27	81
$20 = \Sigma x$	$0 = \Sigma(x - \bar{x})$	$20 = \Sigma(x - \bar{x})^2$	$0 = \Sigma(x - \bar{x})^3$	$164 = \Sigma(x - \bar{x})^4$

$$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{20}{4} = 5$$

$$\therefore \mu_1 = \frac{\Sigma(x - \bar{x})}{4} = 0, \quad \mu_2 = \frac{\Sigma(x - \bar{x})^2}{4} = \frac{20}{4} = 5$$

$$\mu_3 = \frac{\Sigma(x - \bar{x})^3}{4} = 0, \quad \mu_4 = \frac{\Sigma(x - \bar{x})^4}{4} = \frac{164}{4} = 41$$

Example 11: The first four moments of a distribution about the value 4 are– 1.5, 17, –30 and 108. Find the moments abour mean.

Solution: given A = 4

$$\mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30, \mu_4' = 108$$

$\therefore$ First moment about mean = μ_1

$$= A + \mu_1'$$

$$= 4 - 1.5 = 2.5$$

Second moment about mean = μ_2

$$= \mu_2' - (\mu_1')^2$$

$$= 17 - (-1.5)^2 = 14.75$$

Third moment about mean = μ_3

$$= \mu_3' - 3\mu_3'\mu_3' + 2(\mu_3')^3$$

$$= -30 - 3(-1.5) \times (17) + 2(-1.5)^3$$

$$= -30 + 76.5 - 6.75 = 39.75$$

Fourth central moment about mean

$$= \mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2 - 3(\mu_1')^4$$

$$= 108 - 4 \times (-1.5)(-30) + 6(-1.5)^2 \times 17 - 3(-1.5)^4$$

$$= 108 - 180 + 229.5 - 15.1875$$

$$= 142.3125$$

5.6 KARL PEARSON'S BETA (β) AND GAMMA (γ) COEFFICIENT

Prof Karl Pearson defined the following four co-efficients based on the 1st four central moment.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \rightarrow 5.6.1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \qquad \rightarrow 5.6.2$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \quad \left[\because \mu_2 = \sigma^2\right] \qquad \rightarrow 5.6.3$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 \qquad \rightarrow 5.6.4$$

It may be stated here that these co-efficients are pure numbers independent of units of measurement and as such can be conveniently used for comparative studies. In practice, they are used as measure of Skewness and Kurtosis as discussed below.

5.6.1 Co-efficient of Skewness Based on Moments

The co-efficient of Skewness, based on the moments is given by–

$$SK = \frac{\beta_2 + 3\sqrt{\beta_1}}{2(5\beta^2 - 6\beta_1 - 9)}$$

$\therefore$ Sk = 0 if $\beta_1 = 0$ or $\beta_2 + 3 = 0 \Rightarrow \beta_2 = -3$

But $\mu_4 = \frac{1}{n}\Sigma(x - \bar{x})^4 > 0$

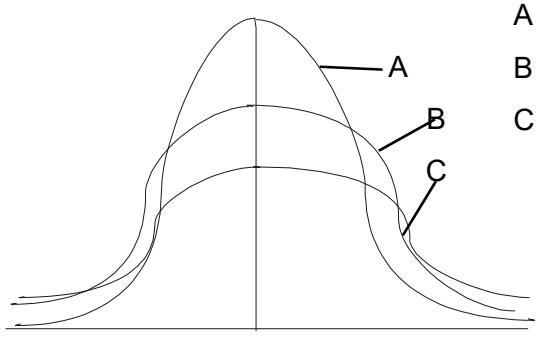
and $\mu_2 = \frac{1}{n}\Sigma(x - \bar{x})^2 > 0$

$\therefore \beta_2 = \frac{\mu_4}{\mu_2^{\,2}} > 0$

Since β_2 can not be negative, SK = 0 if $\beta_1 = 0$.

5.7 KURTOSIS

So far we have studied three measure viz central tendency, dispersion and skewness to describe the characteristics of a frequency distribution. However, even if we know all these three measures, we all not in a position to charcterise a distribution completely. The diagram given below will clarify the situation.



A – Lapto Kurtic
B – Meso Kurtic
C – Platy Kurtic

All these three curves are symmetrical about the mean and have same variation (range). But the shape and nature of the middle part of the curve are different i.e the peakedness of the three curves are not same. So. to identify a distribution completaly, we need one more measure which is called **'Kurtosis'**.

The word **Kurtosis** comes from a Greek word meaning 'humped'. In statistics, **Kurtosis refers to the degree of flatness or peakedness of the frequency curve.**

Curve of type B which is neither flat peaked is known as normal curveand shape of its hump is accepted as a standard one. This curve is termed as **mesokurtic.** The curves of type A, which is more peaked than the normal curve are known as **leptokurtic** and the curves of type C, which are flatter than the normal curve are called **platykurtic**.

5.7.1 Measures of Kurtosis

As a measure of Kurtosis, Karl Pearson gave the coefficient Beta two (β_2) or its derivative Gramma two (δ_2) defined as follows:

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{\mu_4}{6^4} \qquad \rightarrow 5.7.1 \text{ (a)}$$

$$\alpha_2 = \beta_2 - 3 = \frac{\mu_4}{6^4} - 3 \quad \rightarrow 5.7.2 \text{ (b)}$$

For a normal curve (mesakurtic curve), $\beta_2 = 3$ or $r_2 = 0$. For a lepto kurtic curve, $\beta_2 > 3$ or $r_2 > 0$ and for a platy kurtic curve, $\beta_2 < 3$ or $r_2 < 0$.

