UG BA S4-02
(DSC)

FOUR YEAR BACHELOR'S DEGREE COURSE
(CBCS in Accordance with NEP 2020)

SEMESTER IV

BUSINESS STATISTICS

MANAGEMENT

Maniram Dewan School of Management
KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

Prof. Nripendra Nayan Sarma, Moniram Dewan School of Management, KKHSOU
Prof. Munindra Kakati, VC, ARGUCOM
Prof. Rinalini Pathak Kakati, Dept. of Business Administration, GU
Professor Nayan Baruah, Gauhati University
Professor H. C. Gautam, Dept of Commerce, Gauhati University
Dr. S. K. Mahapatra, Gauhati University

Course Coordinator: Harekrishna Deka, K.K.H.S.O.U.

SLM Preparation Team

Units	Contributors
1, 2 & 3	Dr. Padmalochan Hazarika, Gauhati University
5, 6	Dr. Jaydeep Baruah, O.K.D. Institute of Social Change and Development
4	Ajanta Majumdar, Gauhati commerce college
7,13	Dr. Tarakeswar Chaudhury, Retired Professor, Cotton College
12	Dr. Rupam Barman, IIT Guwahati
8	Harekrishna Deka, KKHSOU
9,10, 11	Dr. Rijusmita Sarma, HoD, Department of Statistics L.C.B. College, Guwahati

Editorial Team

Content : Dr. Mitali Deka, NERIM

Structure, Format & Graphics: Harekrishna Deka, K.K..H.S.O.U.

March, 2025

© Krishna Kanta Handiqui State Open University.

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

For the avoidance of doubt, by applying this license KKHSOU does not waive any privileges or immunities from claims that it may be entitled to assert, nor does KKHSOU submit to the jurisdiction, courts, legal processes or laws of any jurisdiction.

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

Headquarters : Patgaon, Rani Gate, Guwahati - 781017

City Office : Housefed Complex, Dispur, Guwahati-781006; Web: www.kkhsou.in

COURSE INTRODUCTION

The course *'Business Statistics'* of BBA 4th semester introduces us to the subject-Statistics. The course consists of units thirteen.

Unit 1 introduces us to some of the basic concepts of Statistics. This unit focuses on the different methods data collection, tabulation of data and presentation of statistical information.

Unit 2 gives us a broad overview of measures of central tendency. In Unit 3 we will discuss the measures of dispersion and skewness. In Unit 4, we will discuss skewness, moments and kurtosis.

In Unit 5 and 6, we will discuss about correlation, scatter diagram, types of correlation, correlation coefficient, regression, linear regression, lines of regression, regression equation etc.

In Unit 7, we will focus on the concepts of probability. Here we will discuss different theorems on probability.

In Unit 8, we discuss the concept conditional probability, Multiplication theorem of probability, Law of probability and Baye's theorem.

In Unit 9, we will discuss random variables and its probability distribution. Unit 10, gives us a broad overview of theoretical probability distributions relating to discrete variables.

In Unit 11, we will discuss about index numbers, types of simple index numbers, construction of simple and weighted index numbers etc.

In Unit 12, we will discuss time series and its various components.

Unit 13 is about sampling. Here we will discuss certain concepts like, sample, population, sampling, random and non-random sampling, and errors in sampling.

While going through a unit, you will notice some along-side boxes, which have been included to hep 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.

BACHELOR OF BUSINESS ADMINISTRATION

Business Statistics

DETAILED SYLLABUS

		Page
UNIT 1:	**INTRODUCTION TO STATISTICS** Meaning and importance of Statistics; Frequency Distribution; Tabulation of Data; Methods of presenting statistical information.	**7-44**
UNIT 2:	**MEASURES OF CENTRAL TENDENCY** Meaning of Measures of Central Tendency; Different Types of Measures of Central Tendency.	**45-69**
UNIT 3:	**MEASURES OF DISPERSION** Meaning of Dispersion; Different Measures of Dispersion; Meaning, Types and Measures of Skewness.	**70-92**
UNIT 4:	**SKEWNESS, MOMENTS AND KURTOSIS** 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.	**93-116**
UNIT 5:	**CORRELATION** Meaning of Correlation; Scatter Diagram and Correlation; Type of Correlation; Correlation Coefficient.	**117-130**
UNIT 6:	**REGRESSION** Meaning of Regression; Line of Regression and Regression Equation; Correlation and Regression Analysis: A Comparison.	**131-140**
UNIT 7:	**PROBABILITY** Random experiment; Definition of Probability; Elementary Theorems on Probability.	**141-159**
UNIT 8:	**CONDITIONAL PROBABILITY** Conditional probability; Multiplication theorem on probability; Total probability; Baye's theorem.	**160-180**

UNIT 9: RANDOM VARIABLES AND ITS PROBABILITY DISTRIBUTION **181-194**

Random Variable: Definition or Random Variable, Discrete Random Variable, Continuous Random Variable; Probability Distribution: Probability Mass Function, Probability Density Function; Mathematical Expectation and Variance: Mathematical Expectation, Properties of Mathematical Expectation, Variance of Random Variable, Properties of Variances, Illustrated Examples; Moments : Raw Moments, Central Moments.

UNIT 10: DISCRETE PROBABILITY DISTRIBUTION **195-220**

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.

UNIT 11: INDEX NUMBERS **221-245**

Concept and 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 12: TIME SERIES **246-263**

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 13: BASIC CONCEPTS IN SAMPLING **264-281**

Definition of Population, Parameter and Sampling; Random and non- random sampling; Sampling with and without replacement; Errors in Sampling.

UNIT 1: INTRODUCTION TO STATISTICS

UNIT STRUCTURE

1.1 Learning Objectives

1.2 Introduction

1.3 Meaning and Definition of Statistics

1.4 Frequency Distribution

- 1.4.1 Classification of Data
- 1.4.2 Types of Classification

1.5 Tabulation of Data

- 1.5.1 Advantages or Importance of Tabulation
- 1.5.2 Different Parts of Table
- 1.5.3 Types of Tables
- 1.5.4 Calculation of Class Frequencies and Formation of Frequency Distribution Table

1.6 Methods of Presenting Statistical Information

1.7 Let Us Sum Up

1.8 Further Readings

1.9 Answer To Check Your Progress

1.10 Model Questions

1.1 LEARNING OBJECTIVES

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

- define what is Statistics
- describe the uses of Statistics
- explain the concept of frequency distribution
- analyse classification and tabulation of data along with the importance of classification and tabulation
- explain graphical presentation of data.

1.2 INTRODUCTION

In this unit we will discuss the importance of statistics in business. It has been ascertained that the word 'statistics' has been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'statistic'. Each of these words means 'a political state'. In old days, the word 'statistics' was used to indicate the quantities like number of soldiers in a state, volume of arms and ammunition, volume of taxes, etc. connected with administrative and political activities of a state. In those days statistics was also called 'The Science of Kings'. But in modern age the scope of statistics is wide enough. Now- a- days the use of statistics is extended to various fields such as agriculture, economics, sociology, psychology, business, management, etc. Below we discuss certain definitions of statistics.

1.3 MEANING AND DEFINITION OF STATISTICS

The word 'statistics' has been defined differently by different authors in different times. This word conveys two different meanings: one when defined in plural sense and the other when defined in singular sense. In plural sense it refers to numerical statements of facts or data. In singular sense it refers to the subject statistics like other subjects such as mathematics, economics, physics, etc. For example, the data relating to the volume of production in different years, the volume of exports and imports in different years, the price of a commodity at different places, etc., refer to statistics used in plural sense. The word 'statistics' when used in singular sense implies statistical methods or the science of statistics.

Defining Data: Since the complexity of business environment makes the process of decision-making difficult thus for decision- makers it is important to develop the ability to extract meaningful information from raw data in order to make better decisions. Decisions have to be based upon data which show relationship between pair of variables, indicate trends and show rates of change in various relevant variables. Thus data can be defined as observations expressed in numerical figures obtained by measuring or counting is called data.

Below we mention some of the definitions of statistics as a subject.

1. "Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry." ..... Selligman
2. "Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty." ..... Wallis and Roberts.
3. "Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data." ..... Croxton and Cowden.

1.4 FREQUENCY DISTRIBUTION

A ***frequency distribution*** is a tabular presentation that generally organizes data into classes in terms of class intervals and shows the number of observations, called frequency, falling into each of these classes. An interval like 'a – b' (read as a to b) defining a class is called a class interval. The number of times an observation or value is repeated is called the ***simple frequency*** or frequency. On the other hand, the number of values included in a class interval is called the frequency of that class interval. Again when we add to the frequency of an observation (class interval), the frequencies of the previous observations (class intervals) then the sum is called the ***cumulative frequency*** of that observation (class interval).

A frequency distribution with one variable is called a univariate frequency distribution. For example, if the students in a class are classified on the basis of their height then it will be a univariate distribution. Here the variable (variate) is the height of the students. There are two other distributions called bivariate frequency distribution and multivariate frequency distribution.

Discrete Frequency Distribution: A discrete variable is one which takes isolated values or specific values. A discrete variable is the result of counting. A frequency distribution prepared for a discrete variable is called a discrete frequency distribution. It may be without class intervals or with class intervals.

An example of discrete frequency distribution without class intervals is given below:

Marks obtained by students (x)	30	33	45	55	63
No. of students (f)	5	7	4	5	6

The following is an example of discrete frequency distribution with class intervals:

Family size	1 - 2	3 - 4	5 - 6	6 - 7	7 - 8
No. of families	5	8	3	2	3

Although the above distribution looks like a continuous distribution yet is not so. The number of members in a family i.e. family size cannot be a fractional number like 1½, 3¾ etc. Hence the variable representing the family size is a discrete variable.

Continuous Frequency Distribution: A continuous variable is one which can take any value integral or fractional within a specified range of values. A continuous variable is the result of measurement. A frequency distribution prepared for a continuous variable in terms of class intervals is called a continuous frequency distribution. [A discrete frequency distribution **without class intervals** is called a simple or ungrouped frequency distribution and a frequency distribution (discrete or continuous) **with class intervals** is also called a grouped frequency distribution.]

An example of continuous frequency distribution is the following:

Height of students (in cm.)	120-130	130-140	140-150
No. of students	10	17	10
Height of students (in cm.)	150-160	160-170	
No. of students	11	9	

In the above example the variable is the height of students and it can take any value between 120cm. and 170cm. For example, the height of a student may be 120.17 cm., 133,255 cm, etc. Hence the variable representing the height of students is a continuous variable.

We have mentioned above that an interval defining a class is called a class interval. For example, in case of a frequency distribution like 0 - 9, 10 - 19, 20 - 29, ……or, 0 - 5, 5 - 10, 10 - 15, …….each is called a class

interval. The end numbers of a class interval are called class limits. The smaller number is called the ***lower class limit*** and the larger number is called the ***upper class limit***.

The length of a class interval is the difference between the lower limit of the class interval and the lower limit of the immediately succeeding class interval. It may be noted that 'length of class interval' is also called 'width of class interval', or 'size of class interval'. If in a frequency distribution, whether discrete or continuous, the length of each class interval is the same then the frequency distribution is said to be a frequency distribution with equal class intervals. For example, frequency distributions involving class intervals 0 - 9, 10 - 19, 20 - 29, 30 - 39, 40 - 49, 50- 59 and 0 - 10, 10 - 20, 20 - 30, 30 - 40, 40 - 50, 50 - 60, frequency distributions with equal class intervals, the length (or width) of the class intervals in each distribution being 10. On the other hand, if in a frequency distribution the lengths of all the class intervals are not the same then such a distribution will be called a frequency distribution with unequal class intervals. An example of such a frequency distribution is : 0 - 10, 10 - 20, 20 - 35, 35 - 55, 55 - 70.