Note : A British statistician W. S. Gosset explained the use of the terms **platykurtic** and **leptokurtic** as: **'platykurtic curves like platyplus, are squat with short tails; leptokurtic curves are high with long tails like the kangaroos noted for leaping.**

Example 12: The first four moments of a distribution about the value 5 and 2, 20, 40 and 50. Do you think that the distribution is platykurtic?

Solution: Here A = 5, $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$, $\mu_4' = 50$

$$\therefore \text{ Mean } (\bar{x}) = \mu_1' + A = 2 + 5 = 7$$

Now, $\mu_2 = \mu_2' - (\mu_1')^2 = 20 - (2)^2 = 16$

$\therefore$ Variance $\mu_2 = 16$

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$

= 40 – 3 x 2 x 20 + 2 $(2)^3$ = – 64..7

$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

= 50 – 4 x 2 x 40 + 6 x 20 x $(2)^2$ – 3$(2)^4$ = 162

$$= \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = 0.63$$

Since $\beta_2 < 3$, the distribution is platykurtic.

Example 13: The first four moments of a distribution about the origin are 1, 4, 10 and 46 respectively. Calculate the central moments and indicates the nature of the distribution.

Solution: given A = 0, $\mu_1' = 1, \mu_2' = 4, \mu_3' = 10, \mu_4' = 46$

$\therefore$ Mean $(\bar{x}) = A + \mu_1' = 0 + 1 = 1$

Variance $(\sigma)^2 = \mu_2 = \mu_2' - (\mu_1')^2 = 4 - (1)^2 = 3$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$

= 10 – 3 x 4 x 1 + $2(1)^3$

= 0

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

= 46 – 4 x 10 x 1 + 6 x 4 x 1 – 3 = 27

$\therefore$ Karl Pearson's moment co-efficient of Skewness is given by

$$\delta_1 = \frac{\mu_3}{\mu_2^{3/2}} = 0, \ \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

Since, $\beta_1 = 0$ the given distribution is symmetrical (normal).

Again, Karl Pearson's measure of Kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{27}{(3)^2} = 3$$

and $\delta_2 = \beta_2 - 3 = 0$

Since, $\beta_2 = 3$, it is mesokurtic (normal) distribution with

Mean $(\bar{x}) = 1$

and S.D. $= \sqrt{3} = 1.732$

Example 14: If $\beta_1 = +1, \beta_2 = 4$ and variance = 9, find the values of μ_3 and μ_4 and comment upon the nature of Kurtosis.

Solution: Given $\beta_1 = +1, \beta_2 = 4$

$$\text{Variance} = \sigma^2 = \mu_2 = 9$$

$$\text{Now, } \beta_1 = +1 \Rightarrow \frac{\mu_3^2}{\mu_2^3} = 1 \Rightarrow \mu_3^2 = \mu_2^3 = 9 \times 9 \times 9 = (3 \times 3 \times 3)^2$$

$$\Rightarrow \mu_3^2 = (27)^2$$

$$\Rightarrow \mu_3 = \pm 27^2$$

Also $\beta_2 = 4$

$$\Rightarrow \frac{\mu_4}{\mu_2^2} = 4$$

$$\Rightarrow \mu_4 = 4 \times 9 \times 9 = 324$$

$\therefore \mu_3 = \pm 27$ and $\mu_4 = 324$

$\mu_2 = 4 > 3$

Hence the given distribution is leptokurtic.



CHECK YOUR PROGRESS

Q.1: Explain meaning of skewness using sketches of frequency curves. How does skewness different from dispersion.

Q.2: Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why?

Q.3: How is Kurtosis measured ?

Q.4: Fill in the blanks:

i) For a symmetrical distribution mean, median and mode are .....

ii) If $\mu_2 > 3$, the curve is called .....

iii) If $\mu_1 = 0$, the distribution is .....

iv) If Mode > Mean, the distribution is ..... skewed.

v) Kurtosis measures ..... of the frequency curve.

vi) Bowley's co-efficient of skewness lies between .....

vii) Relative measure of Kurtosis is .....

Q.5: Say whether the following statements are true or false.

i) For a symmetrical distribution $\beta_1 = 0$.

ii) Bowely's co-efficient of skewness lie between ± 3.

iii) For a negatively skewed distribution mean > mode.

iv) The first moment about the origin L mean.

v) Variance = μ_2

vi) If $\beta_2 > 3$, the curve is platykurtic.

Q.6: Consider the following distribution–

	Distribution A	**Distribution B**
Mean	100	90
Median	90	80
Variance	100	100

i) Both the distribution have same degree of skewness.

ii) Distribution A has the same degree variation as B.

–Say whether the following statements are true or false. Justify your answer.

Q.7: For a distribution $\bar{x} = 30$, $\sigma = 8$ and SK_p= 0.4. Find Median and Mode.

Q.8: For a distribution, Mean = 50, CU = 40%, SK_p = –0.4. Find SD, Median and Mode.

Q.9: For a group of 10 items, $\Sigma x = 452$ and Mode = 43.7. Find the pearsonion co-efficient of Skewness.

Q.10: The measure of skewness for a distribution is – 0.8. If the upper and lower quartiles are respectively 56.6 and 44.1, find Median.

Q.11: Calculate Karl Pearson's co-efficient of skewness from the data given below.

Marks	No. of students
21 – 25	5
26 – 30	15
31 – 35	28
36 – 40	42
41 – 45	15
46 – 50	12
51 – 55	3

Q.12: Calculate Bowley's co-efficient of skewness

Wt (kg):	0–15	15–30	30–45	45–60	60–75	75–90	90 and above
f:	20	30	30	35	45	15	5

Q.13: a) Find the first and second moments about the mean for the numbers 3, 4, 5 and 8.

b) Find the first, second, third moments of the numbers 1, 3, 6, 7, 8 about the value 4.