In a discrete frequency distribution with class intervals, say, 0 - 9, 10 - 19, 20 - 29,……..both the lower and the upper limits of a class (defined by an interval) are included in that class. For instance, in the aforementioned frequency distribution both the lower and the upper limits of any class (class interval) are included in that class. Note that, both the figures 10 and 19 which are respectively the lower and the upper limits of the class defined by the class interval 10 - 19 are included in that class. Here the class defined by the interval 10 - 19 includes the numbers greater than or equal to 10 but less than or equal to 19. The method of classification of data in terms of class intervals in which both the lower limit and the upper limit of any class (class interval) are included in that class (class interval) is known as the inclusive method of classification. On the other hand, in case of a continuous frequency distribution like 0 - 10, 10 - 20, 20 - 30……..while the lower limit of a class is included in that class, the upper limit is not included. For example, the class defined by the interval 0 - 10 includes the numbers which are greater than or equal to zero but less than

10. The method of classification of data in terms of class intervals in which the lower limit of any class included in that class but the upper limit is not included is known as the ***exclusive method of classification.***

A frequency distribution with inclusive method of classification is called an ***inclusive frequency distribution*** and a frequency distribution with exclusive method of classification is called an ***exclusive frequency distribution.***

Thus, a discrete frequency distribution with class intervals is an inclusive frequency distribution and a continuous frequency distribution is an exclusive frequency distribution.

Class Boundaries or Real (True) Limits of a Class Interval: In most cases of measurement of continuous variables, all data are recorded nearest to some unit. For example, when the height of a person is recorded as 160cm., it means that the actual height of a person may be greater than or equal to 159.5cm, but less than 160.5cm. If the data recorded in integral values only are classified by inclusive method, viz., 155 - 159, 160 - 164, 165 - 169, etc., then the group 160 - 164 will include all persons with actual height being greater than or equal to 159.5cm. but less than 164.5cm. The limits 159.5 and 164.5 are called the lower and the upper boundaries or the lower and the upper real limits of the class defined by the inclusive class interval 160-164. Thus while 160 and 164 are respectively the lower and the upper limits of the inclusive class 160 - 164, then 159.5 and 164.5 are respectively called the lower and the upper boundaries of the same inclusive class 160 - 164. It is obvious that in this example the unit of the limits and the boundaries is 'centimetre'. It is to be noted that the lower boundary of any inclusive type class interval is 0.5 less than its lower class limit and upper class boundary is 0.5 higher than the upper class limit. In case of continuous frequency distribution which is an exclusive frequency distribution the upper limit of one class is the lower limit of the next class and in this case, class limits of a class coincide with the class boundaries of that class. Conversion of class limits to class boundaries is necessary when continuous variables recorded in discrete form are classified by using inclusive type class intervals. When class limits of an inclusive frequency

distribution are converted into class boundaries then the inclusive frequency distribution is converted into an exclusive frequency distribution.

Mid-Point or Mid-Value of a Class Interval: The mid-point or the mid-value of a class interval is the point or value exactly at the middle of the class interval. This point or value is obtained as follows:

Mid-point or Mid-value = ½ (Lower class limit + upper class limit)

or

Mid-point or Mid value = ½ (Lower class boundary + upper class boundary)

Thus for the class intervals 10 - 19 and 5 - 10, the mid-points or mid-values are = $\frac{10+19}{2}$ = 14.5 and = $\frac{5+10}{2}$ = 7.5 respectively.

The mid-value may also be calculated as:

Mid-value = Lower limit + ½ (upper limit - lower limit)

Mid-value = Lower boundary + ½ (upper boundary - lower boundary)

Open-end Distributions: An open-end distribution is one in which one or two classes lack one class limit. It is possible that the first class interval may not have the lower limit and the last class interval may not have the upper limit in a frequency distribution. Such a distribution is known as open-end distribution. We present below such as distribution:

Marks in Mathematics	Less than 10	10-20	20-30	30-40	40 and above
No. of students	10	17	15	5	7

In the above distribution the first class interval is presented as "less than 10" and hence the first class does not possess the lower limit. Again, the last class is presented as "40 and above" and hence there is no upper limit of this class. Each of these two class intervals is called an open end class interval.



CHECK YOUR PROGRESS

Q 1: What is frequency distribution?

Q 2: Identity the discrete and continuous frequency distribution.

(i)	x :	5 – 10	11 – 16	17 – 22	23 – 28
	f :	5	2	8	9
(ii)	x :	2 – 8	8 – 14	14 – 20	20 – 26
	f :	4	7	9	10

1.4.1 Classification of Data

Classification of raw data implies the arrangement of raw data into groups or classes on the basis of their similarities and resemblances. Below we quote two definitions of classification:

"Classification is the process of arranging data into sequences and groups according to their common characteristics, or separating them into different or related parts." - Secrist.

"A classification is a scheme for breaking a category into a set of parts, called classes, according to some precisely defined differing characteristics possessed by all the elements of the category". - Tuttle, A.M.

Classification makes the data understandable and it reveals the true significance of the data. In order to analyse the collected data classification is a must. For example, if we classify the data pertaining to the workers working in a factory with respect to gender (male and female), age (between 0 - 10 years, 10 - 20 years, etc.), qualification (below H.S.L.C., H.S.L.C. etc.) etc. then only the data become suitable for analysis.

Advantages or Importance of Classification:

(i) Classification helps in presenting the raw data in a coincise and simple form.

(ii) It facilities comparison by dividing the raw data on the basis of their similarities and resemblances.

(iii) It provides basis for tabulation and analysis of data.

(iv) Classification enables us to identify the possible characteristics in the data.

1.4.2 Types of Classification

There are four types of classification of data which we are going to discuss now–

(i) Geographical Classification: In geographical classification data are classified on the basis of places or geographical locations. For example, if we write the population figures of the different districts of Assam against the corresponding districts then such type of classification of population data will be called geographical classification. Series of data arranged on the basis of places or geographical locations is called spatial series.

Example of geographical classification:

City	: Mumbai	Kolkata	Delhi	Chennai
Population density: (per square km.)	654	685	423	205

(ii) Chronological classification: When the data are classified on the basis of time then it is known as chronological classification. Series of data arranged with respect to time is known as time series. If we write down the savings of a country in different years against the respective years then such type of classification of data relating to the savings of the country will be called chronological classification.

Example of chronological classification:

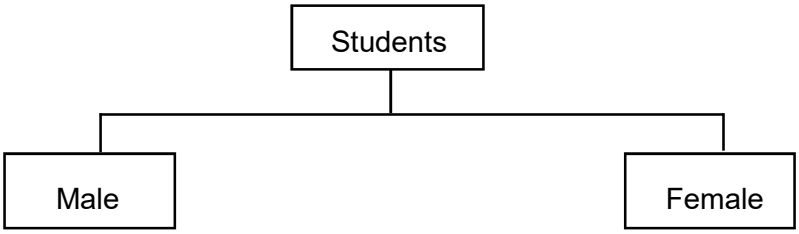
The following is an example of chronological classification

Year	: 1941	1951	1961	1971	1981	1991	2001
Population	: 31.9	36.9	43.9	54.7	75.6	85.9	98.6 (Crore)

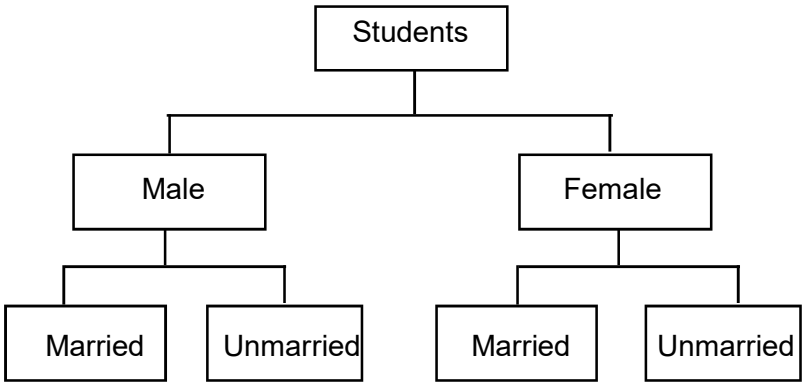
(iii) Qualitative classification or classification according to attributes: An attribute is a non-measureable characteristic like education, sex, religion, caste, etc. If items or individuals

are classified on the basis of one or more than one attributes then such a classification will be called **classification according to attributes**. There are two types of classifications according to attributes. If classification of items or individuals is made one the basis of a single attribute then such a classification is called **simple classification**. On the other hand, if items or individuals are classified on the basis of more than one attributes then such a classification is called **manifold classification**.

If we classify the students of a college on the basis of sex then it will be a case of simple classification. This is shown below:



Again, if we classify the workers in a factory on the basis of their sex and marital status as shown below then it will be a case of manifold classification:



(iv) Quantitative classification or classification according to variables or class intervals: If data are classified on the basis of the measureable characteristics called variables such as height, weight, marks, distance, etc. then such a classification will be called classification according to variables. If in this form

of classification data are organized into classes showing the number of observations falling in each class then such a classification may also be called 'classification according to class intervals'. In this form of classification data are presented in the form of a frequency distribution.

1.5 TABULATION OF DATA

The collected data, after classification, are recorded in rows and columns to give them tabular form. Tabular presentation of data, more conveniently known as tabulation, may be defined as "the orderly or systematic presentation of numerical data in rows and columns designed to clarify the problem under consideration and to facilitate the comparison between the figures".

1.5.1 Advantages or Importance of Tabulation

The following are the advantages of tabulation:

(i) Tabulation simplifies complicated data.

(ii) Tabulation presents quantitative data in a coincise and condenced form.

(iii) It facilitates comparison of data.

(iv) Presentation of data in tabular form provides a basis for analysis and interpretation of such data.

Objectives of Tabulation:

The objectives of tabulation are inherent in advantages of tabulation. Thus the objectives of tabulation are: (i) To simplify complicated data, (ii) To present quantitative data in a coincise and condenced form.

1.5.2 Different Parts of a Table

A table consists of the following parts. These may be considered as the essentials of a satisfactory table.

(i) Table Number: Every table should be identified by a number. It facilitates easy reference. The table number may be given at the beginning of the title of the table, or can be centered above the title of the table.

(ii) Title: A table must have a title which is to be written either below the table number or after the table number in the same line. The title should convey the full description of the contents in the table.

(iii) Stub: The extreme left hand column of the table which contains the headings of the rows is called stub.

(iv) Caption: Caption is the headings for the columns. It is the upper part of the table. There may be sub-heads or sub-captions in each caption.

(v) Body: It is the main part of the table containing the numerical figures.

(vi) Totals: The totals and sub-totals of all the rows and columns should be given in the table.

(vii) Footnote: Any explanatory note concerning the table itself, written directly beneath the table, is called 'footnote'. The purpose of footnote is to clarity some of the specific items given in the table.

(viii) Source: The source or sources of the data embodied in the table should be mentioned beneath the table if data are collected from secondary sources. It is given below the footnote.

1.5.3 Types of Tables

Tables are classified as : (i) Simple Tables and (ii) Complex Tables.

(i) Simple Tables: Tables which are prepared on the basis of only one characteristic of the collected data are known as simple tables. Simple tables are also called one-way tables.

Example 1.

Table

Department wise classification of 100 university students

Department	No. of Students
Physics	30
Chemistry	25
Mathematics	20
Statistics	25
Total	100

(ii) Complex Tables: Tables which show more than one characteristic of the data are known as complex tables. We may have a two-way complex table showing two characteristics, a three-way complex table showing three characteristics and a multi-way complex table showing several characteristics. A two-way, three-way and multi-way complex table may also be called a two-fold, three-fold and multi-fold complex tables respectively.

Example 2. Two-way complex table:

Table

Department wise classification of 100 university students

Department	No. of Students		
	Male	Female	Total
Physics	20	10	30
Chemistry	18	7	25
Mathematics	15	5	20
Statistics	12	13	25
Total	65	35	100

Example 3. Three-way complex table:

Table

Department wise classification of 100 university students

Department	Male			Female			Total		
	Hindu	Others	Total	Hindu	Others	Total	Hindu	Others	Total
Physics	16	4	20	8	2	10	24	6	30
Chemistry	14	4	18	6	1	7	20	5	25
Mathematics	13	2	15	5	0	5	18	2	20
Statistics	6	6	12	10	3	13	16	9	25
Total	49	16	65	29	6	35	78	22	100

Example 4. In a sample study about tea habit in two towns the following information was received.

Town A: Females were 40%; total tea drinkers were 45% and male non-tea drinkers were 20%.

Town B: Males were 55%; male non-tea drinkers were 30%, and female tea drinkers were 15%.

Supplying the missing information, tabulate the above data.

Solution: Let us assume that there were 100 people in each town. We now tabulate the given data below.

Department	Town A			Town B		
	Male	Female	Total	Male	Female	Total
No. of people with tea habit	40	5	45	25	15	40
No. of people without tea habit	20	35	55	30	30	60
Total	60	40	100	55	45	100

Explanation: Let us first consider town A. Since we have assumed total number of people to be 100, hence 40% female will imply 40 females, i.e., there are 40 females in town A. Therefore, the number of males in town A = 100 - 40 = 60. Again, the total number of people (including male and female) in town A with tea habit is 45% i.e., 45. Hence the number of people without tea habit is 20%, i.e., 20. Hence the number of females without tea habit = 55 - 20 = 35. Since the total number of males is 60 and the number of males without tea habit is 20 hence the number of males with tea habit = 60-20 = 40.

Similarly we find the data relating to town B as shown in the table.

1.5.4 Calculation of Class Frequencies and Formation of Frequency Distribution Table

The number of class intervals and the length of class intervals in connection with a frequency distribution to be prepared for raw data is to be determined on the basis of the range of the data. Usually the numbers of class intervals should not be made less than 5 or more than 15. The range of the data is the difference between the largest and the smallest data. The lower limit of the first class interval is usually taken at zero or at a multiple of 5 in case of non-negative observations. For instance, if the smallest observation is zero or a number greater than zero but less than 5 then in case of class intervals of length 5 the first class interval may be taken as 0 - 5 in case of exclusive frequency distribution or as 0 - 4 in case of inclusive frequency distribution. Similarly if the smallest observation in an integer greater than or equal to 30 but less than 40 and if the length of the class interval is chosen to be 10 (so that the number of class intervals lies between 5 and 15) then the first class interval is to be taken as 30 - 39 (in case of inclusive frequency distribution) or as 30 - 40 (in case of exclusive frequency distribution).

Following are the steps that are necessary in forming a frequency distribution table from raw data:

(i) To form the table of frequency distribution involving three columns;

(ii) To form the class intervals and to place them in the first column of the frequency distribution table;

(iii) To make a tally bar or tally mark (a vertical bar) for each observation in the second column (headed by Tally Bars) against the corresponding class interval;

(iv) To draw the fifth tally bar after four tally bars have been drawn in either way IIIII or IIII.

(v) To count the number of tally bars against all class intervals

after all the observations have been exhausted and to place these numbers against the respective class intervals in the third column (headed by the word 'frequency') of the frequency distribution table.

(vi) To add all the (simple) frequencies of the classes and to denote this sum by the letter N.

Example 5: In a particular subject 100 students of a particular class have got the following marks. Form a suitable frequency distribution.

75	80	66	56	50	82	66	45	68	56
80	69	87	74	80	76	81	64	57	68
58	60	73	74	66	87	84	98	79	64
80	63	58	87	96	77	72	74	79	78
86	66	76	76	80	82	63	75	83	34
76	68	74	41	68	40	59	90	35	94
65	88	85	84	79	87	76	65	76	80
75	82	96	95	73	60	52	80	88	63
76	83	60	79	72	77	57	92	67	52
72	86	72	86	73	78	78	79	75	60

Solution: [The above data are called ungrouped or unclassified or raw data. From these data it is not easy to answer such questions as (i) how many students have passed, (ii) how many students have got marks between 70 - 79, (iii) how many students have got letter marks.]

We observe from the given data that the maximum mark is 98 and the minimum mark is 34. For convenience we insert these numbers within two small circles. Now we form the class intervals as 30 - 39, 40 - 49, 50 - 59, ... keeping in mind that the number of class intervals should not be less than 5 and more than 15. (We may also form the class intervals as 30 - 40, 40 - 50, 50 - 60, ... When all the figures are integers we may form the class intervals either by inclusive method or by exclusive method). In case there are some fractional numbers then of course, we must form the class intervals by exclusive method. Now, following the steps in forming a frequency distribution as mentioned in 1.5.4, we get the following frequency distribution table.

Table

Mark wise distribution of students

Class interval	Tally bars	Frequency
30–39	||	2
40–49	|||	3
50–59	𝍸 𝍸 |	11
60–69	𝍸 𝍸 𝍸 𝍸	20
70–79	𝍸 𝍸 𝍸 𝍸	32
80–89	𝍸 𝍸 𝍸 𝍸 𝍸	25
90–99	𝍸 ||	7
	Total	N = 100

From the above table of frequency distribution we can very easily answer the following type of questions:

(i) How many students are getting marks between 40 and 49?

[Ans. 3]

(ii) How many students are getting marks between 70 and 79?

[Ans. 32]

(iii) How many students are getting marks between 50 and 79?

[Ans. 11+20+32=53]

(iv) How many students are getting marks 40 or more?

[Ans. 98]

(v) How many students are getting marks less than 60?

[Ans.16]

(vi) What is the percentage of students getting marks 80 or more?

[Ans. 32%]



CHECK YOUR PROGRESS

Q 3: State whether the following statements are true or false:

(i) An attribute can be measured.

(ii) A discrete variable can take any fractional value.

(iii) A continuous variable may assume any value between two specified limits.

(iv) The smallest possible measurement in a class is called the lower limit of that class.

(v) Class limits and class boundaries of an exclusive class are the same.

(vi) Class limits and class boundaries of an inclusive class are the same.

(vii) An open-end distribution creates problems in the calculation of mid-points.

(viii) The mid-point of a class interval is equal to the lower limit of that class plus half the difference between the lower limit and the upper limit of that class.

(ix) A discrete variable is the result of counting.

(x) A continuous variable is the result of counting.

1.6 METHODS OF PRESENTING STATISTICAL INFORMATION

Diagrams and graphs are the methods which are devised to present huge mass of quantitative data so that they become at once comparable and give pleasure to the eye besides being interesting and easily understandable.

Usually data are presented with the help of the following diagrams:

(a) Bar Diagram (Chart)

(b) Rectangular and Square Diagrams (Chart)

(c) Circular and Pie Diagrams (or Pie Chart)

We shall limit our discussions to various types of bar diagrams as well as pie diagram only as these are the most commonly used diagrams.

(a) Bar Diagram: In bar diagram only the length is considered, the breadth may be of any finite magnitude. The bar diagrams are divided into the following three categories:

(i) Simple Bar Diagram

(ii) Multiple Bar Diagram

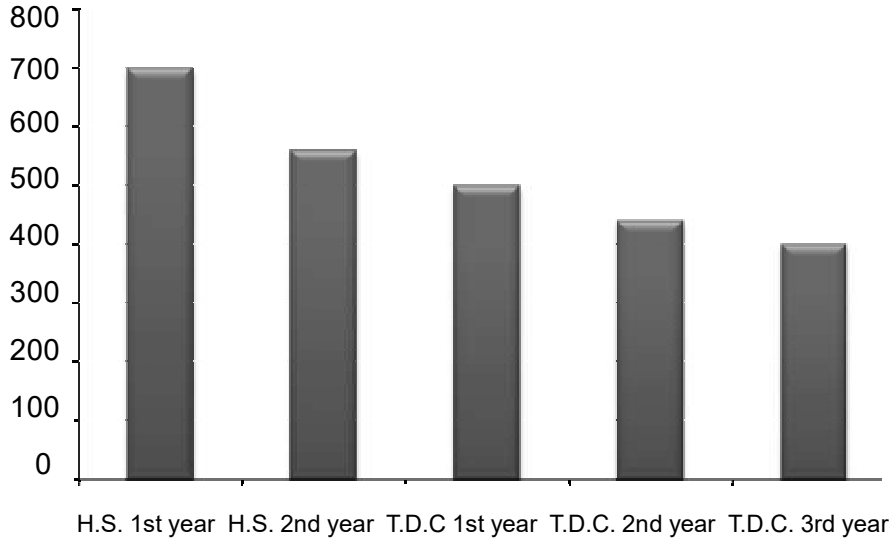
(iii) Sub-Divided or Compound Bar Diagram

(i) Simple Bar Diagram: Only one type of data are presented with the

help of simple bar diagram. For example, the volume of production of rice in Assam during the last five years can be presented with the help of simple bar diagram. In order to draw simple bar diagram, a bar is drawn for each datum. All the bars are on the same general base. The heights of the bars will be as per the magnitudes of the data. The breadth of each bar must be same and the gaps among the bars must be uniform. Generally the gap between two consecutive bars should not be less than half the breadth of a bar. The bars are drawn either on a common horizontal or on a common vertical base. Data can be easily compared with the help of the heights of the bars.

Example 1. Draw a vertical bar-chart for the number of students of a college: H.S. 1st year - 700, H.S. 2nd Year - 560, TDC 1st year - 500, T.D.C. 2nd year - 440 and T.D.C. 3rd year - 400

**** Solution:*** In order to represent the students of the various classes of the college the following vertical simple bar diagram has been drawn:

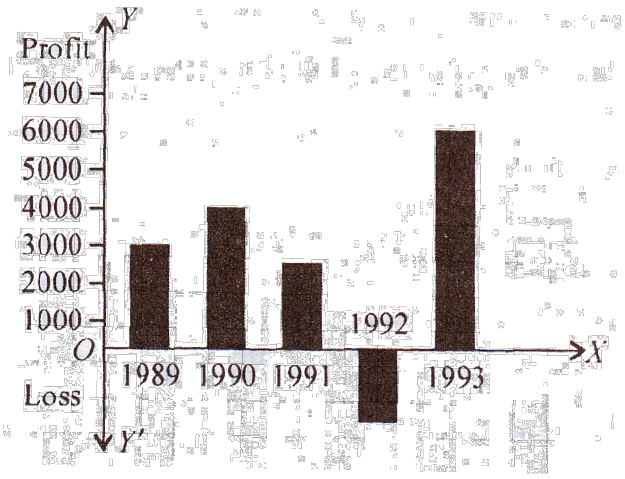


Example 2: The profits and losses of a business concern for the years 1989-1903 are given below:

Year :	1989	1990	1991	1992	1993
Profits (Rs.) :	3000	4000	2500	--	6000
Loss (Rs.) :	--	--	--	2000	--

Represent the above data by a bar-chart

Solution: In order to represent the profit and loss of the business concern we draw the following simple bar diagram.