Q.14: a) The first three moments of a distribution about the value 1 of the variable are 2, 25 and 80. Find the mean, standard deviation and the moment measure of skewness.

b) The first four moments of a distribution about x = 2 are 1, 2.5, 5.5 and 16. Calculate the four moments about ($\bar{x}$) and about zero.

Q.15: a) The central moments of a distribution are given by μ_2 = 140, μ_3 = 148 and μ_4 = 6030, calculate the moment measure of skewness and kurtosis and comment on the shape of the distribution.

b) From the data given below do you think that the distribution is platykurtic?

N = 100, $\Sigma fd = 50$, $\Sigma fd^2 = 1970$, $\Sigma fd^3 = 2948$ and $\Sigma fd^4 = 86, 752$; where d = x – 48.



5.8 LET US SUM UP

In this unit we have learnt about the following–

- Meaning of symmetrical and asymmetrical distribution.
- Meaning of Skewness and types of Skewness.
- Different measures of skewness.
- Meaning of moments, central moments, Raw moments.
- Use of moments for describing various characteristics of frequency distribution.
- Meaning of kurtosts and types of Skewness.
- To measure Skewness and Kurtosis using moments.



5.9 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publication.
2) Arora, P. N., Arora, Sumeet & Arora, S. *Comprehensive Statistical Methods.* S. Chand and Company.
3) Ghosh, R. K. and Saha, S. *Business Mathematics and Statistice.* New Central Book Agency (P) Ltd.
4) Gupra, S. C. *Fundamental of Statistics.* New Delhi: Himalaya Publishing House.
5) Gupta, S. C. and Gupta, Indira. *Business Statistics.* New Delhi: Himalaya Publishing House.
6) Sharma, J. K. *Business Statistics.* New Delhi: Pearson Education Ltd.



5.10 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 4: (i) Equal, (ii) Laptokurtic, (iii) symmetrical, (iv) negatively, (v) peakedness/flatness, (vi) ± 1, (vii) β_2 or r_2.

Ans. to Q. No. 5: (i) True, (ii) False, (iii) False, (iv) True, (v) True, (vi) False.

Ans. to Q. No. 6: (i) True, (ii) False

Ans. to Q. No. 7: Median = 28.93, Mode = 26.8

Ans. to Q. No. 8: SD = 20, Mode = 58, Median = 52.67

Ans. to Q. No. 9: $SK_p = 0.077$

Ans. to Q. No. 10: Median = 55.35

Ans. to Q. No. 11: $SK_p = -0.109$

Ans. to Q. No. 12: $SK_B = -0.1126$

Ans. to Q. No. 13: a) $\mu_1 = 0,\ \mu_2 = 3.5$

b) $\mu_1' = 1,\ \mu_2' = 7.8,\ \mu_3' = 14.2$

Ans. to Q. No. 14: a) mean = 3, SD = 4.58

$\delta_1 = \sqrt{\beta_1} = -0.561$

b) Moments about mean.

$\mu_2 = 1.5,\ \mu_3 = 0,\ \mu_4 = 6$

Moments about zero

$\mu_1' = 3,\ \mu_2' = 10.5,\ \mu_3' = 40.5,\ \mu_4' = 168$

Ans. to Q. No. 15: a) $\delta_1 = 0.0893,\ \beta_2 = 0.3076$

The distribution is approximately symmetrical and platykurtic

b) $\beta_2 = 2.2.14,\ \beta_2 < 3$, distribution is platykurtic.



5.11 MODEL QUESTIONS

Q.1: Averages dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distribution.

Q.2: Define moments establish the relationship between the moments about mean and moments about any orbitrary point.

Q.3: The first two moments of a distribution about the value 5 are 2 and 20. Find the mean and variance.

Q.4: Calculate Karl Peason's Co-efficients of Skewness for the given distribution.

Marks less than:	20	40	60	80	100
No. of students:	18	40	70	90	100

Q.5: Calculate Bowley's Co-efficient of Skewness.

Marks:	0–10	10–20	20–30	30–40	40–50
No. of students:	5	9	12	8	6

Q.6: The arithmetic mean of a certain distribution is 5. The second and the third moments about the mean are 20 and 140 respectively. Find the third moment of the distribution about zero.

*** ***** ***

UNIT 6: CORRELATION

UNIT STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.3 Correlation: Meaning and Concept

6.4 Types of Correlation

6.5 Methods of Measuring Correlation

6.6 Scatter Diagram and Correlation

6.7 Correlation Coefficient

6.8 Underlying Assumptions

6.9 Interpretation of Correlation Coefficient

6.10 Rank Correlation Coefficient

6.11 Let Us Sum Up

6.12 Further Reading

6.13 Answers To Check Your Progress

6.14 Model Questions

6.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of correlation coefficients of correlation and their properties
- discuss different types of correlation and regression
- explain the essential features of correlation and regression analysis
- explain and compare the various aspects of correlation.

6.2 INTRODUCTION

Depending on the number of variables that we study, we may have univariate, bivariate or multivariate series or distribution. Univariate series or distribution is one when we deal with only one variable. For example, if we consider, say, only age of different persons, then we will have a univariate series, i.e. values of only age. Bivariate distribution, on the other hand,

involves two variables (note that 'bi' means two). If, for instance, we take both income and expenditure of persons then distribution thus obtained will constitute a bivariate distribution. Multivariate distribution will, similarly consist of more than two variables. The two concepts-correlation and regression that we will study in this unit are associated with bivariate and/or multivariate series or distributions.



LET US KNOW

A **variable** is an entity that can take any value. Age, income, height, weight, expenditure, marks etc. are examples of variables. Note that for different persons these will take different values. Contrary to it, **constant** is an entity that can assume only fixed value.



ACTIVITY 6.1

Try to list some variables and constants that you find in common day to day life.

.....

.....

6.3 CORRELATION: MEANING AND CONCEPT

In the simplest term, correlation means relation between two variables (note that 'co' simply means two things together). In case of bivariate distributions, many a times, we may be interested in looking at how the two variables are related among themselves. Obviously, when two variables are related, change in any one variable is responded by the other variable showing some changes in it. Correlation analysis tries to measure the 'extent' of such responses among the variables. Thus, correlation is an analysis of co-variation between two (or even more) variables. Two variables are said to be correlated if the change in any one of them results in a corresponding change in the other.

6.4 TYPES OF CORRELATION

Correlation may be classified into the following three types:

a) Positive correlation
b) Negative correlation and
c) Zero Correlation.

a) Positive Correlation: When increase (or decrease) in one variable leads to a corresponding increase (or decrease) in the other variable then it is called positive correlation. The basic idea of positive correlation is that both the variables move in the same direction. For example, there exists positive correlation between the following pairs of variables:

i) Income and expenditures of a group of families
ii) Amount of rainfall and yield of crops
iii) Price and supply of a company
iv) Temperature and sale of ice-creams on different days of a month in summer.

b) Negative Correlation: Negative correlation is, on the other hand, is a situation when increase (or decrease) in one variable is accompanied by a decrease (or increase) in the other variable. Unlike the positive correlation, the fundamental element in negative correlation is that the variables move in opposite directions. For example, following are the situations of negative correlations:

i) Price of a commodity and demand for it.
ii) Sale of woolen garments and the day temperature.
iii) Number of workers and the time required to complete work.

c) Zero correlation: In case, variables are found as uncorrelated i.e., change in one of the variables fail to produce any corresponding change in the other, and then it is called zero correlation. For example, one should expect zero correlation between the heights of workers and the income earned by them, or between price of rice and demand for sugar.

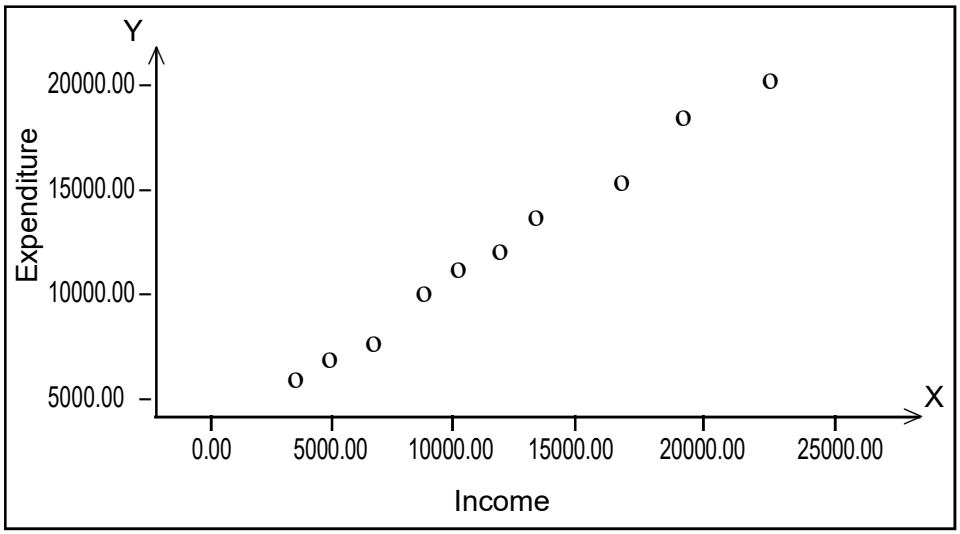
6.5 METHODS OF MEASURING CORRELATION

The following are the methods used to measure the correlation between two variables:

i) Scatter Diagram Method

ii) Karl Pearson's Correlation Method and

iii) Spearman's Rank Correlation Method.



6.6 SCATTER DIAGRAM AND CORRELATION

How can one find whether any two variables are correlated or not? One of the easiest ways for identifying correlation between two variables is scatter diagram. Scatter diagram is a very simple graphical technique of identification of correlation. When we consider respective pairs of values as points and diagram them against X-Y axes then we get the scatter plot of the values. To understand this, let us consider the following bivariate series of income and expenditure of say, ten persons.

Person (No.)	Income (Rs.)	Expenditure (Rs.)
1	9200	6700
2	6860	4256
3	12825	9875
4	17980	12900
5	22125	19700

6	7892	4780
7	4645	3200
8	10925	8645
9	13820	10900
10	24780	21200

Taking income in X-axis and expenditure in Y-axis, the values of the above table can now be plotted in a graph to see whether these values are correlated. The scatter thus obtained will be somewhat like this

It may be seen from the above scatter that points show an upward-rising trend. If the scatter does indicate some sort of trend or pattern then it implies existence of correlation among the variables. If, for example, pattern is such that there is a visible upward trend from left to right, then variables are said to have positive correlation. In case, there is no such identifiable pattern then the variables are said to be uncorrelated.



ACTIVITY 6.2

Note down height and weight of 10 of friends of yours. Then try to draw a scatter plot of the values. What can you say about their possible correlation?

.....

.....



CHECK YOUR PROGRESS

Q.1: Choose the correct answer from the following:

i) Correlation and regression are associated with–

a) Univariate series b) Bivariate series

c) Both bivariate and multivariate series

d) Multivariate series

.....

ii) Scatter plot technique is used to detect–

a) Correlation b) Regression

.....

iii) When variables move in opposite directions then it is–

a) Positive Correlation b) Negative Correlation

.....