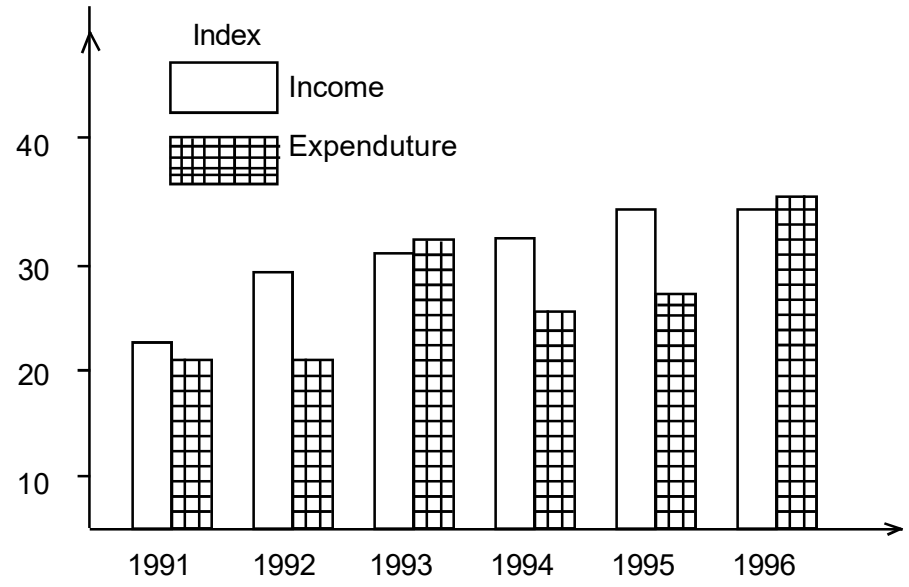
(ii) Multiple Bar Diagram: With the help of multiple bar diagram (multiple bar chart) more than one type of data can be depicted at a time side by side to represent 2, 3 or 4 series of values for comparison. For example, in order to show the profits and losses of a business concern for a few years, a pair of bar diagrams indicating profit and loss of each year are to be drawn side by side. For convenience the adjacent bars are differently shaded. In order that the diagram becomes distinct, gap is maintained between different groups of adjacent bars.

Example 3: The income and expenditure of a business concern (in thousand rupees) have been given below:

Year	1991	1992	1993	1994	1995	1996
Income	22.0	27.3	28.2	30.3	32.7	33.3
Expenditure	19.5	21.7	30.0	25.6	26.1	34.2

Represent the given data by a multiple bar-diagram.

Solution: The income and expenditure of the business concern are represented by a multiple bardiagram as shown below:



Scale : Length of one small square on the

Y – axis = Rs. 2000

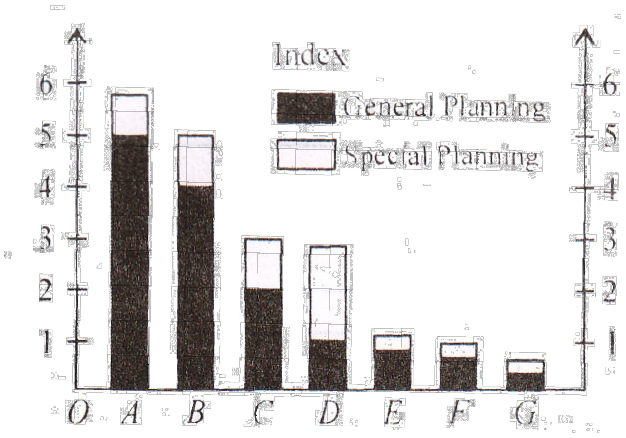
(iii) Sub-divided Bar Diagram: Sub-divided bar diagrams are used to present data which are to be shown in parts or which are totals of various subdivisions. Here a bar is drawn indicating the total corresponding to a particular item and then the bar is sub-divided into various parts or components in proportion to the various subdivisions of the quantity of the item. To distinguish various components from one another, different colours or shades may be given. The diagram may be drawn on absolute figures or percentages.

Example 4: The list of expenditure on various departmental heads during the Fourth Five Year Plan of a certain state is given below. Represent these data by a sub-divided bar diagram.

	Items	General planning (Rs in lakh)	Special Planning (Rs. in lakh)	Total
A.	Irrigation and Power	5074	226	5300
B.	Social Welfare	3647	1245	4892
C.	Agriculture	2170	1030	3200
D.	Communication and Transportation	1075	1740	2815
E.	Industry and Minerals	1210	100	1310

F.	Cooperatives	738	377	1115
G.	Miscellaneous	86	212	368

* ***Solution:*** The following sub-divided bar diagram is drawn for the given data.



A : Irrigation and Power,

B : Social Welfare,

C : Agriculture,

D : Communication and Transportation,

E : Industry and Minerals,

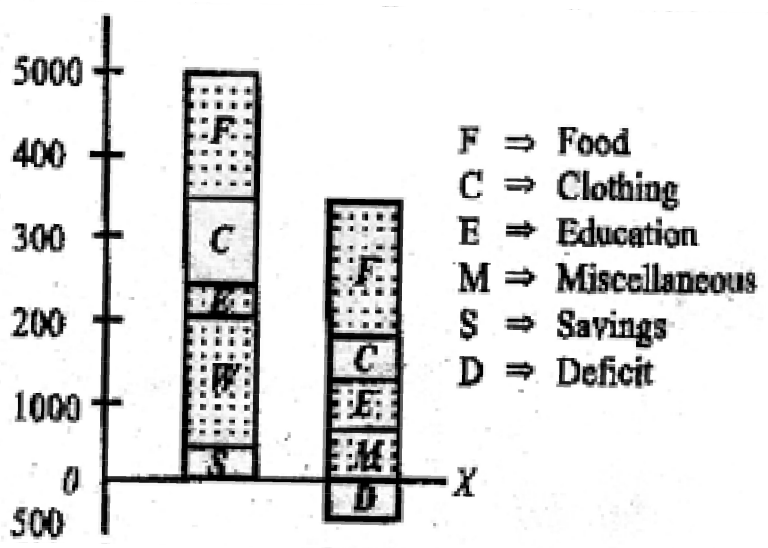
F : Cooperatives,

G : Miscellaneous.

Example 5: Represent the following data by a suitable diagram.

Items of expenditure	Family A	Family B
	(Income Rs. 5000)	(Income Rs. 3000)
Food	1500	1500
Clothing	1250	600
Education	250	500
Miscellaneous	1900	700
Savings or Deficit	+ 100	- 300

* ***Solution:*** The given data may be represented by a sub-divided bar diagram as shown below:



Pie Diagram (Pie Chart): A pie-diagram is a circle of radius neither too larger nor too small whose area is divided into as many different sectors as there are components of the whole data. This is done by drawing straight lines from the centre to the circumference of the circle. The area of the circular lamina represents the whole data and it is equivalent to 360^0 at the centre. The area of each sector is proportional to the value of the corresponding components of the data. The area of a sector is proportional to the angle at the centre.

Pie diagram is very useful in drawing comparison among the various components or between a part and the whole.

Example 6: The proposed expenditures on various items during a five year plan of a government are as follows:

Head	Amount of money (Rs. in crores)
Agriculture	12,000
Industry and Mines	9,000
Irrigation and Power	6,000
Education	8,000
Communication	5,000

Draw a pie chart for the data.

Solution: We have calculated below the various angles of pie-diagram:

Expenditure heads	Amount of expenditure (Rs. in crores)	Magnitude of angle (in degree)
Agriculture	12,000	108
Industry and Mines	9,000	81
Irrigation and Power	6,000	54
Education	8,000	72
Communication	5,000	45
Total	40,000	360

Calculation of angles:

Angle representing Rs. 40,000 = 360^0

$\therefore$ " " Rs. 1,000 = $\frac{360}{40} = 9^0$

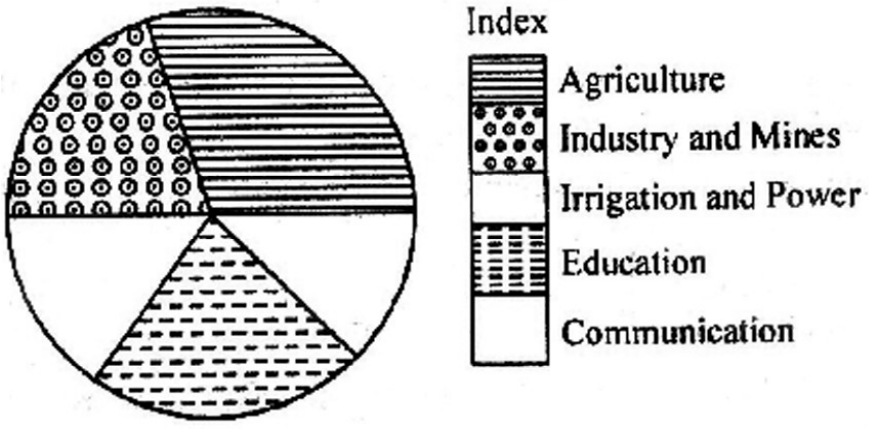
$\therefore$ " " Rs. 12,000=9 x12=108^0

" " Rs. 9,000=9 x 9=81^0

" " Rs. 6,000=9 x 6=54^0

" " Rs. 8,000=9 x 8=72^0

" " Rs. 5,000=9 x 5=45^0



Example 7: In an election 72,000 votes were casted. Out of four candidates the first got 24,000 votes, the second got 20,000 votes, the third got 18,000 votes and the fourth got 10,000 votes. Draw a pie-chart for these data.

*** Solution:** Let, the total of 72,000 votes be represented by 360^0.

The angle corresponding to 1 vote = $\left(\frac{360^0}{72,000}\right) = \left(\frac{1^0}{200}\right)$

The angle corresponding to 24,000 votes got by the

1st candidate = $\left(\frac{1}{200}\times 24{,}000\right)$ = 120^0

The angle corresponding to 20,000 votes got by the

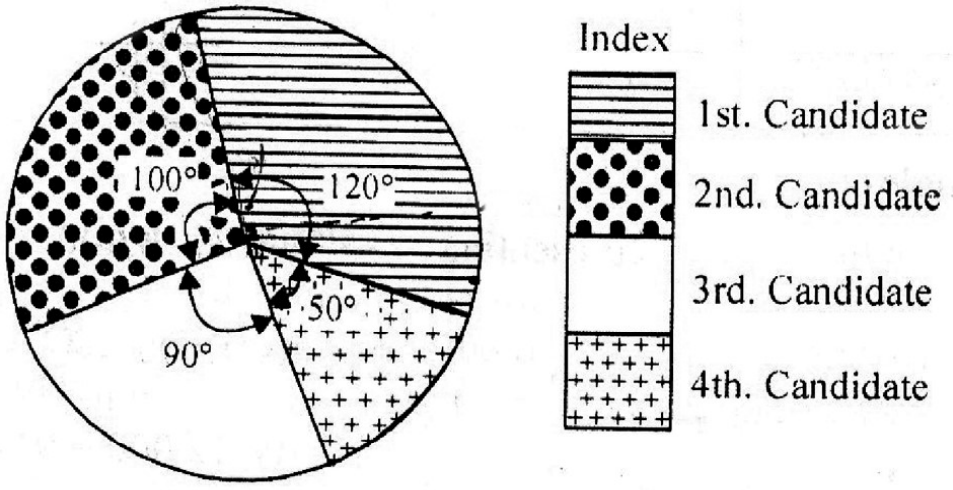
2nd candidate = $\left(\frac{1}{200}\times 20{,}000\right)$ = 100^0

The angle corresponding to 18,000 votes got by the

3rd candidate = $\left(\frac{1}{200}\text{x}18{,}000\right)$ = 90^0

The angle corresponding to 10,000 votes got by the

4th candidate = $\left(\frac{1}{200}\text{x}10{,}000\right)$ = 50^0.



Advantages of Diagrams:

(i) Diagrammatic representation of statistical data makes it easy to understand the characteristics of such data.

(ii) A diagram or chart sometimes clarifies a complex problem and other reveals hidden facts which cannot be otherwise easily detected.

(iii) Many complex theories of economics can be easily explained with the help of diagrams.

Limitations of Diagrams:

(i) Diagrams do not show details. In a table one may show information covering a large number of items or on related topics by providing

additional rows and columns. This is not possible in case of a diagram.

(ii) In a table statistical data can be shown exactly. But statistical data can be shown only approximately in a diagram.