6.7 CORRELATION COEFFICIENT

Scatter plots are particularly useful to identify whether the variables are correlated and in case they are then what is the type of such relation. It, however, can't quantify or quantitatively measure the 'amount', 'extent', 'magnitude' or 'degree' of relation that exists between the variables. British statistician Karl Pearson (1867-1936) introduced a simple technique for measuring the magnitude of such correlation. His measure is called Pearson's Product-Moment Correlation Coefficient or simply Pearson' Correlation Coefficient and usually denoted by 'r' (standing for 'relation') which is given as:

$$r_{x,y} = \frac{Cov(x,y)}{\sigma_x.\sigma_y}$$

Here, cov (x, y) means ''covariance of x and y'' and σx and σy are standard deviations of x and y respectively. Formulas for obtaining these are as bellow:

$$Cov(x, y) = \frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum\left(x-\bar{x}\right)^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum\left(y-\bar{y}\right)^2}$$

Putting these in the original formula and simplifying, we will have,

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\sum\left(x-\bar{x}\right)^2\left(y-\bar{y}\right)^2}}$$

This can be further simplified and can be written as:

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{\left[n\sum x^2-\left(\sum y\right)^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}}$$

This formula is relatively easier to calculate and most commonly used to obtain the Pearson's coefficient of correlation; 'n' here means total number of observations.

Here is procedure of finding the 'r' for a series involving two variables x and y:

- Obtain total of x i.e. Σx
- Obtain total of y i.e. Σy
- Get product of x and y for all pairs and then get the total of xy, i.e. Σxy
- Get x^2 and y^2 for all values of x and y and then get their totals, i.e. $\Sigma x^2 y^2$
- Put these values to the formula and get the r.



LET US KNOW



Carl Pearson, later known as Karl Pearson (1857-1936) was born to William Pearson and Fanny Smith, was also an accomplished historian and Germanist. Karl Pearson was educated privately at University College School, after which he went to King's College, Cambridge in 1876 to study mathematics. He then spent part of 1879 and 1880 studying medieval and 16th century German literature at the universities of Berlin and Heidelberg. He graduated from Cambridge University in 1879 as Third Wrangler in the Mathematical Tripos. In 1911 he founded the world's first university statistics department at University College London. Besides, correlation coefficient, chi-square is another significant tool contributed by Pearson.



EXERCISE 6.1

A) Given the following pairs of values:

Capital employed (lakhs of Rs.)	12	13	15	16	18	29	32
Profits (lakhs of Rs.)	6	5	7	8	12	11	17

a) Draw a scatter diagram.

b) Do you see any correlation between capital employed and volume of profits? What can be concluded about their correlation if it exists?

B) Calculate Pearson's coefficient of correlation for following data:

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

6.8 UNDERLYING ASSUMPTIONS

The most fundamental assumption of Pearson's correlation coefficient is that it supposes that the variables are quantitative. Further, it assumes a linear relation between the two variables. Therefore, the value of 'r' indicates extent of linearity that exists between the two variables. In case, the value of 'r' is found to be equal to 'zero', it may be best understood as 'absence of linear relations' among the variables. It may, however, be possible that the variables have other non-linear types of relations.

6.9 INTERPRETATION OF CORRELATION COEFFICIENT

Once the value of 'r' is obtained, it is very important to correctly interpret the result. Interpretation of the value of 'r' depends on two things– sign and magnitude. When 'sign' is positive correlation is said to be positive, and when 'sign' happens to be negative correlation is also said to be negative.

On the other hand, the magnitude of the value of correlation coefficient can not exceed the value of 1. The value of 'r', therefore, will always lie in between –1 and +1. More closer is the value to 1, greater is the

extent of linear correlation. Conversely, further away is the value from 1 greater is amount of deviation from (linear) correlation. Many scholars provide ranges of values of 'r' for meaningful interpretation. Most commonly it is held that–

When the value of 'r' is 1 it is called perfect correlation. For value ranging from 0.8 to 1, variables are said to be "highly correlated". When the value lies in between 0.6 to 0.8, variables are called "fairly correlated". For value of 'r' in between 0.3 to 0.6, it is called moderate correlation. For values of 'r' less than 0.3 it is called low correlation. When the value of 'r' is equal to 0, the variables are said to be linearly uncorrelated or linearly independent.



CHECK YOUR PROGRESS

Q.2: Choose the correct answer from the following:

i) Correlation analysis is based on the assumption that–

a) Variables are quantitative

b) Variables are qualitative

.....

ii) Interpretation of r depends on–

a) Only sign b) Only magnitude

c) Both sign and magnitude

.....

iii) When value of r falls within the range of 0.6 to 0.8 it is called–

a) Fair Correlation b) Moderate Correlation

c) Low Correlation c) Perfect Correlation

.....

iv) In case of perfectly positive correlation, the value of r is

a) –1 b) +1 c) 0.8 d) 0

.....

6.10 RANK CORRELATION COEFFICIENT

Many a times, we come across variables, which can not be quantitatively measured. However, these can be 'ranked' and accordingly 'ordered'. For example, beauty, intelligence, honesty etc. In fact, most of the qualitative attributes are of this type. It is difficult to measure exact amount of honesty that a person has, or exact amount of intelligence that he possesses. It is, nevertheless, very much possible to find out which person is more honest, or intelligent or beautiful so on and so forth. When we try to do such comparisons, we in fact, attach some ranks to individual and then try to arrange in some order depending on the ranks. Such variables which can't be directly measured but can be ordered are called ordinal variables. When we have a bivariate distribution involving two such ordinal variables then how do we find correlation? In this case Pearson's correlation coefficient can not be applied as it is based on the fundamental assumption of quantitative variable.