Uses of Diagrams:

Since representation of statistical data by diagrams become appealing to the eyes and easily understandable as such diagrammatic representation of data has gained much importance in Economics, Business, and Commerce. Diagrams have become indispensable in various advertisements published by various business establishments. For busy traders diagrams are very important because they do not find time to study the characteristics of statistical data by other means. They can have a bird's eye view of the business situation by having a look on the diagrams.

Presentation of Data by Graphs : Frequency distributions can be described with the help of graphs. Graph is more distinct in comparison to diagram and it is comparatively easier to draw graph than diagram. Data of frequency distributions and time series data are generally represented by graphs. Generally diagrams are drawn to make comparative study. But diagrams do not depict the relationships among the variables to be studied. With the help of graphs the relationships among the variables can be studied. The importance of graph is more than that of diagram in statistical analysis.

In the practical field various types of graphs are used. These graphs can broadly be divided into two groups:

(a) Graphs of frequency distribution

(b) Graphs of time series

Graphs of frequency distribution: In order to make frequency distribution easily understandable three types of graphs are usually drawn. These are:

(i) Histogram or column diagram

(ii) Frequency Polygon

(iii) Cumulative frequency curve or Ogive.

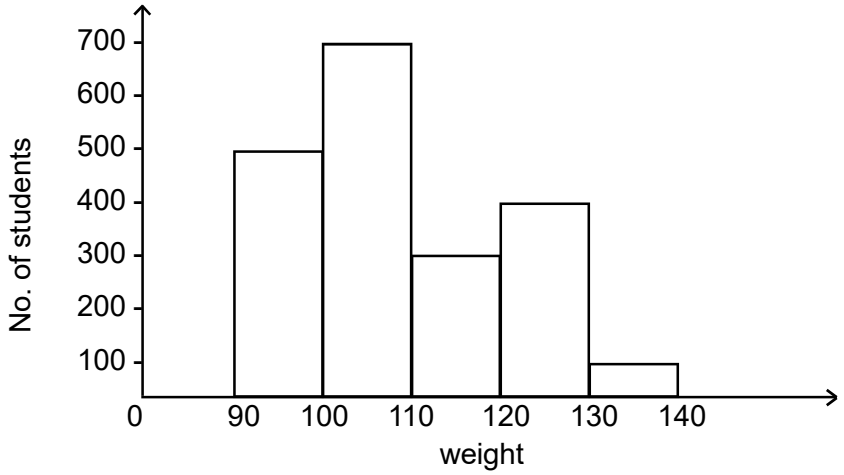
(i) Histogram or column diagram: The graph by which the frequencies of various class intervals of a frequency distribution with the help of adjacent vertical rectangles are shown is called a histogram or column diagram.

First of all, the actual class-intervals are to be marked on the x-axis choosing a suitable scale. Then taking these as bases, rectangles are to be drawn continuously on this basis. When the class intervals are equal then the heights of the rectangles will be proportional to the frequencies of the corresponding class intervals, but if the class intervals are not equal then the heights of the rectangles will be proportional to the ratios between the corresponding class frequencies and the lengths of the class intervals.

Example 8: Draw a histogram to represent the following distribution:
(Frequency distribution of weights of 2000 students)

Weights (in lbs)	90-100	100-110	110-120	120-130	130-140
No. of students	500	700	300	400	100

*** Solution:** The class intervals of the above frequency distribution are of equal length. First of all, the class intervals are marked on the X-axis and then rectangles of heights proportional to the frequencies of the class intervals are drawn on them.



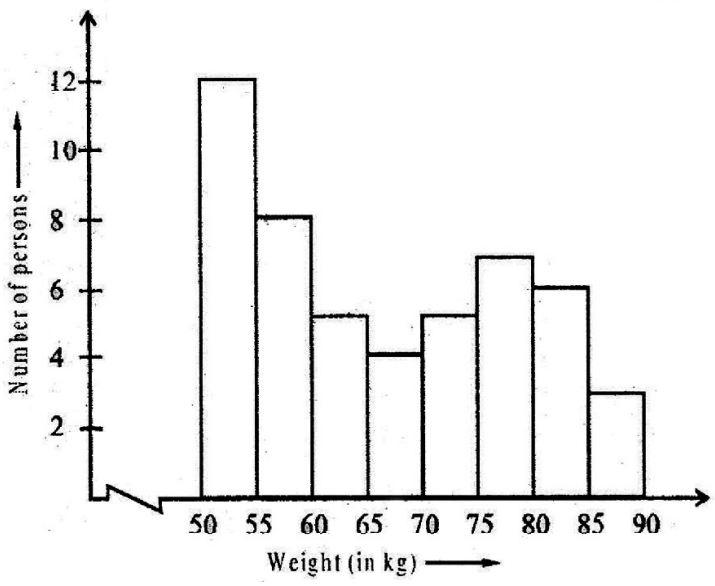
Example 9: The following is the distribution of weights (in kg) of 50 persons:

Weights (in kgs)	50-55	55-60	60-65	65-70	70-75	75-80	80-85	85-90
Number of persons	12	8	5	4	5	7	6	3

Draw a histogram for the above data.

*** Solution:** We represent the class limits along the X-axis on a suitable scale and the frequencies along the Y-axis on a suitable scale.

Since the scale on the X-axis starts at 50, a kink (break) is indicated near the origin to signify that the graph is drawn to scale beginning at 50, and not at the origin.



(ii) Frequency Polygon: The graph obtained by joining the mid-points of the upper horizontal sides of the adjacent rectangles of a histogram by line segments is called a frequency polygon. Frequency polygon can be drawn directly without obtaining indirectly from histogram. For this we are to take the mid-points of the class intervals on the X-axis and the frequencies of the class intervals are to be plotted against the corresponding class mid-points along the Y-axis. Then these plotted points are to be joined by line segments. Then the end points of the graph so obtained are to be joined with the mid-points of the two frequency less class intervals which preceed and succeed the first and the last class intervals respectively. They graph obtained thereby is the desired frequency polygon.

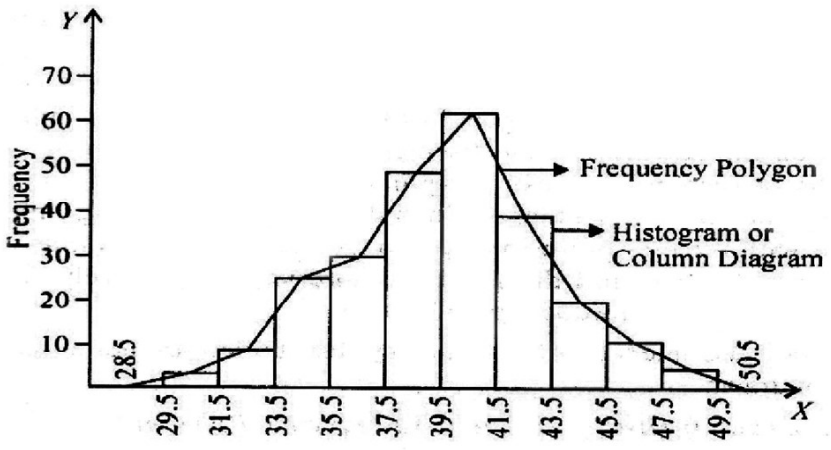
Note: In case the class intervals of a frequency distribution are not all equal then frequency polygon is usually not drawn.

Example 10: Draw (i) Histogram and (ii) Frequency Polygon for the following frequency distributions:

Weekly Wages (Rs.)	No of Workers
30-31	2
32-33	9
34-35	25
36-37	30
38-39	49
40-41	62
42-43	39
44-45	20
46-47	11
48-49	3

*** Solution:** The following table is to be prepared before drawing histogram and frequency polygon.

Class intervals	Real class limits	Mid-point	Frequency
30-31	29.5-31.5	30.5	2
32-33	31.5-33.5	32.5	9
34-35	33.5-35.5	34.5	25
36-37	35.5-37.7	36.5	30
38-39	37.5-39.5	38.5	49
40-41	39.5-41.5	40.5	62
42-43	41.5-43.5	42.5	39
44-45	43.5-45.5	44.5	20
46-47	45.5-47.5	46.5	11
48-49	47.5-49.5	48.5	3



Note: Here we have drawn the frequency polygon by joining the mid-pints of the upper horizontal sides of the rectangles by line segments. In order to draw the frequency polygon directly we are to take the mid-points of the class intervals on the X-axis and the frequencies of the corresponding class intervals on the Y-axis. Then we are to plot the points (30.5, 2), (32.5, 9), ……(48.5, 3) on graph paper and then we are to join these points by line segments. In both the methods, the end points of the graph so drawn are to be joined with the mid-points of the class intervals preceeding the first class and succeeding the last class. (The frequency of each of these two class intervals will be zero). The graph so obtained is the frequency polygon to be drawn.

(iii) Cumulative Frequency Curve or Ogive: The graphical representation of a cumulative frequency distribution is known as the Cumulative frequency curve or ogive. First of all, from the given raw data or from the given frequency distribution, a cumulative frequency table is constructed. Then the upper class limits (or the upper class boundaries) of the class intervals are marked on the X-axis and the points are plotted by showing the corresponding cumulative frequencies on the Y-axis. Then these points are connected by drawing smooth curve with free hand. This curve is the cumulative frequency curve or the ogive. Depending upon the pattern of drawing the graph we get two types of ogives namely, less-than ogive and more-than ogive. The methods of drawing less-than ogive and more-than ogive are explained below. With the help of ogive one can determine median, quartiles, deciles etc.

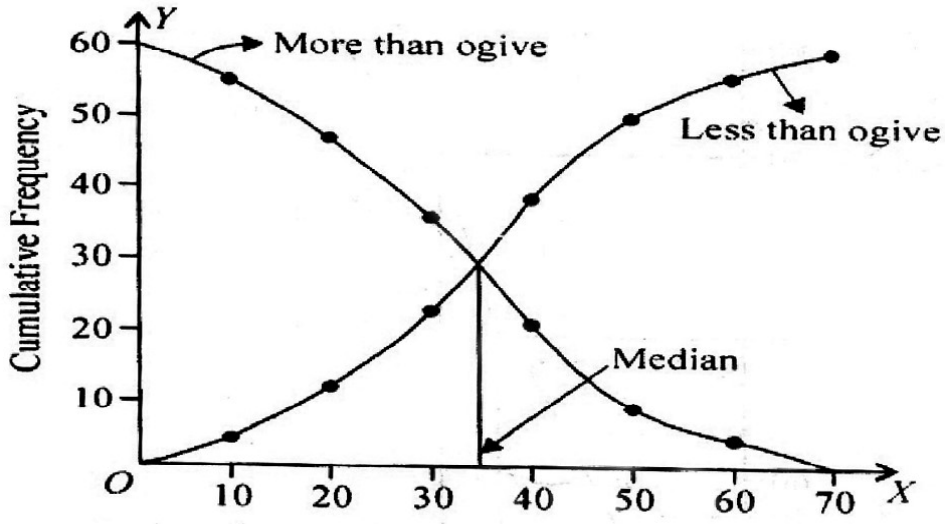
Example 11: Draw (i) less than ogive and (ii) more than ogive for the following frequency distributions:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	4	8	11	15	12	6	3

**** Solution:*** The class intervals of the given frequency distribution are as per the exclusive method. Hence the lower and the upper limits of that class.