Charles Edward Spearman, a British psychologist, developed a formula in 1904 for obtaining a coefficient of correlation of ranks of two attributes. To obtain the correlation of ranks the two attributes are to be ordered and then their differences of ranks are to be calculated. If say 'd' is the difference of ranks between the two attributes then the rank correlation coefficient is given by–

$\sigma = 1 - \frac{6\sum d^2}{n(n^2-1)}$ n is the number of observation

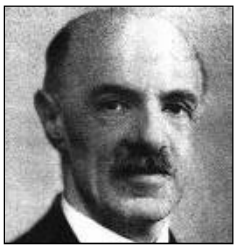
Procedure to obtain the rank correlation coefficient is fairly simple and easy:

- Arrange the two attributed in same order by their ranks.
- Subtract each pair of ranks to get their differences, i.e. d and then get d^2.
- Get the total of d^2 i.e. Σd^2.
- Put the values thus obtained in the formula and get the coefficient.

Just like the value of r, the value of also lies in between –1 and +1. Interpretation is also same as r except the foundational assumption of linearity.



LET US KNOW



Charles Edward Spearman, (September 10, 1863 - September 17, 1945) was an English psychologist known for work in statistics. He was a pioneer of factor analysis. He also did seminal work on models for human intelligence. Although Spearman achieved most recognition in his day for his statistical work, he regarded this work as subordinate to his quest for the fundamental laws of psychology, and he is now similarly renowned for both.



CHECK YOUR PROGRESS

Q.3: Choose the correct answer from the following:

i) Rank Correlation analysis is used when

a) Variables are quantitative

b) Variables are qualitative

.....

ii) Spearman's rank correlation does not consider linearity

a) True b) False

.....

iii) Just like the value of r, Rank correlation also falls within +1 and –1

a) True b) False

.....



ACTIVITY 6.3

Ranking of 10 students by two teachers A and B are shown below. Obtain the rank correlation coefficient for the data and figure out the problem in decision making

A:	3	5	8	4	7	10	2	1	6	9
B:	6	4	9	8	1	2	3	10	5	7

.....

.....



6.11 LET US SUM UP

In this unit, we have discussed the following:

- Depending on the number of variables, we have univariate, bivariate or multivariate distributions.
- Correlation and regression are associated with bivariate and/or multivariate distributions
- Correlation is the relation between two variables.
- Scatter diagram is a simple graphical method of identifying correlation. For correlated variables scatter plot exhibits some observable pattern or trend.
- Correlation can be of two types– positive when variables move in the same direction and negative when the variables move in the opposite directions.
- Correlation coefficient is based on assumptions of that variables are quantitative and linearly related.
- Interpretation of the correlation coefficient depends on both sign and magnitude of the value of r.
- Rank correlation is applied for qualitative variables which are not measurable but ordinal.
- Like Pearson's correlation coefficient, Spearman's rank correlation coefficient also lies in between +1 and –1.



6.12 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



6.13 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: i) c) Both bivariate and multivariate series
ii) a) Correlation
iii) b) Negative Correlation

Ans. to Q. No. 2: i) a) Variables are quantitative
ii) c) both sign and magnitude
iii) a) Fair Correlation
iv) b) +1

Ans. to Q. No. 3: i) a) Variables are quantitative
ii) a) True
iii) a) True



6.14 MODEL QUESTIONS

Q.1: What is meant by correlation? What are the properties o the coefficient of correlation?

Q.2: What are the methods of calculating coefficient of correlation?

Q.3: Calculate Karl pearson's coefficient correlation from the data given below

X :	2	4	6	8	10
Y:	20	18	16	14	12

Q.4: Find Karl Pearson's coefficient of correlation between x and y from following data giving test scores of 10 candidates in mathematics and statistics and interpret:

Scores in Mathematics:	98	70	40	20	85	75	95	80	10	5
Scores in Statistics:	85	65	32	30	80	60	61	55	54	65

Q.5: Find the coefficient of correlation from the following data give below:

X:	12	20	15	18	33	24	30	12	15	22
Y:	30	35	28	36	29	39	30	25	30	38

*** ***** ***

UNIT 7: REGRESSION

UNIT STRUCTURE

7.1 Learning Objectives

7.2 Introduction

7.3 Meaning and Concept of Regression

7.4 Linear Regression

7.5 Line of Regression and Regression Equation

7.6 Coefficient of Regression

7.7 Relation between Correlation and Regression Coefficients

7.8 Correlation and Regression Analysis: A Comparison

7.9 Let Us Sum Up

7.10 Further Reading

7.11 Answers To Check Your Progress

7.12 Model Questions

7.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of regression and regression, lines and equation of regression, and their properties
- discuss different types of regression
- explain the essential features of correlation and regression analysis
- explain and compare the various aspects of regression analysis.

7.2 INTRODUCTION

In the previous unit, we introduced the concept correlation between two variables where it was discussed whether the variables move in the same direction or in the opposite direction and the extent of association between the two variables under study. On the other hand regression analysis is a statistical technique that expresses the relationship between two or more variables in the form of an equation to estimate the value of a variable, based on the given value of another variable. In business, several times it

becomes necessary to have some forecast so that it can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets.

7.3 MEANING AND CONCEPT OF REGRESSION

Regression is one of the most important and powerful statistical techniques widely used in both science and social science studies. The term regression was introduced by Francis Galton. In a famous scientific paper ‘‘Family Likeness in Stature’’ published in Proceedings of Royal Society, London in 1886, he first talked about ‘‘regression to mediocrity’’.

Today, broadly speaking, regression analysis is referred to the study of the dependence of one variable on one or more other variables, with a view to estimating and/or predicting the average value of the former in terms of fixed or known values of the latter.