Intervals	Frequency	Cumulative frequency	
		Less-than	More-than
0-10	4	4	59
10-20	8	12	55
20-30	11	23	47
30-40	15	38	36
40-50	12	50	21
50-60	6	56	9
60-70	3	59	3

From less-than cumulative frequency column we find that the cumulative frequency of a particular class implies the number of observations included upto the upper limit of that class. For example, the cumulative frequency of the class 30-40 is 38. This means that there are 38 observations less than 40. Again, from the more than cumulative frequency column we find that the cumulative frequency of a class implies the number of observations of the given distribution greater than or equal to the lower limit of that class interval. Thus, the cumulative frequency 59 of the first class (shown by 0-10) tells us the fact that the number of observations of the given distribution greater than or equal to 0 is 59. In case of less than ogive, upper class limits of the class intervals are marked on the X-axis and then points are plotted by taking the corresponding cumulative class frequencies on the Y-axis. The points are joined by a smooth curve with free hand. Then the lower end point of the curve is joined with the lower limit of the first class (class interval). By looking at the graph of the more-than ogive drawn below students can easily guess the procedure of drawing a more than ogive.



In order to get the median of the distribution, a perpendicular is to be drawn on the X-axis from the point of intersection of the less-than ogive and the more-than ogive. The distance between the origin and the foot of the perpendicular upon the X-axis is the median. In case of the given frequency distribution we get the median to be approximately equal to 35.6 marks.

Note: The number of observations less than 0 is zero. Hence the lower end point of the less-than ogive is at the origin. Again, the number of observations greater than 70 is also zero. Hence the lower end point of the more-than ogive is at the point corresponding to 70 on the X-axis.

Example 12: The following data relate to sales of 100 companies. Draw the two ogives. Determine the number of companies whose sales are (i) les than Rs, 13 lakhs, (ii) more than Rs. 36 lakhs.

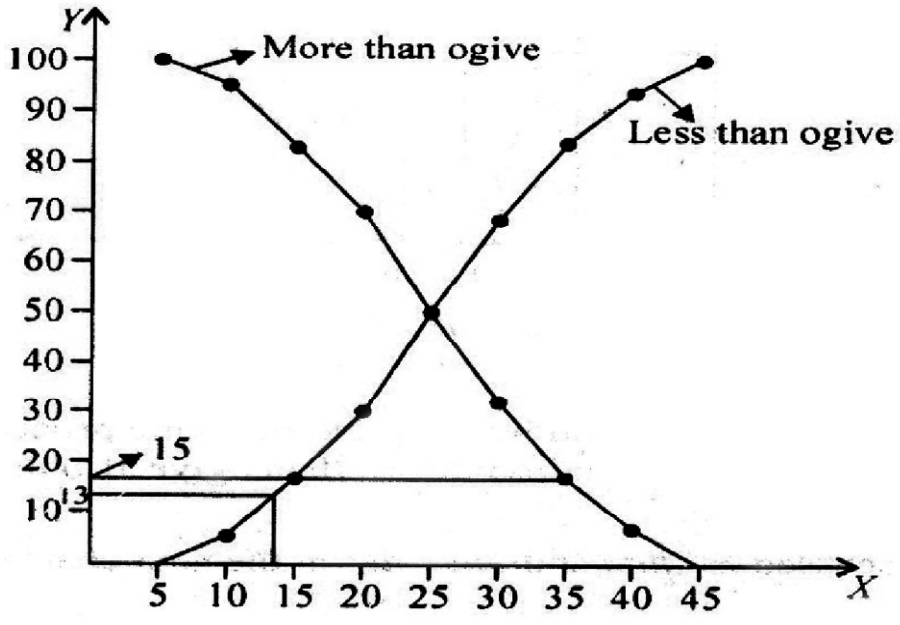
Sales (Rs. in lakhs)	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
No. of Companies	5	12	13	20	18	15	10	7

Solution: The class intervals of the given frequency is as per the exclusive method of classification and hence the lower and the upper limits of the class intervals may be taken as the actual lower and upper limits.

Class Intervals	Frequency	Cumulative frequency	
		Less-than	More-than
5-10	5	5	100
10-15	12	17	95
15-20	13	30	83
20-25	20	50	70
25-30	18	68	50
30-35	15	83	32
35-40	10	93	17
40-45	7	100	7

In case of less-than ogive the cumulative frequencies are plotted on a graph paper against the upper limits of the corresponding classes and in case of more than ogive the cumulative frequencies are plotted on a graph paper against the lower limits of the corresponding classes.

Taking the length of each small square on the X-axis as one unit and the length of each small square on the Y-axis as 5 units, the two ogives (Less-than and More than) have been drawn. It has been found that (i) the number of companies with sale proceeds more than 13 lakhs = 13 and (ii) the number of companies with sale proceeds more than 36 lakhs = 15.





CHECK YOUR PROGRESS

Q 4: What will be the most suitable diagrams to represent the following sets of data?

(i) Plan outlays for Assam in different five year plans;

(ii) Birth rates and death rates in different states of the North Eastern region according to 2001 census

(iii) Sectoral composition of India's GDP in a given year.

Q 5: Fill in the blanks with suitable word/words:

(i) Frequency polygon can be drawn with the help of _____.

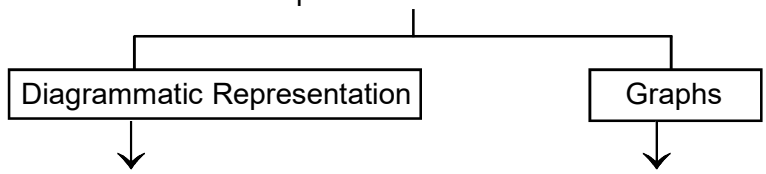
(ii) The ogives of less than type and more than type distributions intersect at_____.

(iii) Pie-chart is always_____.



1.7 LET US SUM UP

- Statistics is the numerical representation of data.
- Statistics can be used in economics, insurance companies, business, banks etc.
- Frequency distribution is a tabular presentation of data.
- Data can be classified on geographical, chronological, qualitative and quantitative factors.
- Statistical data representation



(a) Bar Diagram
(b) Pie Diagram

(a) Histogram
(b) Frequency Polygon
(c) Cumulative Frequency Curve



1.8 FURTHER READINGS

1) Business Mathematics, Qazi Zameeruddin, V K Khanna, and S K Bhambri, Vikas Publishing House Pvt Ltd, Second Reprint, 1995.

2) Agarwal, D. R. (2006). Business Statistics. Delhi, Vrinda Publications.

3) Gupta S.C. (1994). Fundamentals of statistics. New Delhi, Himalayan Publishing House.

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

5) Sharma, J. K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

6) Verma, A. P. (2007). Business Statistics. Guwahat, Asian Books Private Limited.



1.9 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Frequency distribution is a tabular presentation of data into classes in terms of class intervals.

Ans to Q No 2: (i) Discrete (ii) Continuous.

Ans to Q No 3: (i) F, (ii) F, (iii) T, (iv) T, (v) T, (vi) F, (vii) T, (viii) T, (ix) T, (x) F.

Ans to Q No 4: (i) Simple bar diagram, (ii) Multiple bar diagram, (iii) Pie diagram,

Ans to Q No 5: (i) Histogram, (ii) Median, (iii) Circular.



1.10 MODEL QUESTIONS

Q 1: Define frequency distribution. Explain the construction of a frequency distribution.

Q 2: Marks obtained by 50 students are given below. Prepare a frequency distribution table by taking class intervals each of length 10 by inclusive or exclusive method:

34	54	10	21	51	52	12	43	48	36
48	22	39	26	34	19	10	17	47	28
13	30	30	60	15	59	07	18	40	49
40	51	55	32	41	22	30	35	53	25
14	18	19	40	43	04	17	45	25	43

Q 3: Age at death of 50 persons of a town are given below:

36	48	50	45	49	31	50	48	43	42
37	32	40	39	41	47	45	39	43	47
38	39	37	40	32	52	56	31	54	36
51	46	41	55	58	31	42	53	32	44
53	36	60	59	41	53	58	36	38	60

(a) Arrange the data in frequency distribution having 10 class intervals.

(b) Obtain the percentage frequency in each class interval.

Q 4: Construct a frequency distribution table for the following data. Answer the following questions from the table:

(i) P.C. of students securing more than 30 marks,

(ii) P.C. of students securing less than 30 marks,

Marks secured: 5, 27, 23, 32, 49, 19, 50, 38, 37, 25, 30, 29, 42, 37, 18, 12, 13, 9, 18, 27, 50, 32, 47, 48, 29, 30, 32, 37, 27, 41.

Hint: Here range = 50 - 5 = 45. Length of class intervals = = 4.5 i.e., 5. The class intervals are: 5 - 10, 10 - 15,……50 - 55. No. of students getting marks more than 30 = (5+4+2+3+2) - 2 = 14. [Ans. (i) 46.67%, (ii) 46.67%]

Q 5: What do you mean by classification of data? What are the objectives of classification?

Hint: For objectives of classification see advantages of classification. You are to write down the objectives as: (i) to present the raw-data in a concise and simple form, (ii) to facilitate comparison by dividing the raw-data on the basis of their similarities and resemblances, etc.

Q 6: Discuss the different types of classification.

Q 7: Distinguish between classification according to attributes and classification according to variables.

Q 8: Discuss the importance of classification in statistical analysis.

Q 9: What is tabulation? What are its objectives?

Hint: For objectives see advantages of tabulation. You are to write down the objectives as:

(i) To simplify complicated data, (ii) To present quantitative data in a concise and condensed form, (iii) To facilitate comparison of data, (iv) To provide basis for analysis and interpretation of data.

Q 10: In a sample study about the tea habits in two villages the following data were obtained:

Village A: 70% persons were males. 80% were tea drinks and 62% were male tea drinkers.

Village B. 55% persons were males. 35% were tea drinkers and 52% were male tea drinkers. Tabulate the above supplying the missing link.

Q 11: Prepare a table from the data given below:

In 1987, out of a total of 1,750 workers of a factory 1200 were members of a trade union. The number of women employees was 200, of which 175 did not belong to the trade union.

In 1988, the number of union members increased to 1580, of which 1290 were men.

On the other hand, the number of non-union workers fell down to 208, of which 180 were men.

Q 12: What are the different types of diagrams commonly used for the presentation of statistical data?

Q 13: What is graphical representation of data? What are the various kinds of such graphs?

Q 14: Write short notes on (a) Histogram, (b) Frequency polygon, (c) Pie-diagram.

Q 15: The profits and losses of a business concern (Rs. in lakhs) are given below:

Year	1999	2000	2001	2002	2003
Profit (Rs. in lakhs)	36	-	22	38	-
Loss (Rs. in lakhs)	-	10	-	-	15

Represent the above data by a bar-chart.

Q 16: The following table shows total cost (in Rs.) and its component parts in two consecutive years:

	1990	1991
Direct material	4,000	5,500
Direct labour	5,000	6,000
Direct expenses	1,200	1,500
Overheads	2,500	2,000

Represent the data by a suitable diagram.

Hint: The data may be presented by a sub-divided bar-diagram.

Q 17: Represent the following data by means of a pie-diagram. The data relate to the expenditures of a state during the Second Five Year Plan (Rs. in crores) on different heads:

Industry	127.00
Irrigation	92.50
Agriculture	100.0
Transport and communication	92.50
Others	480.00

Q 18: A table of weekly wages of some tea-garden workers is given below. Draw a histogram for these data.

Salary (Rs.)	200-210	210-220	220-230	230-240	240-250
No. of employees	8	15	21	35	62

*** ***** ***

UNIT 2: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE

2.1 Learning Objectives

2.2 Introduction

2.3 Measures of Central Tendency

- 2.3.1 Definition
- 2.3.2 characterstics of a good average
- 2.3.3 '$\sum$ ' Symbol
- 2.3.4 Different Types of Measures of Central Tendency

2.4 Let us Sum Up

2.5 Further Readings

2.6 Answer to Check Your Progress

2.7 Model Questions

2.1 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

2.2 INTRODUCTION

In unit 1 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.

2.3 MEASURES OF CENTRAL TENDENCY OR AVERAGES

2.3.1 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.

2.3.2 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 algebraic treatment.