To understand consider this - suppose x and y are two variables such that values of y depends on values of x. Then regression analysis will try to estimate or predict the average value of y that it will take for any given value of x.

7.4 LINEAR REGRESSION

We have understood that regression is the technique of estimating average value of the dependent variable corresponding to any given value of the independent variable when the two variables are dependent upon each other i.e. the two variables are ‘related’ to each other. Now we know that relation may be either linear or non-linear. Linear regression is the technique of regression when variables are linearly related i.e. for each successive change in the independent variable there is a ‘constant’ change in the dependent variable.



LET US KNOW

A dependent variable is a variable that depends on value of some other variable. Independent variable is a variable that does not depend on values of the other variable.

Linearity on the other hand means that when independent variable changes the dependent variable changes by constant rate. Such changes produce a straight line when plotted.



ACTIVITY 7.1

Try to list five dependent variables and five independent variables. Also find out which depends on what.

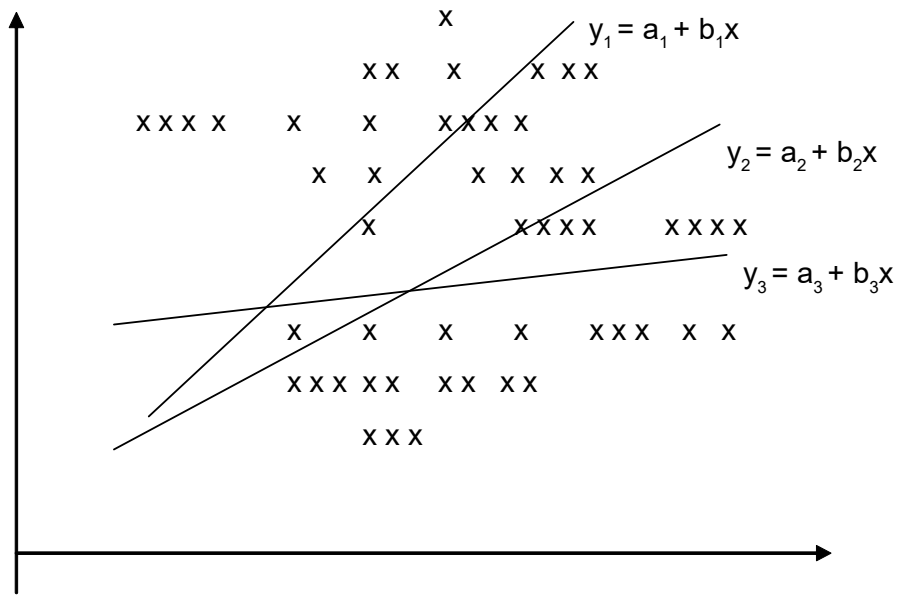
.....

.....

7.5 LINE OF REGRESSION AND REGRESSION EQUATION

Let us consider two variables *x* and *y*, which are linearly related. Relation is such that value of *y* depends on value of *x*. The relation may be expressed as y = a + bx, where *a* and *b* are two constants usually called parameters.

Note that value of *y* now depends on values of *a* and *b* given the value of *x*. For different values of the parameters *a* and *b*, we will have different lines that will pass through the points showing pairs *x* and *y*. Look at the following figure:



In the figure you will notice that three lines passing through the points produces three different values of *y* depending on three different values *a* and *b*. Now since all three values of *y* are 'estimated' value therefore there will be an 'error' of estimation. Error is simply the difference between the 'actual' and 'estimated' values of *y*. That is:

$e = y - \hat{y}$; where $\hat{y}$ is the estimated value of y.

Ideally it would be better if 'error' *r* be zero or minimum. Now since $\hat{y}$ depends on values of *a* and *b* we need to estimate *a* and *b* in such a way that *e* becomes minimum. Suppose, we estimate values of *a* and *b* as $\hat{a}$ and $\hat{b}$ so that we have $\hat{y}$ as:

$\hat{y} = \hat{a} + \hat{b}x$ and resultant *e* is minimum.

Then this equation is called the regression equation and the line that this equation produces is called the line of regression or simply regression line. In other words regression equation is the equation that estimates dependent variable in such a way that the difference between the estimated and actual value of *y* or error in short, is minimum.

The line of regression often interpreted as line of "best fit". Among the all possible lines (look at the figure above showing three such lines) the line of regression fits i.e. touches most of the points (and that is why error is minimum).

When dependent and independent variables are clearly identified then we have single line of regression. When it is not clear will have two lines of regression one showing *y* as dependent and other showing *x* as dependent.

It is useful to remember that in case the variables are perfectly related i.e. if correlation coefficient is either –1 or +1 both the lines will coincide. When the variables are linearly independent regression lines will be perpendicular to each other. Also, both the lines of regression pass through the mean values of the variables.

7.6 COEFFICIENT OF REGRESSION

Consider the regression equation:

$$\hat{y} = \hat{a} + \hat{b}x$$

In this equation $\hat{a}$ is called 'constant' term of the regression equation and $\hat{b}$ is called the 'coefficient' of regression. How can one obtain the $\hat{b}$? In fact, $\hat{b}$ is simply given by

$$\hat{b} = \frac{Cov(x,y)}{\sigma_x}$$

Formal derivation of $\hat{b}$ involves partial differentiation. However, the following formula is used to estimate the regression coefficient–

$$\hat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Once the $\hat{b}$ is obtained then the constant can be obtained by solving the following equation:

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

Where $\bar{y}$ and $\bar{x}$ are means of *y* and *x* respectively. The regression coefficient indicates the amount of variation in dependent variable explained by the independent variable.