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

2.3.3 '$\sum$' Symbol

In order to denote sum (i.e., a total of certain quantities) the Greek letter '$\sum$' (capital sigma) is used. For example, if variable x takes the values x_1, x_2, x_3,……..x_n then the sum of these values of the variable i.e. (x_1, x_2, x_3,……..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'.

2.3.4 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.

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

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

In case of ungrouped frequency distribution table

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

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

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

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_2 + \cdots + f_n x_n}{f_1 + f_2 + f_3 + \cdots + f_n}$$

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

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

where $\mathrm{N} = f_1 + f_2 + f_3 + \cdots + f_n$

- **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 frequencydistribution $\mathrm{A.M.}(\bar{x})$ is defined as

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

where $\mathrm{N} = f_1 + f_2 + f_3 + \cdots + f_n = \sum f$ 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 Σfx = 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 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

$$Mean\,(\bar{x}) = \frac{\sum fx}{N} = \frac{1670}{50} = 33.4$$

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	Frequency	d=x-A	$d' = \frac{d}{h}$	fd'
	(x)	(f)	A=147	h=5	
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, $\text{A.M.}(\bar{x}) = A + \frac{\sum fd'}{N} \times h$

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

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

=147–1.65

= 145.35

i.e. the required A.M. = **145.35c.m.**

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 $\sum(x - \bar{x}) = 0$ where $\bar{x}$ = Σx/n and in case of frequency distributions $\sum f(x - \bar{x}) = 0$ where $\bar{x}$ = Σfx/N, N= Σf.

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

$\Rightarrow \bar{u} = 2.5\bar{x} = 2.5x2.5 = 6.25$

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

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

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

Example 6: Calculate Arithmetic 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 Arithmetic 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		$\Sigma fx = 695.0$		$\Sigma fd = -70$

By short cut method

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

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

= 25.5–2.33

= 23.17 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

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

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

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



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:

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

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.

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 $\frac{N}{2}+1$ = 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.

$$\text{Median (Me)} = L+\frac{\frac{N}{2}-f_C}{f}\times I$$

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

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

Here L = $65,\ f = 30,\ f_c = 38,\ N = 111,\ 1 = 5$

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

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

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

$$= 65 + 2.92$$

$$= 67.92$$

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.

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.

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

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 l = Length of the model class

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

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

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

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

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

= 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. Karl 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:	12	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



2.4 LET US SUM UP

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



2.5 FURTHER READINGS

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.



2.6 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

2.7 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 3: MEASURES OF DISPERSION

UNIT STRUCTURE

3.1 Learning Objectives

3.2 Introduction

3.3 Dispersion

3.3.1 Different Measures of Dispersion

3.4 Let us Sum Up

3.5 Further Readings

3.6 Answer 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 dispersion or variation
- describe the various methods of obtaining dispersion for various types of distribution

3.2 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.

3.3 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".

3.3.1 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 absolute 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 = 39cm.

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 Series

Inrerquartile Range = $Q_3 - Q_1$

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

Where Q_1 = size of $\left(\frac{n+1}{4}\right)^{th}$ 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 $\left(\frac{N+1}{4}\right)^{th}$ 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 = $Q_3 - Q_1$

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

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

= Value of $\frac{543 + 1}{4}^{th}$ item

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

= Value of 136 th item

= 40 years

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

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

= Value of 3 x 136 th item

= 408th item which is 60 years

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{60 - 40}{2} = \frac{20}{2} = 10 \text{ years}$$

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

∴ Q_1 lies in the group 32–34

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

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

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

= 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

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

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

$$= 38+0.75$$

$$= 38.75$$

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

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

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

$$= 2.85$$

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

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

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

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

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

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

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 $(\overline{x})$ can be denoted by the symbol M.D. $(\overline{x})$. 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 4. 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	Σfx=576		Σf|d|=36

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

(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 $x_1, x_2, \ldots, x_n$ and if $\bar{x}$ be the arithmetic mean of these values, then

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

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:

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

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

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, 6 Q.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 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 specially 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:

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

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 5. 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 $x = \bar{x}, (\bar{x} = 17)$	$\left(x - \bar{x}\right)^2$
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 $\left(x - \bar{x}\right)^2$ 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		$\Sigma d' = -5$	$\Sigma d'^2 = 85$

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

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

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

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

$$= 2.87 \times 2$$

= 5.74

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

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

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 6. 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		$\sum fd' = -9$	$\sum fd'^2 = 35$

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

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

$$= \sqrt{3.5 - 0.81} \quad = \sqrt{2.69}$$

$$= 1.64$$

Example 7. 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.($\bar{x}$) = $A + \frac{\sum fd'}{N} \times h$

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

= 32.5 - 2.99 = 29.51

The required A.M. = 29.51 years.

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

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

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

= $\sqrt{2.5191} \times 5$

= 1.5871 x 5 = 7.936

∴ The required S.D. = 7.936 years.

Example 8. 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.($\bar{x}$)= $A + \frac{\sum fd'}{N} \times h$

$= 64.5 + \frac{27}{75} \times 10$

= 64.5 + 3.6 = 68.10

∴ The required A.M. = Rs. 68.10

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

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

$= \sqrt{1.6933 - 0.1296} \times 10$ $= \sqrt{1.5637} \times 10$

= 1.2505 x 10 = 12.505 = 12.51

∴ The required S.D. = Rs. 12.51

Example 9. 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

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

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

= $\sqrt{1.4105} \times 10$

= *11.876*

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

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

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

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

= *55 – 0.277 = 54.723*

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

$\therefore$ C.V. = $\frac{11.876}{54.723} \times 100\%$ = ***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.



3.4 LET US SUM UP

- In this unit we have learnt the meaning of measure of dispersion.
- 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.



3.5 FURTHER READINGS

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: 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).



3.7 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.

*** ***** ***

UNIT 4 : SKEWNESS, MOMENTS AND KURTOSIS

UNIT STRUCTURE

4.1 Learning Objectives

4.2 Introduction

4.3 Skewness

4.4 Measures of Skewness

- 4.4.1 Karl Pearson's Co-efficient of Skewness
- 4.4.2 Boweley's Co-efficient of Skewness

4.5 Moments

- 4.5.1 Moments About Mean
- 4.5.2 Moments About Arbitrary Point A
- 4.5.3 Relation Between Central and Raw Moments

4.6 Karl Pearson's Beta (β) and gamma (γ) coefficient

- 4.6.1 Coefficient of Skewness Based on Moments

4.7 Kurtosis

- 4.7.1 Measures of Kurtosis

4.8 Let Us Sum Up

4.9 Further Readings

4.10 Answers To Check Your Progress

4.11 Model Questions

4.1 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.

4.2 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 spread 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.

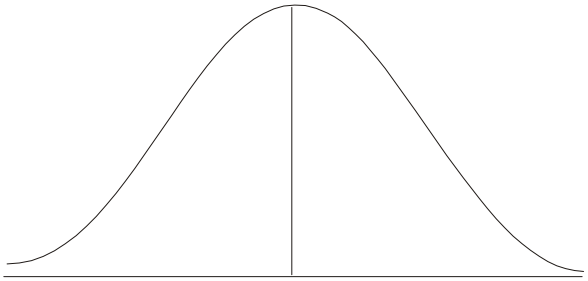
4.3 SKEWNESS

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. For which the curve has a longer tail towards the right is said to be **positively skewed** (Fig 4.2) and if the longer tail lies towards the left, it is said to be negatively skewed (Fig 4.3).

Note : 1) For a symmetrical distribution, a Bell-shaped (normal curve) is

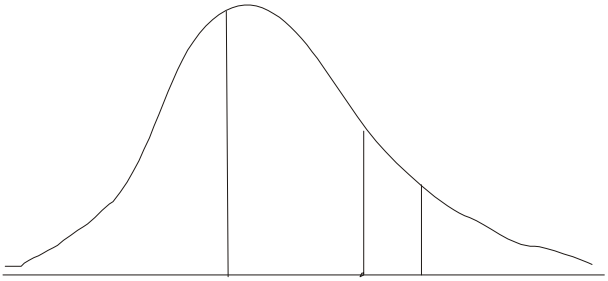
obtained for which Mean = Median = Mode i.e. these three values fall at the same point (Fig 4.1)



Mean = Mode = Median

(Fig 4.1)

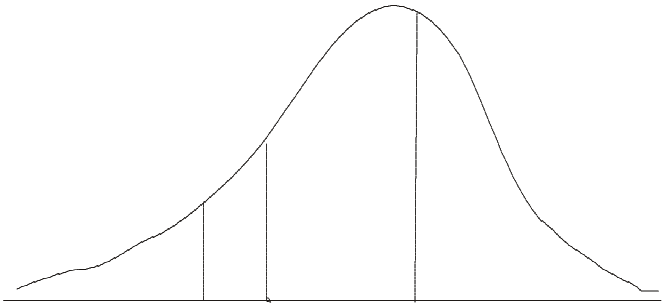
(Symmetrical Distribution)



$\bar{x}$ > Median>Mode

(Fig 4.2)

(Positively skewed distribution)



$\bar{x}$ < Median < Mode

(Fig 4.3)

(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.

$$Q_3 - \text{Median} \neq \text{Median} - Q_1$$

4.4 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.

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

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

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

In that case,

$$Sk_p = \frac{3(Mean - Median)}{S.D.}$$

Remarks :

1. Theoretically, Karl Pearson's Co-efficient of Skewness lies between ± 3. But in practice these limits are rarely 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.

4.4.2 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

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

Remarks :

1. The values of Sk_{KB} 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

$$SK_p = \frac{\bar{x} - Mode}{\sigma}$$

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

$\Rightarrow$ 3.2 = 30 – Mode

$\Rightarrow$ Mode = 30 – 3.2 = 26.8

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

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

$\Rightarrow$ 3.2 = 90 – 3 Median

$\Rightarrow$ 3 Median = 86.8

$\Rightarrow$ 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,

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

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

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

$\therefore$ 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.

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

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 based on quartiles is 0.6. If the sum of upper and lower quartiles 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 $\Sigma fd' = 22 \quad \Sigma fd'^2 = 130$

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)	(f)	Mid value(x)	d = x–A	$d^{/}$=d/5	fd^1	fd^{12}	Class Boundarie
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				4= $\Sigma fd^{/}$	208	
	= N					= Σfd^{12}	

Let A = 147 Here i = 5

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

$$= 147 + \frac{4 \times 5}{100} = 147.2 \ cm$$

$$\sigma = \sqrt{\frac{\Sigma fd^{12}}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

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

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

= 7.21 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 cm

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

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	50 & above
No. of employee	13	29	46	60	112	94	45	21

Solution : This is an open-end 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 coefficient of Skewness

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

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

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

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

= 34 + 2.77 = 37.77 yrs

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

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

= 42 + 2.93 = 42.93 yrs

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

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

4.5 MOMENTS

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

4.5.1 Moments About Mean

Let a variable x takes the values $x_1, x_2, \ldots\ldots x_n$. 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 μ_r [where μ is the letter (mu) of the Greek alphabet] is defined as

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

Similarly, for a frequency distribution of a variable x

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

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

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

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

Putting r = 1 in (1) and (2) we get

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

$\mu_1 = \frac{1}{N}\Sigma f(x-\bar{x})^r = 0$ $\rightarrow$ 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.

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

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

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

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

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

4.5.2 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

$$\mu'_r = \frac{\Sigma(x-A)^r}{n} \qquad \rightarrow \text{For individual series}$$

and $$\mu'_r = \frac{\Sigma f(x-A)^r}{N} \qquad \rightarrow \text{For frequency distribution.}$$

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.