7.7 RELATION BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

It may be seen that both coefficient of regression and coefficient of correlation are related to each other. We know that:

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}; \text{ and}$$

$$\hat{b} = \frac{Cov(x,y)}{\sigma_x}$$

It is thus clear that–

$$r = \frac{\hat{b}}{\sigma_y}$$

This relation is practically very useful for analysis. Remember that, in case we don't know clearly which one is dependent and which one is independent variable then we will have two lines of regression and hence two coefficients of regression. Let us assume that the two possible coefficients of regression are $\hat{b}$ and $\tilde{b}$ respectively. Then they satisfy the following property–

$$r^2 = \hat{b} \times \tilde{b}$$

This follows from the above property that both regression coefficients have same sign and if value of any one is greater than unity value of the other must be less than unity.

7.8 CORRELATION AND REGRESSION ANALYSIS: A COMPARISON

Correlation analysis simply detects whether two variables are having any relations or not. Such detection simply looks for possible movements in variables and then some identifiable patterns in the movements. On the other hand, regression analysis tries to estimate or predict average value of one variable depending on given values of the other variable. It also tries to explain the extent of variation caused by a variable. In this sense regression analysis has greater explanatory abilities than correlation analysis.



CHECK YOUR PROGRESS

Q.1: Choose/Tick the correct answer from the following:

A) Regression analysis tries to estimate ..... values.

a) Average b) Median

c) Mode d) Range

.....

B) The line of regression is called line of ..... fit.

a) Worst b) Highest

c) Best d) Lowest

.....

C) Regression equation estimates the values such that error is–

a) Minimum b) Maximum

c) Normal d) Abnormal

.....

D) When correlation is 0, then the angle between the lines of regressions–

a) 90^0 b) 30^0 c) 45^0 d) 0^0

.....

E) Lines of regression pass through ..... values of the variables.

a) Maximum b) Minimum

c) Mean d) Median

.....

F) Regression coefficients are of opposite sign.

a) True b) False

.....

G) Both the regression coefficients can be greater than unity.

a) True b) False

.....



7.9 LET US SUM UP

In this unit, we have discussed the following:

- Regression is a method of estimating average value of dependent variable given the value of independent variable.
- Line of regression gives the line of best fit and regression equation defines the line of regression. The line of regression passes through the means of the variables.
- The regression coefficient indicates the amount of variation in dependent variable explained by the independent variable.
- Correlation analysis simply detects whether two variables are having any relations or not. On the hand regression analysis tries to estimate or predict average value of one variable depending on given values of the other variable.



7.10 FURTHER READING

1) Agarwal, D. R. (2006). *Business Statistics*. Delhi: Vrinda Publications.
2) Gupta S. C. (1994). *Fundamentals of Statistics*. New Delhi: Himalayan Publishing House.
3) Rajagopalan, S. P. & Sattanthan R. (2009). *Business Statistics and Operations Research*. New Delhi: Tata McGraw-Hill
4) Sharma, J. K. (2007). *Business Statistics*. New Delhi: Pearson Education Ltd.
5) Verma, A. P. (2007). *Business Statistics*. Guwahat: Asian Books Private Limited.



7.11 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: A) (a) Average, B) (c) Best, C) (a) Minimum, D) (a) 90^0, E) (c) Mean, F) (b) False, G) (b) False



7.12 MODEL QUESTIONS

Q.1: Explain the concept of regression and bring out its utility in business and commerce.

Q.2: What is a regression line? Why can we have two lines of regression? Explain the relationship between lines of regression and correlation coefficient.

Q.3: Why the regression line is called line of best fit? What do you understand by line of best fit?

Q.4: What is regression coefficient? What does it indicate?

Q.5: Fit a line of regression to following data, taking Y as dependent variable and estimate Y for X is equal to 20.

Y:	1	2	4	5	7	8	9
X:	1	3	4	8	9	11	14

Q.6: If followings are the two lines of regression find out the correlation coefficient and means of x and y

$8x - 10y + 66 = 0$ and $40x - 18y = 214$

Q.7: Distinguish between correlation and regression analyses.

*** ***** ***

Name of the Paper: .....
Course Code: .....

Centre for Internal Quality Assurance (CIQA)
Krishna Kanta Handiqui State Open University
City Office: Housefed Complex, Guwahati-781006

Learner's Feedback on Course

Dear Learner,

Regarding the course as mentioned above, we would like to know your opinions and comments so as to improve the quality of self learning materials in future. Please respond to the following statements by ticking the number you feel most reflect your opinion. After completion of the additional comments, please detach the page and send/mail the same to us at the address given below.

The Director, Centre for Internal Quality Assurance, KKHandiqui State Open University
Housefed Complex, Dispur, Guwahati-781006
(E-mail id: ciqa@kkhsou.in)

1) Approximately how many hours did you spend for studying the units in the course? ☐☐☐

2) Please give your opinions (by ✓ mark) to the following items based on your reading of the block:

Sl. No.	Statements	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I)	The SLMs of the course fulfil the learning objectives set out in the beginning of each unit					
II)	The units of the SLM could satisfy my academic needs and expectations					
III)	The Unit writers have excellent knowledge about the course contents					
IV)	Language and contents of the units were presented at a level which I could readily understand					
V)	Ample opportunity for participation in the activities provided in the units					
VI)	Used enough Illustrations (Diagrams, tables etc.) for conceptual clarity					
VII)	Quality of content is engaging, relevant, and up-to-date					
VIII)	The self check questions are very helpful					
IX)	The Possible/model questions and the answers to check my progress have benefited me a lot					

Additional Comments: *(Please feel free to provide your open comments)*

1) Which aspects of the SLM, according to you, worked well?

   .....

   .....

2) What sort of changes/improvements do you feel KKHSOU could implement to improve the overall quality of the SLM?

   .....

Thank you for taking the time to complete this form.