4.5.3 Relation Between Central and Raw Moments

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, $$\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$$

$$= \frac{1}{n}\Sigma\{x - A - (\bar{x} - A)\}^r$$

$$= \frac{1}{n}\Sigma\left(x - A - \mu'_1\right)^r$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\ \mu'_1 \Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\ \Sigma(x-A)^{r-2} + \ldots\ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \frac{1}{n}\left[\Sigma(x-A)^r - {}^rC_1\ \mu'_1 \frac{1}{n}\Sigma(x-A)^{r-1} + {}^rC_2(\mu'_1)^2\ \Sigma(x-A)^{r-2} + \ldots\ldots\ldots + (-1)^r(\mu'_1)^r\right]$$

$$= \mu'_r - {}^rC_1\ \mu'_1 - \mu'_{r-1} + {}^rC_2\,(\mu'_1)^2\,\mu'_{r-2} + \ldots\ldots\ldots + (-1)^r(\mu'_1)^r$$

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

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

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 about 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) x (17) + 2 (– 1.5)³

= – 30 + 76.5 – 6.75 = 39.75

Fourth central moment about mean

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2 - 3(\mu_1')^4$$

$= 108 - 4 \times (-1.5)(-30) + 6(-1.5)^2 \times 17 - 3(-1.5)^4$

$= 108 - 180 + 229.5 - 15.1875$

$= 142.3125$

4.6 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.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} \quad \left[\because \mu_2 = \sigma^2\right]$$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

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.

4.6.1 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 β_1 = 0 or $\beta_2 + 3 = 0 \Rightarrow \beta_2 = -3$

But $\mu_4 = \frac{1}{n}\Sigma(x - \bar{x})^4 > 0$

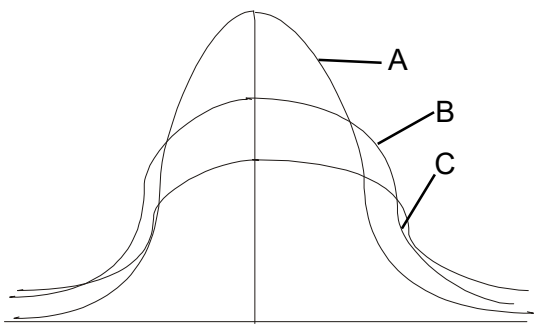
and $\mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 > 0$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^{\,2}} > 0$$

Since β_2 can not be negative, SK = 0 if $\beta_1 = 0$.

4.7 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 curve and 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**

4.7.1 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:

$$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{\mu_4}{6^4} \qquad (1)$$

$$\alpha_2 = \beta_2 - 3 = \frac{\mu_4}{6^4} - 3 \qquad (2)$$

For a normal curve (mesokurtic 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, $\mu_1' = 2, \mu_2' = 20, \ \mu_3' = 40, \ \mu_4' = 50$

$\therefore$ 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

$$\delta_1 = \frac{\mu_3}{\mu_2^{3/2}} = 0, \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

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 $\beta_1 = +1, \ \beta_2 = 4$

Variance $= \sigma^2 = \mu_2 = 9$

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 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 ($\overline{X}$) and about zero.

Q 15: (a) The central moments of a distribution are given by (), 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.



4.8 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.



4.9 FURTHER READINGS

1) Gupta S. C. *Fundamental of Statistics.* Himalaya Publishing House.

2) Gupta S. C. and Gupta Indira. *Business Statistics.* Himalaya Publishing House.

3) P. N. Arora, Sumeet Arora, S. Arora, *Comprehensive Statistical Methods.* S. Chand and Company.

4) R. K. Ghosh and S. Saha. *Business Mathematics and Statistics.* New Central Book Agency (P) Ltd.

5) Agarwal D. R. *Business Statistics*. Delhi, Vrinda Publication.

6) Sharma J. K. *Business Statistics.* Pearson Education Ltd.



4.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 4: (i) Equal (ii) Leptokurtic (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.



4.11 MODEL QUESTIONS

Q 1: Averages dispersion, skewness and kurtosis are complementary to one another in understanding a frequency distribution. Discuss the Statement.

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 5: CORRELATION

UNIT STRUCTURE

5.1 Learning Objectives
5.2 Introduction
5.3 Correlation: Meaning and Concept
5.4 Types of correlation
5.5 Methods of Measuring correlation
5.6 Scatter Diagram and Correlation
5.7 Types of Correlation
5.8 Correlation Co-efficient
5.9 Underlying Assumptions
5.10 Interpretation of Correlation Co-efficient
5.11 Rank Correlation Co-efficient
5.12 Let Us Sum Up
5.13 Further Readings
5.14 Answers To Check Your Progress
5.15 Model Questions

5.1 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
- explain the essential features of correlation analysis
- explain and compare the various aspects of correlation

5.2 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 concept - correlation 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 1

A. Try to list some variables and constants that you find in common day to day life.

.....

.....

.....

5.3 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.

5.4 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.

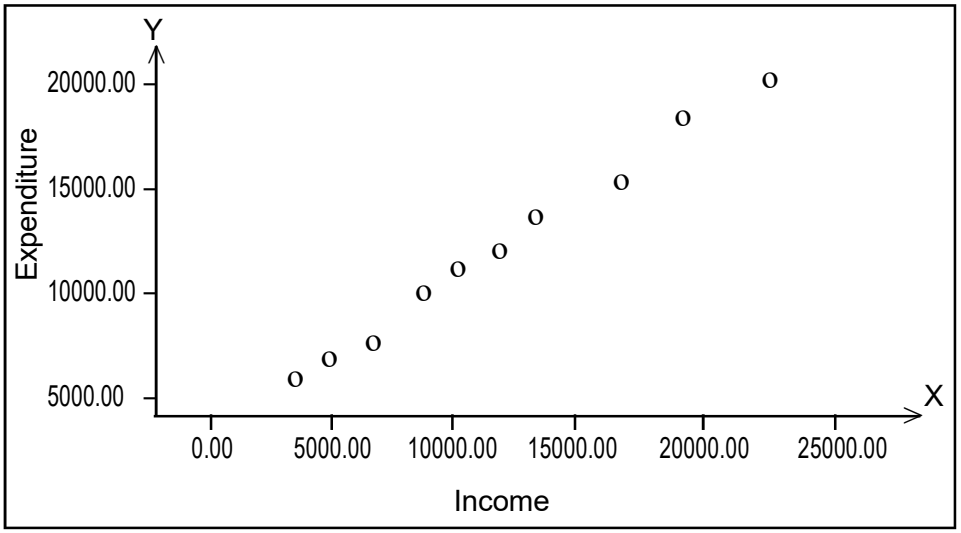
(iii) Number of workers and the time required to complete work.

(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.

5.5 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.



5.6 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 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?

5.7 TYPES OF CORRELATION

Depending upon the pattern that the scatter exhibits, we may have two distinct types of correlation - positive and negative.

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. Such a movement will result in scatter-plot to show an upward pattern of points or dots from left to right.

Negative correlation is, on the 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. Obviously, scatter-diagram for negatively correlated variables will be a downward slopping pattern from left to right.

In case, variables are found as uncorrelated i.e. change in one of the variables fail to produce any corresponding change in the other, then it is called zero correlation. The scatter exhibits no identifiable pattern in case the variables are uncorrelated.



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

.....

5.8 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 σ_x and σ_y are standard deviations of x and y respectively. Formulas for obtaining these are as below,

$$Cov(x,y)=\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

$$\sigma_x=\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}$$

$$\sigma_y=\sqrt{\frac{1}{n}\sum(y-\bar{y})^2}$$

Putting these in the original formula and simplifying, we will have,

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

This can be further simplified and can be written as

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2\ ]}}$$

This formula is relatively easier to calculate and most commonly used to obtain the Pearson's coefficient of correlation; Here 'n' 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. $\sum x$
- Obtain total of y i.e. $\sum y$
- Get product of x and y for all pairs and then get the total of xy i.e. $\sum xy$
- Get x^2 and y^2 for all values of x and y and then get their totals i.e. $\sum x^2, \sum y^2$
- 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 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

5.9 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.

5.10 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

.....

5.11 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

$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$, n is the number of observation

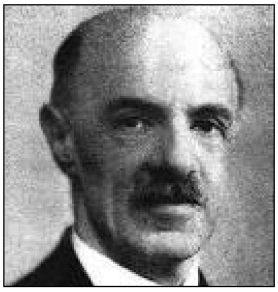
Procedure to obtain the rank correlation coefficient is fairly simple and easy:

- 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. $\sum d^2$
- 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 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



5.12 LET US SUM UP

- Depending on the number of variables, we have univariate, bivariate or multivariate distributions.
- Correlation is 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.



5.13 FURTHER READINGS

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.



5.14 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

Ans to Q No 4:

A. a) Average

B. c) Best

C. a) Minimum

D. a) 90^0

E. c) Mean

F. b) False

G. b) False



5.15 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 6: REGRESSION

UNIT STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.3 Meaning and Concept of Regression

6.4 Linear Regression

6.5 Line of Regression and Regression Equation

6.6 Coefficient of Regression

6.7 Relation between Correlation and Regression Coefficients

6.8 Correlation and Regression Analysis: A Comparison

6.9 Let Us Sum Up

6.10 Further Readings

6.11 Answers To Check Your Progress

6.12 Model Questions

6.1 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

6.2 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.

6.3 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.

6.4 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 4

A. Try to list five dependent variables and five independent variables. Also find out which depends on what.

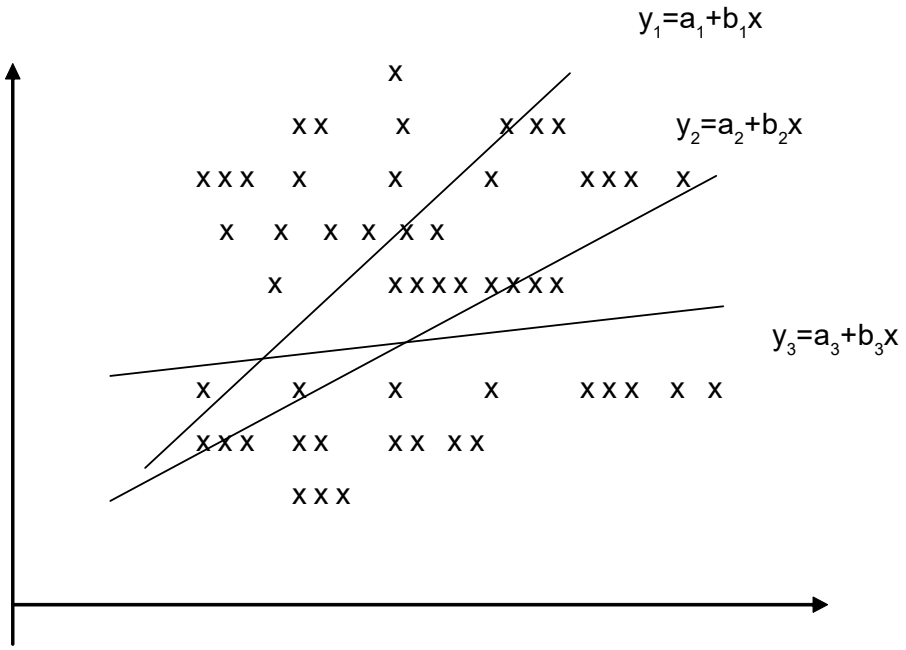
.....

.....

6.5 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.

6.6 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.

6.7 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

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$; 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 $\hat{b}_1$ and $\hat{b}_2$ respectively. Then they satisfy the following property

$$r^2 = \hat{b_1} \times \tilde{b_2}$$

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.

6.8 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 4: Choose 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 regression s

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

.....



6.9 LET US SUM UP

- 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.



6.10 FURTHER READINGS

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.



6.11 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

Ans to Q No 4:

A. a) Average

B. c) Best

C. a) Minimum

D. a) 90^0

E. c) Mean

F. b) False

G. b) False



6.12 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.

*** ***** ***